Proposing a Sparse Weighted Directed Network Construction Method and a Novel Mutual-Information based Sparse Feature Selection Algorithm for Multivariate Time-series Analysis & its Application in Medical Diagnostic Problems

by

RAHILSADAT HOSSEINI

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August, 2018

Supervising Committee:

Shouyi Wang (Committee Chairman)

Hanli Liu

Victoria Chen

Jay Rosenberger

To my parents ,Fataneh and Javad, and my husband, Ashkan,

for their endless love & support

TABLE OF CONTENTS

Acknowledgments

I would like to thank my supervisor Dr. Shouyi Wang for providing me with priceless research opportunities to collaborate with prestigious research labs and also for his invaluable advice during the course of my doctoral studies. I wish to thank my academic adviser Dr. Sheik N Imrhan for the guide and support.

I am very grateful to have Professor Victoria Chen in IE department. Her undeniable care and support toward each and every student and valuable expert knowledge has made her a role model and indispensable for the department. It was a great honor to meet with Professor Jay Rosenberger and Professor Bill Corley, key members of Cosmos lab in the IE department, and participate in high-level classes taught by them.

I would also like to extend my appreciation to Dr. Hanli Liu and Dr. Fenghua Tian, from the department of Bioengineering, at University of Texas at Arlington, for the excellent and extensive advice during various meetings and collaborations. I thank Dr. Liu for her interest in my research and for taking time to serve in my dissertation committee and leading my collaboration with department of pathology and urology in University of Texas, south western (UTSW). I thank Dr. Tian for the constructive discussions and comments also providing me the opportunity to collaborate with Speech, Language & Hearing Sciences at Purdue University.

I am especially grateful to Dr. Bridget Walsh from College of Heath and Human Sciences, at Purdue University for her interest in my research and for the helpful discussions and invaluable comments and taking the time to critically evaluate the manuscript we wrote.

Abstract

The main purpose of this study is feature engineering/learning from multivariate (MV) time-series to achieve a more interpretable model by dimension reduction. This aim is fulfilled in 2 main parts. In part 1, we proposed a network estimation approach namely SWDN which stands for sparse weighted directed network. In this approach, the directed subgraph of the underlying network was detected by maximum spanning tree (MST) algorithm that created a null model of connections with maximum inter-dependence (pairwise correlation or mutual information) forming the backbone structure of the MV time-series as an empirical reference. The edge weights were estimated using the linear conditional Gaussian parameters with the maximum likelihood. The efficiency of the proposed method (SWDN) was evaluated on the publicly available simulated fMRI data-set generated based on BOLD with different simulation parameters and in comparison with other network construction methods, it was verified to outperform Granger and lag-based methods under some circumstances. We applied SWDN as a feature extraction tool, and classified Parkinson's Disease (PD) fMRI data by finding the discriminative patterns between estimated network of PDs vs controls and achieved 75 % test accuracy via N-fold cross-validation.

In part 2, we made an extensive feature analysis framework for MV time-series. This framework consisted of extensive features extraction, post processing and a novel proposed feature selection technique based on mutual information and sparsity learning with embedded group structure. The multivariate time-series in this part was functional near infrared spectroscopy (fNIRS) which is a noninvasive neuroimaging technique for brain activity monitoring. We applied the proposed supervised exten-

sive sparse feature learning method on two data-sets to extract and select features and by applying machine learning and data mining approach and algorithms to classify participants with brain disorder/disease from the controls.

CHAPTER 1

Introduction

The main goal of this dissertation is feature learning and analysis of the multivariate time-series (MVTS), which is conducted in two main section. Section one is about structure learning or network construction of the MVTS which provides the connectivity analysis and reduced dimension. Section two is about a feature learning/engineering framework consisting of data pre-processing, exclusive feature extraction, post-processing and proposing the new feature selection technique based on the mutual information and sparse group structure of the multivariate time-series. The further application of these methods is in application of data mining and machine learning algorithms in order to classify binary labels. The proposed methods is applied on the brain neuroimaging data for evaluation like fMRI and fNIRS with the goal of efficient network construction and classification of binary labels for brain disease/disorder respectively.

The proposed methods have broad application in various fields with MVTS data. However our focus was on medical data and particularly brain informatics. Neuroimaging techniques like EEG, fNIRS, and fMRI provide a complex MVTS data in a subject-based or trial-based depending of the design of experiment which were suitable examples to apply and evaluate the efficacy of the proposed methods with practical goals. Next, we discuss the existing problems in the medical support system and how the proposed methods can be solutions along with the short descriptions of the embedded articles which represent the effectiveness, applicability and generalizability of the proposed methods.

1.1 Decision Support Systems in Medical Diagnostic Problems

Medical decision support systems are considered challenging because they are dependent on the subjective data of the patient and judgment of the assessor. Since 1970, application of computer systems and AI technology in medical field is growing rapidly [1,2]. One main reason is increasing facility in gathering the exact and abundant data with the new technology and the ability to easy storage, for example the soft-wares that can easily monitor the cognitive load and record the measurements for diagnosis like focus/attention, stress/anxiety. On one hand the data is increasing which can result is big models, on the other hand the need for sparse models are undeniable since smaller models are more interpretable and more generalizable. Therefore, the need for development of the methods which reduce the dimension of the data and can find the strong connection of the available variables in order to find the significant ones is also growing.

In this study we particularly focus on the application of the proposed methods on the brain activity measurements with 2 different techniques: functional MRI and functional near infrared spectroscopy (fNIRS). There are other techniques for brain data collection like electroencephalographic (EEG), and magnetoencephalographic (MEG), blood oxygen level dependent (BOLD) which we did not use for analysis. Generally, the challenges in development of a decision support system for medical data can be categorized in 3 groups: (1) probabilistic nature of the statistical modeling cause the output of the model to be fuzzy, there is no rigid simple yes or no. (2) Uncertainty in the gathered data is spite of targeting for a specific output. For example, a patient visiting a doctor for a speculated disease, may carry some other symptoms which one is not even aware of. However, the side problem can mislead the model. Another example is the brain computer experiments. Although the measurements are recorded after a specific stimulus on the brain, there is always uncertainty

about the data revealing the preplanned causal effects of the designed experiment because of the highly complex structure of the brain. (3) the validation of such models is very challenging. Most popular evaluation approach is cross-validation, however having and an additional test set is always more desirable. Moreover, imbalances/skewed classes is another source of challenge and makes the validation of the model in complex systems to be very difficult. Regarding the brain neuro-imaging data; the main challenge in the analysis of the collected data by any of the mentioned techniques, is high dimensional feature space while small-sample size; which cause over-fitting, poor generalization and non-interpretable solutions. This complex problem can be viewed and solved from different angles like network analysis, graphical modeling, Bayesian learning, dimension reduction using linear algebra techniques like principal component analysis (PCA) and independent component analysis (ICA), all with the goal of building an interpretable machine learning system. In this study we propose 2 methods as the solution to this problem and apply them on the simulated and real/experimental data. First, a network construction method based on the maximums spanning tree (MST) to detect the null model as a sub-graph of the underlying network, the model is evaluated on the simulated fMRI data and a real data from Parkinson Disease (PD), the results are shown in the article 1 (in Chapter 2). Second, we propose a data mining approach with the application of machine learning algorithms to exclusively extract features from the high-dimension MVTS and apply a mutual information based sparse group lasso feature selection technique to find the most significant features from the most significant groups of the voxels (channels) on the brain known as region of interest (ROI). These discovered features are known as biomakres which can be used in diagnostic and treatment-tracking of the patients with the brain disease/disorder. The result of the application of the second method on the children who stutter is shown in article 2 (in Chapter 3) and on the veterans

with post-traumatic stress disorder (PTSD) is shown in article 3 (Chapter 4). The machine learning approach for the further application of the proposed methods is defined to be binary classification. Clinical decision making in the field of machine learning can be categorized mainly to two groups: 1- supervised, when we have the known classes/labels of the samples/measurements and 2- unsupervised, when we need to detect the clusters of the similar samples. In this thesis, we mainly focused on the binary classification decision making.

1.2 Research Objectives and Contributions

This PhD dissertation consists of two main sections listed below:

- Chapter 2: structure learning / network construction of the multivariate time-series and introducing its two applications. First, connectivity analysis of the MVTS, detecting the weighted directed network (subgraph) with the most likelihood as a null model of the underlying network. Second, using the learned structure as the extracted features for classification using maximum margin to find discriminative patterns between binary labels.
- Chapter 3 and 4: exclusive feature learning framework including the proposing novel feature selection algorithm based on mutual information and sparse group lasso. This approach combines the nonlinear and linear dependence of the features to the class and results in a sparse selection which can be introduced as non-localized biomarkers which are data-driven-based instead of prior-knowledge-based.

1.3 Authorship

The first and primary author of the 3 articles embedded in the dissertation is Rahilsadat Hosseini, who conducted all the advanced data analytics under the supervision of Dr Shouyi Wang in COSMOS lab, at IMSE department in UTA. I applied data mining and machine learning techniques to solve the problem and propose new techniques.

In the first paper namely "Construction of Sparse Weighted Directed Network (SWDN) from the Multivariate Time-series & its Application as Feature Extraction for the Binary Classification", in Chapter 2, the coauthor was Dr. Shouyi Wang, my supervisor, who guided me through the method development and evaluation, suggesting similar articles to read and apply moreover, giving effective critics to make the paper more consistent and stronger.

In the second paper, namely "An fNIRS-Based Feature Learning and Classification Framework to Distinguish Hemodynamic Patterns in Children who Stutter", in Chapter 3, the main collaborator was Dr. Bridget Walsh, who had the NIH grant to design the experiment and collect the fNIRS data from children who were stuttering and who were controls and the group who were treated. She provided me with expert knowledge of the main problem definition and explanation/interpretation of the achieved results. Dr. Tian was the other collaborator who guided me through the analysis.

In the third paper, namely "Biomarker Identification of Post-traumatic Stress Disorder from Functional Near-Infrared Spectroscopy using a Novel Mutual Information-Guided Sparse Feature Learning Approach", in Chapter 4, the collaborators were Dr. Hanli Liu and Dr. Fenghua Tian from bioengineering department at UTA. They collected the fNIRS data from veterans with PTSD and controls and provided me

with expert knowledge about nature and description of fNIRS data and the disorder, moreover, verification of the biological interpretation of the achieved results.

REFERENCES

[1] E. Coiera, *Guide to Health Informatics*, ser. Get Through. Arnold, 2003. [Online]. Available: https://books.google.com/books?id=ptUM_lbU59QC

[2] R. A. Miller, "Medical diagnostic decision support systems–past, present, and future: a threaded bibliography and brief commentary." *J Am Med Inform Assoc*, vol. 1, no. 1, pp. 8 – 27, 1994.

[3] R. Hosseini, B. Walsh, F. Tian, and S. Wang, "An fnirs-based feature learning and classification framework to distinguish hemodynamic patterns in children who stutter," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 6, pp. 1254–1263, June 2018.

CHAPTER 2

Construction of Sparse Weighted Directed Network (SWDN) from the Multivariate Time-series & its Application as Feature Extraction for the Binary Classification

The short version of the paper namely "Construction of Sparse Weighted Directed Network (SWDN) from the Multivariate Time-series" was submitted to the 11th International Conference on Brain Informatics (BI 2018, December 7-9, Arlington, Texas, USA)

However, the long version of the paper in Elsevier format for the journal of NeuroImage is included in this dissertation.

Construction of Sparse Weighted Directed Network (SWDN) from the Multivariate Time-series & its Application as Feature Extraction for the Binary Classification

Rahilsadat Hosseini, and Shouyi Wang

701 S Nedderman Dr. University of Texas at Arlington, 76013, Arlington, TX

Abstract

There are many studies focusing on network detection in multivariate (MV) time-series data. A great deal of focus have been on estimation of brain networks using fMRI, fNIRS and EEG. We propose a sparse weighted directed network (SWDN) estimation approach which can detect the underlying minimum spanning network with maximum likelihood and estimated weights based on linear Gaussian conditional relationship in the multivariate time series. Considering the brain neuro-imaging signals as the multivariate data, we evaluated the performance of the proposed approach using the publicly available fMRI data-set and the results of the similar study which had evaluated popular network estimation approaches on the simulated fMRI data. Moreover, we applied the proposed network construction method as a feature extraction technique from fMRI data to classify the patterns of the Parkinson Disease.

Keywords: multivariate time-series, sparse weighted directed network (SWDN), feature extraction, classification, fMRI

1. Introduction

MV time-series analysis is used to investigate the concept of connectivity in dynamic systems like physiological time series. Connectivity analysis can detect coupling which means the presence or absence of interactions between the processes and identify causality which means the presence of driver-response relationships. There are different approaches to transform MV time-series in to a network through mapping algorithms. A classic but popular approach is considering each one of time-series as a node, and the weight of the edge connecting nodes would be interdependency between pairwise data [1] like correlation matrices [2]. Another recent approach is mapping the time-series into abstract graphs [3] for example the visibility algorithms that is applied on uni-variate time-series [4]. Another way to perform connectivity assessment include linear MV autoregressive (MVAR) process, and deriving measurements like coherence, the partial coherence, the directed coherence and so on from the frequency domain. Dynamic dependence model (the extensions of multi-regression dynamic models (MDM) [5]) is another available approach that map time-series to directed graphical models in which causality over time is decided based on the contemporaneous values of each one of the time-series as the predictor in a conditional relationship. Later sparsity was induced to this network using sequential Bayesian mixture modeling [6].

There is a growing interest in brain network estimation. Brain connectivity/network reveals the linking patterns in the brain which happens in different layers from neurons to neural assemblies and brain structures. Brain connectivity involves 3 concepts: neuroanatomical or structural connectivity (pattern of anatomical links), functional connectivity (usually understood as statistical dependencies) and effective connectivity (referring to causal interactions). Generally brain network estimation can be conducted in two main approaches, first, pairwise connectivity analysis like correlation, second, a convoluted approach to consider all the nodes globally like Bayes net modeling. Different methods can be applied on various brain imaging techniques. For example, MEG analysis of functional connectivity patterns based on the mutual information between wavelet time-series [7]. Another example is fNIRS (hemodynamic signals, such as HbO, HbR, and HbT responses) functional and effective connectivity analysis via Granger causality methods [8, 9, 10], pair-wise temporal correlation [11], frequency specific characteristics based on spontaneous oscillation in the low-frequency range [12], Dynamic Causal Modeling (DCM) [13] i.e. fitting differential equation or state space models of neuronal activity to brain imaging data using Bayesian inference [14], statistical parametric mapping (SPM) applying the general linear model (GLM) and random field theory[15] and fast causal inference algorithm [16]. The other significant group of studies focus on the analysis of fMRI network [17], using Granger causality [18], dynamic causal modeling [19], structure learning of sparse Markov networks specifically Gaussian graphical models incorporated with variable selection using block coordinate descent method [20], a regularized regression (Elastic Net) [21] and exploit the interactions by sparse Markov random field classifiers and linear methods, such as Gaussian Naive Bayes and SVM [22].

Smith et. al. [23] generated various fMRI simulations based on BOLD and evaluated the efficacy of different network construction methods. The 28 simulations varied based on simulation factors including number of nodes, session duration,

TR/neural lag, noise, haemodynamic response function (HRF) standard deviation and other factors like shared inputs, global mean confound, bad ROI (mixed and new random), backward connections, cyclic connections, stronger connections, more connections, non-stationary & stationary connections and only-one-strong input. The tested network modeling techniques were correlation and partial correlation, regularized inverse covariance (ICOV), mutual information, Granger causality (conditional, pairwise, directed and causality difference) and related lag-based measures, PDC (partial directed coherence), DTF (directed transfer function), coherence, generalized synchronization (Gen Synch), Patel's conditional dependence measures, Byes Net and LiNGAM (Linear, Non-Gaussian, Acyclic causal Models). The four evaluation metrics were defined as follows: Z-score true positive (TP), Z-score false positive (FP), c-sensitivity i.e. the fraction of TPs that are estimated with higher connection non-normalized strengths than the 95th percentile of the raw non-normalized FPs and total number of true connections and lastly, d-accuracy i.e. mean fractional rate of detecting the correct directionality of true connections. Evaluation of the network methods are summarized as follow: first-rank performing methods with c-sensitivity about 90% were: Partial correlation, ICOV and the Bayes net. The second-rank with 70-80% were: full correlation and Patel's. The third rank with 50% were: MI, Coherence and Gen Synch. The forth rank with poor performance of under 20% were: the lag-based methods (Granger, PDC and DTF) and LiNGAM. Regarding the detection of the direction of the connection, none of the methods were accurate except Patel's with 65%. The effect of factors are summarized as follow: longer duration resulted in higher c-sensitivity and had strong dependency with detection of directionality. Duration was more effective than TR and TR was more effective than noise level. Bad RIO was significantly deteriorating. The number of nodes and the addition of a global mean confound had complex patterns of effect.

In this study, we aim to learn the structure of a multivariate time-series and construct a graphical data-driven model using minimum spanning tree, maximum likelihood and linear conditional Gaussian dependance. The biggest challenge in structure learning when having no prior knowledge about the structure, is finding the highest score structure which is NP-hard. A very complex yet powerful approach is Bayesian learning in which each variable is assumed to have a specific distribution and variables are conditioned on each other, and final model is selected with methods like Monte Carlo Markov chain (MCMC) to sample from the posterior distribution and maximizing expected posteriors or BIC. However there are less computationally-complex approaches which are popular and commonly applied for example correlation, regularized inverse covariance, mutual information, Granger causality and so on.

The first purpose of this study is to apply the proposed network construction method to a variety of MV time-series in order to evaluate the efficacy of the method in comparison with other network estimation methods. As an example of the multivariate time-series, we applied the method to estimate functional connectivity in fMRI measurements which shows the temporal statistical correlation among neural assemblies. The fMRI data was publicly available from study [23] that consisted of 28 sessions of BOLD simulated fMRI data, each simulation had different properties including, number of nodes, session duration, TR (repetition time), neural lag, noise, HRF standard deviation. We exploited the results from study [23] and compared the performance of SWDN with similar evaluation metrics including relative sensitivities to finding the presence of a direct network connection, ability to find the direction of the connection, and robustness against various network challenges. Another purpose besides connectivity analysis from the network construction of the MV time-series is using structure-learning as a feature extraction technique and build a network-based feature-space for predictive models.

2. Method

2.1. Data Description

As it has been explained in the previous section, we used the public fMRI data to evaluate the network detection. The BOLD timeseries fMRI data was generated based on dynamic causal modeling (DCM) in 28 sessions with 50 subjects, varying time-stamp points and simulated with different properties. The session properties are retrieved from the study [23] and is summarized in Table (1).

In the second part, we applied the proposed method on the experimental fMRI data collected from the participants with and without Parkinson Disease (PD) and estimated the underlying network for each subject. Next, with the aim of classification, we used the estimated network weights as the extracted features to find the discriminative patterns between controls and PDs. The fMRI measurements consisted of 21 controls and 25 participants with PD, each subject had 264 channels (nodes) with 300 number of time-points. All of the computations in this study were conducted in Matlab version 2017.

2.2. Maximum Spanning Tree (MST), Adjacency Matrix and Graph

Maximum spanning tree is the same as minimum spanning tree but with the selection of edges with maximum weigh at each iteration. Minimum spanning tree as a sub-network containing the strongest connections, has successfully been applied to detect the null model of connections that form the backbone structure of the brain to create an empirical reference [24], moreover to capture network alterations due to aging and disease in functional and structural imaging data [25, 26, 27]. We implemented the Prim's minimum spanning tree algorithm to find the underlying network. Prim's algorithm solves the problem of finding acyclic set connecting all vertices in V with the minimal weight, $w(T) = \sum_{(u,v)\in T} w(u,v)$, for a given connected undirected graph $G = (V, E)$, where each edge (u, v) has a weight $w(u, v)$. Prim's algorithm starts with a spanning tree, containing arbitrary vertex and no edge, it repeatedly adds

Table 1: Summary of the session properties of the simulated fMRI

Sim1, 5Nd, 200NTp	Baseline	Sim15, 5Nd, 200NTp	Stronger connection
Sim2, 10Nd, 200NTp	Baseline	Sim16, 5Nd, 200NTp	More connections
Sim3, 15Nd, 200NTp	Baseline	Sim17, 10Nd, 200NTp	Reduced noise
Sim4, 50Nd, 200NTp	Baseline	Sim18, 5Nd, 200NTp	Removed all HLV
Sim5, 5Nd, 1200NTp	1 hour session	Sim19, 5Nd, 2400NTp	Increased neural lag
Sim6, 10Nd, 1200NTp	1 hour session	Sim20, 5Nd, 2400NTp	Neural lag and removed HLV
Sim7, 5Nd, 5000NTp	4 hour session	Sim21, 5Nd, 200NTp	2-group
Sim8, 5Nd, 200NTp	Shared input	Sim22, 5Nd, 200NTp	Nonstationary connection strength
Sim9, 5Nd, 5000NTp	Shared input	Sim23, 5Nd, 200NTp	Stationary connection strength
Sim10, 5Nd, 200NTp	Global mean confound	Sim24, 5Nd, 200NTp	Only one strong external input
Sim11, 10Nd, 200NTp	Bad ROI - mixed	Sim25, 5Nd, 100NTp	Reduced noise
Sim12, 10Nd, 200NTp	Bad ROI - mixed	Sim26, 5Nd, 50NTp	2.5 min session
Sim13, 5Nd, 200NTp	Backward connectionn	Sim27, 5Nd, 50NTp	Reduced noise
Sim14, 5Nd, 200NTp	Cyclic connection	Sim28, 5Nd, 100NTp	Reduced noise

Sim: simulation, Nd: number of Nodes, NTp: Number of time-points, HLV: haemodynamic lag variability

edges with minimum weight and grows the spanning with a vertex not in the tree in a greedy way. We defined a priority queue for the vertices not in the tree, using a pointer from adjacency matrix as the list of entry, in order to find the minimal edge connected to the tree. The key of the vertex is weight of the edge connecting it to the tree. This greedy algorithm works in $O((|V| + |E|) \log |V|) = O(|E| \log |V|)$ running time while loop runs $|V|$ times.

In maximum spanning tree, the set is found by vertices with maximum weight. Weight is calculated as the multivariate linear or nonlinear dependance matrix using pairwise mutual information (MI) and correlation. As described in the Algorithm (1), the tree starts with connecting all vertices ($v_{1...n} \in V$) to the root, then a queue is listed for entering the nodes (Q). The child (F) of the root, is decided by having the maximum weight ($keys(v)$) among all edges (E), if the weight of the new edge is greater than the current weight ($W(F, v) > keys(v)$). The selected (v) is then removed from the Q. This repeats till the queue become empty and predecessor (p) for all ($v \in Q$) is decided while root is the only vertex without parent. This procedure iterates for ($\forall root \in V$) and the best tree is selected based on the maximum likelihood of the ($data|G, model$).

Algorithm 1 Maximum spanning tree (MST), an implementation of Prim's minimum spanning tree

$Q = V - \{root\}$
$p(v) = root \quad \forall v \in Q$
$keys(v) = W(root, v) \quad \forall v \in Q$
while $Q \neq \emptyset$ **do**
  $F = \underset{v \in Q}{\operatorname{argmax}}\, keys(v)$
  $Q = Q - \{F\}$
  for $v \in Q$ **do**
    if $W(F, v) > keys(v)$ **then**
      $p(v) = F$
      $keys(v) = W(F, v)$

The constructed tree is a compact joint representation over unstructured variable representation, much less smaller network. This tree network helps to speed up enumeration and eliminate variable and is the basis to construct the adjacency matrix. Adjacency matrix is a 0-1 matrix that takes a n-by-n weight matrix and returns a list of the maximum weight spanning tree. If there is a predecessor, then $pred(i) = 1$, otherwise it is zero. It can be either symmetric or non-symmetric. We converted the adjacency matrix to a graph by defining a matrix in size of adjacency matrix but with 2 columns. Column one defines the existence of an edge (binary), column 2 defines the parent node, each row represents the child node. Adjacency matrix is used to calculate the linear parameters of the conditional Gaussian graphical model. There are two assumptions for simplicity and being allowed to use linear systems: all nodes follow Gaussian distribution [28] and child-parent (edge) have linear relationship [13].

2.2.1. Conditional Gaussian Distribution

We used the constructed graph to detect edges between children and parents, and to fit linear Gaussian between them. Next, we estimated parameters of the linear Gaussian model (β) using Equation (1, 2) where $C_{M\times1}$ represents the child variable with M examples and $U_{M\times N}$ represents N parents ($U_1, \ldots, U_n$) each with M examples.

$$C|U \sim N(\beta(1) * U_1 + \beta(n) * U_n + \beta(n+1), \sigma^2) \quad (1)$$

$$\sigma = \sqrt{cov(C) - \sum(\sum(\beta * \beta'. * cov(U)))} \quad (2)$$

In Equation (3), (A) represents the expectations matrix and is required to solve the linear system ($A \times \beta = B$) where (B) as the right hand side of the equation follows Equation (4).

$$A = \begin{bmatrix} E[U_1] & E[U_2] & \dots & E[U_n] & 1 \\ E[U_1 * U_1] & E[U_2 * U_1] & \dots & E[U_n * U_1] & E[U_1] \\ \vdots & \vdots & \dots & \vdots & \vdots \\ E[U_1 * U_n] & E[U_2 * U_n] & \dots & E[U_n * U_n] & E[U_n] \end{bmatrix} \tag{3}$$

$$B = \begin{bmatrix} E[X] \\ E[X * U_1] \\ \vdots \\ E[X * U_n] \end{bmatrix} \tag{4}$$

We used log-likelihood to evaluate the data given the model and graph structure ($data|G, P$), where P is the structure array of estimated parameters (β) for the linear Gaussian. The selected model is the one with the maximum likelihood.

$$E(x) = \beta(0) + \beta(1) * U(1) + ... + \beta(n) * U(n) \tag{5}$$

$$p(v) = \sum(G == v)/|V| \tag{6}$$

$$\text{log-likelihood}(u|v) = p(v) * \log(\sum(\exp((x - E(x))^2/2\sigma^2 - log(\sqrt{\pi}\sigma)))) \tag{7}$$

2.3. Network Evaluation Metrics

The four evaluation metrics were defined as follows: (1) normalized true positive (Z-score TP) i.e. normalized weight of the true connections (correctly detected edge when it existed in the ground-truth network), (2) normalized false positive (Z-score FP) i.e. normalized weights of the network for edges that are defined but should have been empty based on the ground-truth, (3) c-sensitivity i.e. the fraction of TPs that are estimated with higher connection non-normalized strengths than the 95^{th} percentile of the raw non-normalized FPs and total number of true connections and (4) d-accuracy i.e. mean fractional rate of detecting the correct directionality of true connections which can be calculated as difference of normalized weights in $node_{ij}$ and $node_{ji}$.

2.4. Max Margin Optimization of the weights

The constructed network was used as the feature extraction tool from the MV time-series. In this approach, MV timesseries of each subject was mapped to a square matrix of pairwise weights where non-zero value verified the existing of a linear relationship between parent and child node. Next, linear Gaussian coefficients were used as decision variables in a max margin optimization problem for classification task.
In Equation (8), $\beta_{i,j}$, represents the linear coefficient between child-node (i) and parent-node (j), C is a non-negative tuning parameter, M is the width of the margin; ϵ_i are error or slack variables that allow individual observations to be on the wrong side of the margin of the hyperplane, weights shown as $\hat{\omega}_{0,...N}$, when having n channels, $N = n^2$, and p is the sample size. To solve this problem, we used radial basis function (RBF) kernel and applied SVM-RBF models.

$$\max_{\omega_0,\omega_1,...,\omega_n,\epsilon_1,\epsilon_2,...\epsilon_p} M \tag{8}$$

$$\text{subject to } \sum_{j=1}^{n} \omega_j^2 = 1,$$

$$y_i(\omega_0 + \omega_1\hat{\beta}_{i1} + \omega_2\hat{\beta}_{i2} + ... + \omega_n\hat{\beta}_{in}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{p} \epsilon_i \leq C,$$

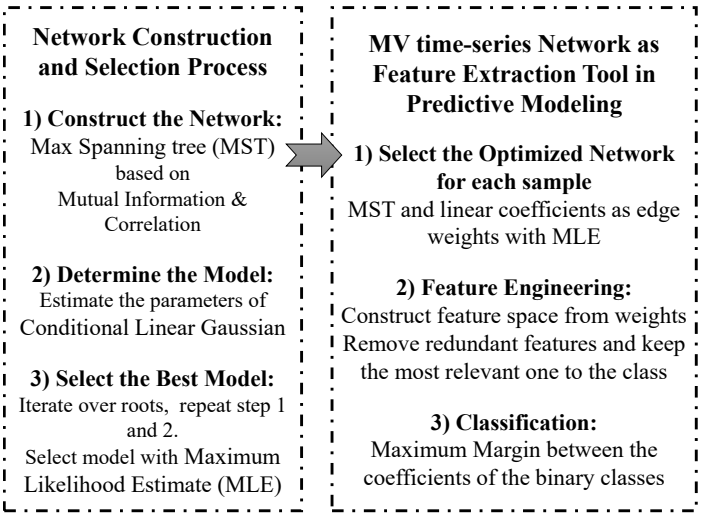


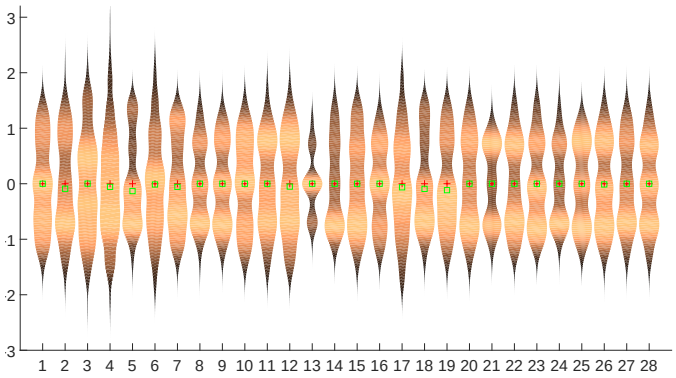
Figure 1: Summary of the process of MV time-series network construction and selection and its application as a feature extraction tool in the predictive modeling

The summary of the process for network construction and optimization and its application as the feature space for classification is shown in Figure (1). In step 2 of the predictive modeling application, in Figure (1), the feature selection method is described shortly, which is the criteria of max-relevance, and min-redundancy (mRMR) [29], the number of selected features is decided via cross-validation.

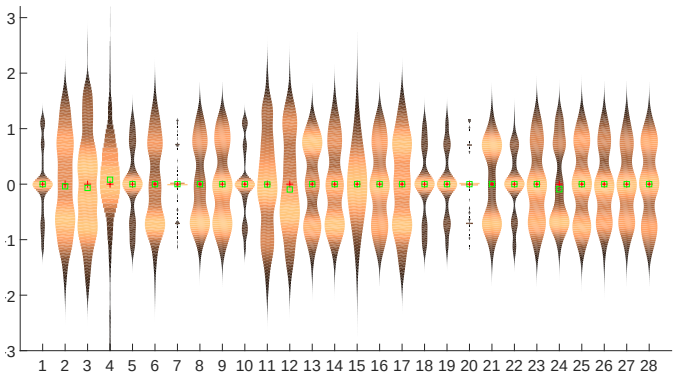
3. Results

3.1. Comparing SWDN with other Network Methods

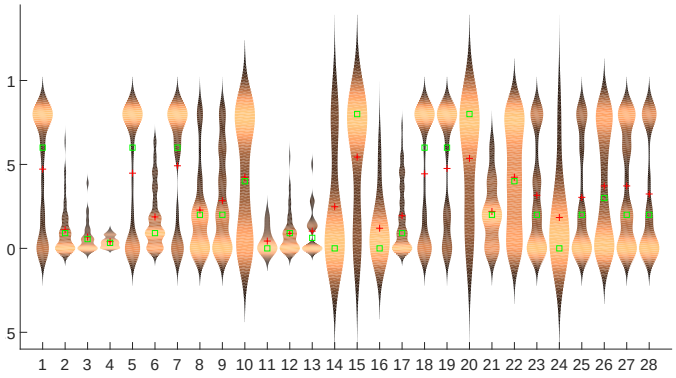
In Figure (2a, 2b, 2c and 2d) performance metrics Zscore TP, Zscore FP, c-sensitivity and d-accuracy are respectively calculated for SWDN. Based on the violin plots (vertical histograms that can depict multimodality), SWDN performs well in sessions 1, 5, 7, 15, 19 and 20, in which there is least overlap between distributions in Figure (2a and 2b) meaning that the c-sensitivity distribution shown in Figure (2c) has higher values with average mean and variance of 0.48 and 0.34 respectively. The sessions with very poor performance are 3, 4, 11, 12, 13, 16 and 24 in which c-sensitivity distribution has average mean and variance of 0.09 and 0.12 respectively.



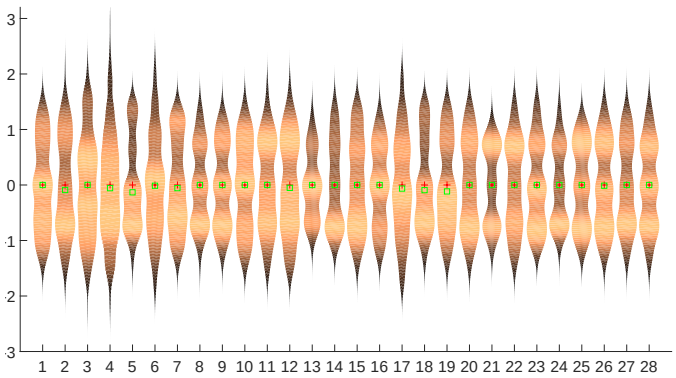
(a) Zscore TP



(b) Zscore FP

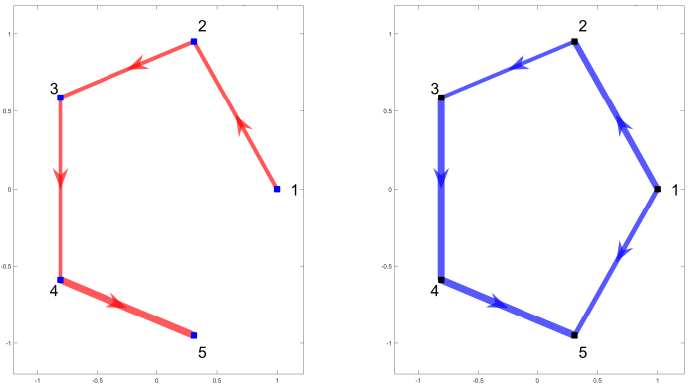


(c) C-sensitivity

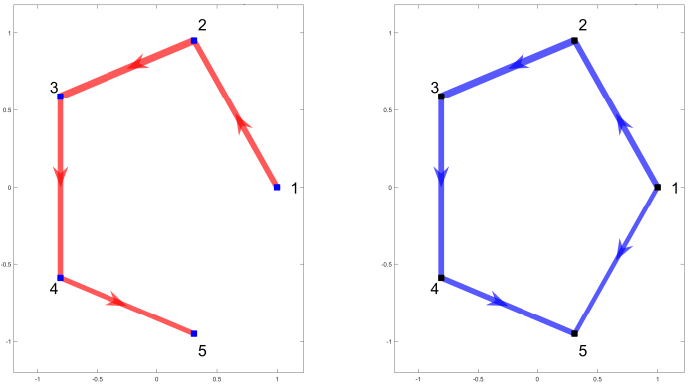


(d) D-accuracy

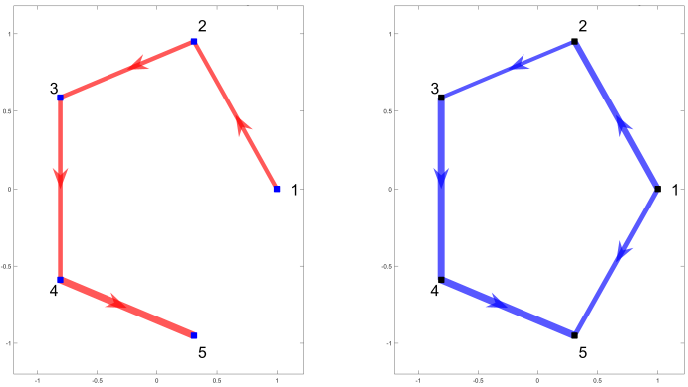
Figure 2: X-axis represents session IDs starting from 1 to 28, Y-axis represents the weights of all edges from all subjects.



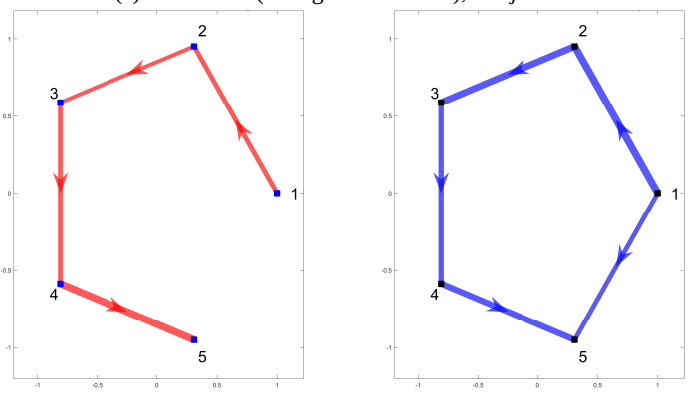
(a) Session 1 (5 nodes), Subject 1



(b) Session 7 (4 hours session), Subject 4

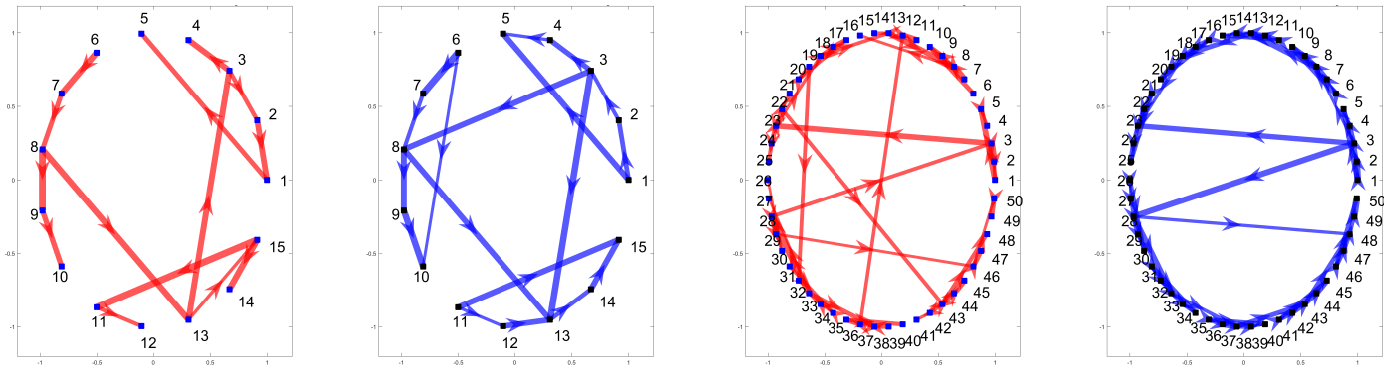


(c) Session 15 (stronger connection), Subject 1



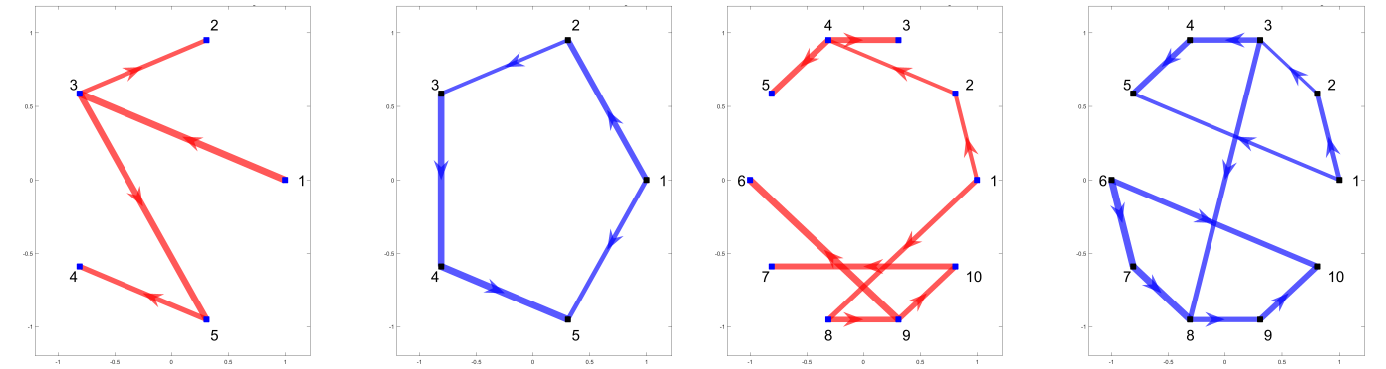
(d) Session 20 (neural lag and removed HLV), Subject 6

Figure 3: The best constructed network among subjects in a session by SWDN (in red, on the left) vs. the ground-truth network (in blue, on the right). These are the examples of excellent performance of SWDN under the promising circumstances like small network, longer duration, existence of strong connection and neural lag with removed HLV. Edge width is proportional to edge weight.



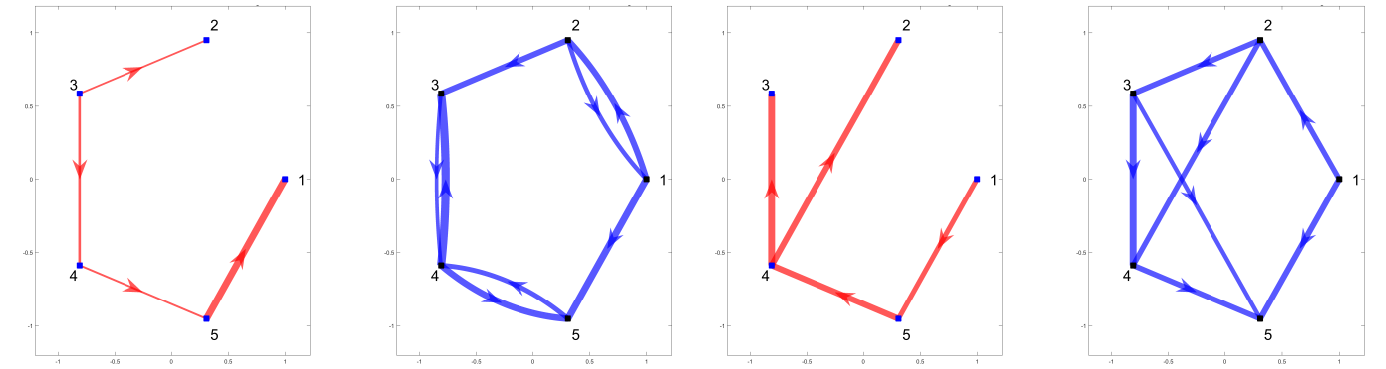
(a) Session 3 (15 nodes), Subject 2

(b) Session 4 (50 nodes), Subject 5



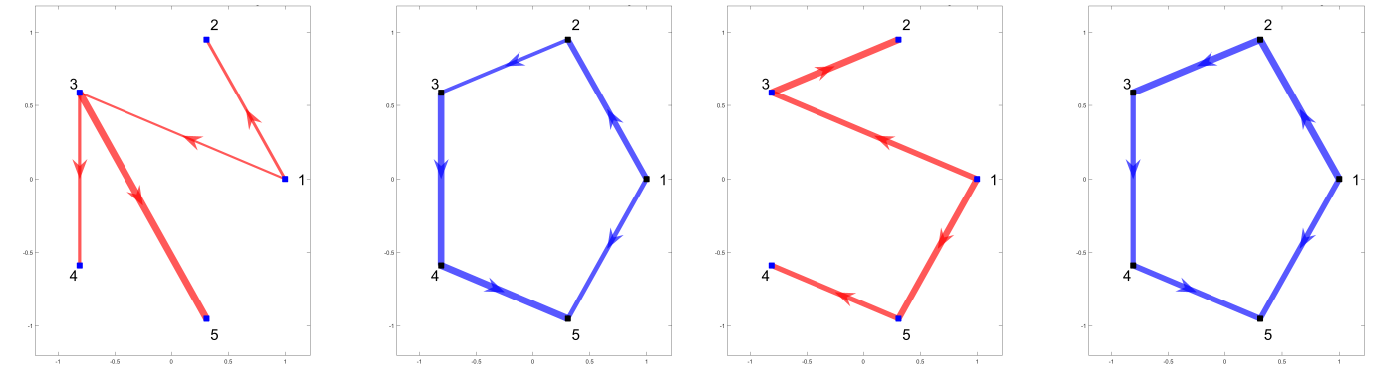
(c) Session 9 (shared input), Subject 1

(d) Session 11 (bad ROI), Subject 1



(e) Session 13 (backward connection), Subject 6

(f) Session 16 (more connections), Subject 3



(g) Session 24 (one strong external connection), Subject 1

(h) Session 26 (2.5 min session (shorter duration)), Subject 2

Figure 4: The worst constructed network among subjects in a session by SWDN (in red, on the left) vs. the ground-truth network (in blue, on the right). These are the examples of poor performance of SWDN under the challenging circumstances like bigger networks, shared input, bad ROI (mixed nodes), existence of backward connection, many connections and one strong external connection in the ground-truth network and shorter duration. Edge width is proportional to edge weight.

The capability of the proposed method (SWDN) in detection of the ground-truth network is compared with other networks' capabilities which are retrieved from the study [23] and the results are summarized below. In each item, the italic font sentences summaries the results taken from the study [23] and the words in bold font state the performance of SWDN in each stated simulation session.

- *In simulation 1, 2 and 3, Partial correlation, ICOV and the Bayes net performed about 90% of c-sensitivity, while lag-based methods (Granger, etc.) less than 20%.* **The proposed method (SWDN) performed with average of 53% c-sensitivity with the standard deviation of .35 in sim1 and2 and outperformed lag-based methods including Granger but significantly reduced sensitivity to 17% in sim3.**

- *In simulation 4, full correlation, ICOV and Patel's all performed excellently.* **In simulation 3 and 4 number of nodes increased to 15 and 50, SWDNs performance decreased, especially in sim4 to 10%**

- *In simulation 5 and 6, the duration was increased to 60 minutes, which caused the single lag-based method to reach higher sensitivity but poor d-accuracy suggested that it was not a trustworthy result. LINGAM was performing better in sim5 because of more time-points which improved better function for temporal ICA however, 10% reduced in sim6 because of more time-points.* **SWDN performed 50% in sim5 but decreased in sensitivity in sim6 (24%) since there were more nodes.** *In Sim7 with 5 nodes and 250 min duration, LINGAM outperformed all other methods among all sessions (90%).* **SWDN increased sensitivity in sim7 to 55% because of the longer duration and less number of nodes comparing to sim5 and 6.**

- *In simulation 8 and 9, shared inputs deteriorated all estimation methods to 60% and below.* **SWDN was not designed to capture shared inputs because it was based on minimum spanning tree (maximum of one parent for a child), therefore as expected the results was low and about 28-34% c-sensitivity.**

- *In simulation 10 with global mean confound, there was the same results as sim1.* **SWDN achieved 48% c-sensitivity, outperforming lag-based methods but behind Bayes and Partial correlation.**

- *In simulation 11 and 12 with bad ROI (mixed or new random), the results were extremely bad, all the methods lower than 20% in sim11 but much better in sim12.* **SWDN also performed poorly about 10% in both simulations.**

- *In simulation 13 with backward connections, all methods reduced sensitivity significantly with best method to be Bayes with 60% and Coherence, Gen Synch and MI like the correlation measures, ICOV and Patel's, all being around 50%.* **SWDN was not promising with 11%.**

- *In simulation 14, with cyclic connections, there were same results as sim1 but reduced d-accuracy.* **SWDN was not designed to capture cyclic connections and achieved 30% c-sensitivity.**

- *In simulation 15, with stronger connection, Partial correlation, ICOV and the Bayes net methods achieved 90%. Full correlation and Patel's fell to around 60%. MI, Coherence and Gen Synch were unchanged, and, Partial MI increased to 85%. Lag-based methods were performing very poorly (less than 30%).* **SWDN achieved 60% indicating that it outperformed lag-based methods in capturing stronger connections.**

- *In simulation 16, there were similar results as sim1 but lower sensitivity.* **Expectedly, SWDN could not estimate network with many connections because it was designed to capture sparse network with the most significant edges, therefore performing poorly about 17%.**

- *In simulation 17, which could be compared with sim15, Partial correlation, ICOV and the Bayes net had excellent performance, while lag-based methods (Granger, etc.) less than 20%. MI and Coherence increased to 70s%.* **SWDN with poor performance of 25% indicated that in comparison with sim15 it performed equally or better than Granger methods.**

- *Simulation 18 was similar to sim1 after removal of haemodynamic lag variability. The results were unchanged, all lag-based methods performed very poorly both with respect to c-sensitivity and d-accuracy.* **SWDN performed better than lag-based methods with 50% sensitivity indicating that existence of lag did not affect its performance.**

- *In simulations 19 and 20 with added neural lag, Partial correlation and ICOV achieved highest sensitivity 90s%, with some of the Granger approaches achieving 80s%.* **SWDN achieved 53 to 60% sensitivity.**

- *Simulation 21 tested the sensitivity of the methods at detection of changes in connection strength among subjects. And introduced the most sensitive method as Patel's, with t=7.4, Full correlation, Partial correlation, ICOV, Gen Synch and most of the Bayes net methods.* **SWDN was not very sensitive with 27%.**

- *In simulation 22, there were non-stationary connection strengths. Bayes net methods, correlation and ICOV achieving the c-sensitivity (78% and 70% respectively). Coherence measures were expected to be promising but they were not.* **SWDN was promising gaining 48% sensitivity.**

- *In simulation 23, there was stationary connection strength. Partial correlation and ICOV performed the best, but the Bayes net methods did not perform so well, falling to 60%.* **SWDN performed poorly around 37%. The decrease**

in performance was similar to the Bayes Net methods comparing to sim22.

- *In simulation 24 which was similar to sim15, but with only one strong external input, none of the methods had a c-sensitivity greater than 50%, and none had d-accuracy greater than 61%. Best performing methods Partial correlation and ICOV =5 resulted in 40s% and the Bayes net models performed badly in 20s%.* **SWDN performed similar to Bayes Net with 24% c-sensitivity.**

- *In simulations 25, 26 (shorter duration), 27 (shorter duration and reduced noise level) and 28 (reduced noise level), the three best-performing models resulted in 70s%, 50s%, 70s% and 80% sensitivity respectively.* **SWDN achieved 36, 43, 43, and 38% c-sensitivity respectively.**

In Figure (3) and (4) the excellent and poor performance of SWDN are respectively depicted via examples among subjects in each session. The examples are selected to visualize the strength and weakness of the method more clearly. SWDN is more efficient in estimation of the ground-truth network when it has less number of nodes and longer duration while it performs poorly with bigger networks and shorter duration. It was expected to have low sensitivity in estimation of specific networks (with complex simulation parameters) like backward, more connections, shared input and bad ROI because of the nature of spanning tree which is acyclic network with fixed number of edges in which each child node has maximum of one parent.

3.2. PD Classification using SWDN as feature extraction

We constructed the feature space from the weights of the network which are the linear conditional Gaussian parameters generated by SWDN. Next, we applied the N-fold cross-validation with $N = 5$, to divide data to training and testing and reported the classification performance as the average of the N-times repeated test-set. We used the sensitivity (true positive rate) and specificity (true negative rate), as the evaluation metrics to assess the performance of the classification. In this problem, we defined sensitivity as the accuracy of PD class and specificity as the accuracy of control class and reported total accuracy as the average of sensitivity and specificity.

The best SVM classification performance for the real dataset (PD) was 75.33% accuracy (the average of sensitivity: 86.66% and specificity: 64%). The proposed method, SWDN, outperformed the reported accuracy in study [30] for about 5%, in which Daehan Won [30] tried different classifiers on PD data and achieved maximum of 70% accuracy with 60 features and utilizing the sparse selection of nodes and edges in a leave-one-out cross-validation. This comparison suggested that SWDN successfully detected the backbone brain structure which was altering between PD and controls.

4. Discussion

The best performing sessions based on c-sensitivity shown in Figure (2c) for SWDN were sim1, 5, 7, 15, 18, 19 and 20 which all had network size of 5 which verified that SWDN is more effective with smaller networks. Sim5 ,7, 19 and 20 had longer duration, 1, 4 and 2 hours respectively, which also verified more efficacy of the method with longer duration. In sim15, SWDN was capable of capturing the strong connection. In sim18, 19 and 20, removed haemodynamic lag variability (HLV) and increased neural lag affected the SWDN's performance positively.

The sessions with poor performance were sim3, 4, 11, 12 and 13. Sim3 and 4 had network size of 15 and 50, which explains the poor performance of the method because of bigger size of the networks. Sim11 and 12 had bad ROI which caused the deterioration in performance and the results were consistent with other methods [23]. Sim13 had backward connections and SWDN was not capable of detection of such connections based on the nature of the spanning tree algorithm. Although SWDN does not capture all the edges in the ground truth, it is capable of detecting the most significant sub-graph. Moreover, while it does not reach to the level of the sensitivity of the computationally complex methods like Bayes Net, it can achieve a valuable performance only having low computational complexity.

With respect to the estimation of network connection directionality, SWDN was poor and was not able to detect higher than random accuracy. This conclusions is consistent with the results taken from study [23] which stated the d-accuracy of the methods to be at chance level (50%).

We achieved 75% accuracy in classification of PD data. Since there were 264 nodes in the network of each subject, it would be considered as a very big network, therefore, SWDN might not be the best performing method. However, Bayes net for such a big network is also very complex and time-consuming and not feasible. On the other hand the achieved results should be equal or better that the possible results with Granger and lag-based methods since SWDN outperformed lag-based methods in some sessions. Moreover, longer duration (more number of time-points, ≥ 300) of fMRI data could improve the current results, since SWDN was more efficient in longer duration which was the similar conclusion for other methods [23]. However, we achieved a comparable result with another study on the same data and it verified the capability of the SWDN to detect the empirical reference network as a null model connection forming backbone structure of the human brain which was sensitive to alterations in network topology between classes.

5. Conclusion & Future Studies

SWDN's main strength comes from the underlying network construction to be minimum spanning tree which generates a unique, acyclic, strongest sub-network with fixed number of connections. Minimum spanning tree is an unbiased method, which avoids several methodological biases like arbitrary thresholding and is insensitive to alterations in connection strength or link density. All of these advantages made SWDN capable of capturing the strongest sub-graph of the underlying network based on the MST algorithm.

SWDN as a network estimation method, was more compatible with smaller networks meaning less number of nodes (refer

to the good performance in sim1) and less number of edges (refer to poor performance in sim16) and longer duration of simulation (refer to the good performance in sim5, 7, 19 and 20). In 2 of the simulations (sim22 and 24), its performance was similar to Bayes Net, however it was much less complicated computationally and therefore less time-consuming. It outperformed lag-based methods like Granger (refer to sim1, 2, 14, 15, 17 and 18).

Moreover, SWDN as a feature extraction tool performed promisingly by capturing the alternating networks between class of PD and controls with 75% accuracy. This result was comparable to the achieved result (70% accuracy) in the study of the same data-set (PD) by optimization and machine learning approach for the analysis of complex network [30].

SWDN based on minimum spanning tree is not expected to estimate the dense networks or networks with backward or cyclic connections. The SWDN is a general network modeling framework that can incorporate more graphical structures in addition to minimum spanning tree. For example, we will extend the SWDN to model more generalized network that can have more than one parent-node, backward connections and cyclic structures.

References

[1] L. Lacasa, V. Nicosia, and V. Latora. Network structure of multivariate time series. *Scientific Reports*, 5, 2015.

[2] Y. Hu, H. Zhao, and X. Ai. Inferring weighted directed association network from multivariate time series with a synthetic method of partial symbolic transfer entropy spectrum and granger causality. *PLOS ONE*, 11(11):1–25, 11 2016.

[3] J. Zhang and M. Small. Complex network from pseudoperiodic time series: Topology versus dynamics. *Phys. Rev. Lett.*, 96:238701, Jun 2006.

[4] G. Gutin, T. Mansour, and S. Severini. A characterization of horizontal visibility graphs and combinatorics on words. *Physica A: Statistical Mechanics and its Applications*, 390(12):2421 – 2428, 2011.

[5] O. Anacleto, C. Queen, and C. J. Albers. Multivariate forecasting of road traffic flows in the presence of heteroscedasticity and measurement errors. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 62(2):251–270, 2013.

[6] Z.Y. Zhao, M. Xie, and M. West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32(3):311–332, 2016. asmb.2161.

[7] F. Siebenhhner, S.A. Weiss, R. Coppola, D.R. Weinberger, and D.S. Bassett. Intra- and inter-frequency brain network structure in health and schizophrenia. *PLOS ONE*, 8(8):1–13, 08 2013.

[8] B.R. White, A.Z. Snyder, A.L. Cohen, S.E. Petersen, M.E. Raichle, B.L. Schlaggar, and J.P. Culver. Resting-state functional connectivity in the human brain revealed with diffuse optical tomography. *NeuroImage*, 47(1):148156, 2009.

[9] Chang-Hwan Im, Young-Jin Jung, Seungduk Lee, Dalkwon Koh, Do-Won Kim, and Beop-Min Kim. Estimation of directional coupling between cortical areas using near-infrared spectroscopy (nirs). *Opt. Express*, 18(6):5730–5739, Mar 2010.

[10] Zhen Yuan. Combining independent component analysis and granger causality to investigate brain network dynamics with fnirs measurements. *Biomed. Opt. Express*, 4(11):2629–2643, Nov 2013.

[11] Fumitaka Homae, Hama Watanabe, Takayuki Otobe, Tamami Nakano, Tohshin Go, Yukuo Konishi, and Gentaro Taga. Development of global cortical networks in early infancy. *Journal of Neuroscience*, 30(14):4877–4882, 2010.

[12] Shuntaro Sasai, Fumitaka Homae, Hama Watanabe, and Gentaro Taga. Frequency-specific functional connectivity in the brain during resting state revealed by nirs. *NeuroImage*, 56(1):252 – 257, 2011.

[13] S. Tak, A.M. Kempny, K.J. Friston, A.P. Leff, and W.D. Penny. Dynamic causal modelling for functional near-infrared spectroscopy. *NeuroImage*, 111(Supplement C):338 – 349, 2015.

[14] Penny W Friston KJ1, Harrison L. Dynamic causal modelling. *Neuroimage*, 19(4):1273–302, Aug 2003.

[15] S. Tak, M. Uga, G. Flandin, I. Dan, and W.D. Penny. Sensor space group analysis for fnirs data. *Journal of Neuroscience Methods*, 264(Supplement C):103 – 112, 2016.

[16] Samuel Antonio Montero-Hernandez, Felipe Orihuela-Espina, Javier Herrera-Vega, and Luis Enrique Sucar. Causal probabilistic graphical models for decoding effective connectivity in functional near infrared spectroscopy. In *FLAIRS Conference*, 2016.

[17] D. Olivier. fmri connectivity, meaning and empiricism: Comments on: Roebroeck et al. the identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *NeuroImage*, 58(2):306 – 309, 2011.

[18] G. Deshpande, K. Sathian, and X. Hu. Assessing and compensating for zero-lag correlation effects in time-lagged granger causality analysis of fmri. *IEEE Transactions on Biomedical Engineering*, 57(6):1446–1456, June 2010.

[19] K. Friston. Dynamic causal modeling and granger causality comments on: The identification of interacting networks in the brain using fmri: Model selection, causality and deconvolution. *NeuroImage*, 58(2):303 – 305, 2011.

[20] Jean Honorio, Dimitris Samaras, Irina Rish, and Guillermo Cecchi. Variable selection for gaussian graphical models. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 538–546, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[21] Melissa K. Carroll, Guillermo A. Cecchi, Irina Rish, Rahul Garg, and A. Ravishankar Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112 – 122, 2009.

[22] Irina Rish, Benjamin Thyreau, Bertrand Thirion, Marion Plaze, Marie laure Paillere-martinot, Catherine Martelli, Jean luc Martinot, Jean baptiste Poline, and Guillermo A. Cecchi. Discriminative network models of schizophrenia. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 252–260. Curran Associates, Inc., 2009.

[23] S.M. Smith, K.L. Miller, G. Salimi-Khorshidi, M. Webster, C.F. Beckmann, T.E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fmri. *NeuroImage*, 54(2):875 – 891, 2011.

[24] E. van Dellen, I E. Sommer, MM. Bohlken, P. Tewarie, L. Draaisma, A. Zalesky, M. Di Biase, JA. Brown, L. Douw, WM. Otte, R CW. Mandl, and CJ. Stam. Minimum spanning tree analysis of the human connectome. *Human Brain Mapping*, 39(6):2455–2471.

[25] C.J. Stam, E.C.W. van Straaten, E. Van Dellen, P. Tewarie, G. Gong, A. Hillebrand, J. Meier, and P. Van Mieghem. The relation between structural and functional connectivity patterns in complex brain networks. *International Journal of Psychophysiology*, 103:149 – 160, 2016. Research on Brain Oscillations and Connectivity in A New Take-Off State.

[26] C.J. Stam, P. Tewarie, E. Van Dellen, E.C.W. van Straaten, A. Hillebrand, and P. Van Mieghem. The trees and the forest: Characterization of complex brain networks with minimum spanning trees. *International Journal of Psychophysiology*, 92(3):129 – 138, 2014.

[27] P. Tewarie, E. van Dellen, A. Hillebrand, and C.J. Stam. The minimum spanning tree: An unbiased method for brain network analysis. *NeuroImage*, 104:177 – 188, 2015.

[28] A. Shmuel, E. Yacoub, D. Chaimow, N.K. Logothetis, and K. Ugurbil. Spatio-temporal point-spread function of fmri signal in human gray matter at 7 tesla. *NeuroImage*, 35(2):539552, 2007.

[29] H. Peng, C. Ding, and L. Fulmi. Feature selection based on mutual information; criteria of max- dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[30] D. Won. *Optimization and Machine Learning Frameworks for Complex Network Analysis*. PhD dissertation, University of Washington, 2016.

CHAPTER 3

An fNIRS-Based Feature Learning and Classification Framework to Distinguish Hemodynamic Patterns in Children who Stutter

An fNIRS-Based Feature Learning and Classification Framework to Distinguish Hemodynamic Patterns in Children who Stutter

Rahilsadat Hosseini, Bridget Walsh, Fenghua Tian, and Shouyi Wang

***Abstract*—Stuttering is a communication disorder that affects approximately 1 % of the population. Although 5-8 % of preschool children begin to stutter, the majority will recover with or without intervention. There is a significant gap, however, in our understanding of why many children recover from stuttering while others persist and stutter throughout their lives. Detecting neurophysiological biomarkers of stuttering persistence is a critical objective of this study. In this study, we developed a novel supervised sparse feature learning approach to discover discriminative biomarkers from functional near infrared spectroscopy (fNIRS) brain imaging data recorded during a speech production experiment from 46 children in three groups: children who stutter (n = 16), children who do not stutter (n=16), and children who recovered from stuttering (n =14). We made an extensive feature analysis of the cerebral hemodynamics from fNIRS signals and selected a small number of important discriminative features using the proposed sparse feature learning framework. The selected features are capable of differentiating neural activation patterns between children who do and do not stutter with an accuracy of 87.5 % based on a five-fold cross-validation procedure. The discovered set cerebral hemodynamics features are presented as a set of promising biomarkers to elucidate the underlying neurophysiology in children who have recovered or persisted in stuttering and to facilitate future data-driven diagnostics in these children.**

***Index Terms*—stuttering, functional near-infrared spectroscopy (fNIRS), speech production, children, data mining, feature extraction and selection, biomarkers, mutual information, sparse modeling**

I. Introduction

Stuttering is a communication disorder characterized by involuntary disruptions in the forward flow of speech. These disruptions, referred to as stuttering-like disfluencies, are recognized as repetitions of speech sounds or syllables, blocks where no sound or breath emerge, or prolongation of speech sounds. In recent years, there has been considerable progress toward understanding the origins of a historically enigmatic disorder. Past theories of stuttering attempted to isolate specific factors such as anxiety, linguistic planning deficiencies, or muscle hyperactivity as the root cause of stuttering (for review, see [1]). More recently, however, stuttering is hypothesized to be a multifactorial disorder. Atypical development of the neural circuitry underlying speech production may adversely impact the different cognitive, motor, linguistic, and emotional processes required for fluent speech production [2], [3].

The average age of stuttering onset is 33 months [4]. Although, 5-8 %, of preschool children begin to stutter, the majority (70-80 %) will recover with or without intervention [5], [4]. Given the high probability of recovery, parents often elect to postpone therapy to see if their child's stuttering resolves. However, delaying therapy in children at greater risk for persistence allows maladaptive neural motor networks to form that are challenging to treat in the future [6], [4]. The lifelong implications of stuttering are significant, impacting psychosocial development, education, and employment achievement [7], [8], [9], [10].

There is a significant gap in our understanding of why so many children recover while others persist in stuttering. Established behavioral risk factors for stuttering persistence include one or more of the following: positive family history, later age of onset (i.e. stuttering began after 36 months), time since onset, sex–boys are more likely to persist, and type and frequency of disfluencies [4]. Combining behavioral risk factors with objective, physiological biomarkers of stuttering may constitute a more powerful approach to help identify children at greater risk for chronic stuttering. Detecting such physiological biomarkers of stuttering persistence is a critical objective of our research [11], [12].

In our earlier study, Walsh et al. (2017) [13] recorded cortical activity during overt speech production from children who stutter and their fluent peers. During the experiment, the children completed a picture description task while we recorded hemodynamic responses over neural regions involved in speech production and implicated in the pathophysiology of stuttering including: inferior frontal gyrus (IFG), premotor cortex (PMC), and superior temporal gyrus (STG) with functional near-infrared spectroscopy (fNIRS), which is a safe, non-invasive optical neuroimaging technology that relies upon neurovascular coupling to indirectly measure brain activity. This is accomplished using near-infrared light to measure the relative changes in both oxygenated (Oxy-Hb) and deoxygenated hemoglobin (Deoxy-Hb), two absorbing chromophores in cerebral capillary blood [14]. fNIRS offers significant advantages including its relatively low cost and greater tolerance for movement, making it a more child-friendly neuroimaging approach. fNIRS has been used to assess the regional activation, timing, and lateralization of cortical activation for a diverse number of perceptual, language, motor, and cognitive investigations (for review, [15]).

Using fNIRS to assess cortical activation during overt speech production, we found markedly different speech-evoked hemodynamic responses between the two groups of children during fluent speech production [13]. Whereas controls showed clear activation over left dorsal IFG and left

PMC, characterized by increases in Oxy-Hb and decreases in Deoxy-Hb, the children who stutter demonstrated deactivation, or the reverse response over these left hemisphere regions. The distinctions in hemodynamic patterns between the groups may indicate dysfunctional organization of speech planning and production processes associated with stuttering and could represent potential biomarkers of stuttering.

Although different brain signal patterns can be observed for stuttering and control group in our previous studies, there is still a lack of reliable quantitative tools to evaluate stuttering treatment and recovery process based on brain activity patterns. In our previous studies, we have extensive research efforts on specialized machine learning (ML) and pattern recognition techniques for multivariate spatiotemporal brain activity pattern identification under different brain states [16], [17], [18], [19]. In this study, we aimed to detect neurophysiological biomarkers of stuttering using advanced ML techniques. In particular, we performed ML models for two experiments. In experiment (1), we made an extensive feature extraction from fNIRS brain imaging data of 16 children who stutter and 16 children in a control group collected in our previous study [13]. Next, we developed a novel supervised sparse feature learning approach to discover a set of discriminative biomarkers from a large set of fNIRS features, and construct a classification model to differentiate hemodynamic patterns from children who do and do not stutter. In experiment (2), we applied the constructed classification model on a novel test set of fNIRS data collected from a group of children who had recovered from stuttering and underwent the same picture description experiment. Using the novel test set with children's data that was not used to develop the initial algorithms allowed us to assess the model generalization with the discovered biomarkers from experiments (1) to (2). We elected to include children who had recovered from stuttering in the test group for theoretical and clinical bearings. Young children who begin to stutter are far more likely to recover than persist. It is important to assess the underlying neurophysiology of different stuttering phenotypes to learn, for example, whether recovered children's hemodynamic patterns would classify them with the group of controls or with the group of stuttering children. These proof-of-concept experiments represent a critical step toward identifying greater risk for persistence in younger children near the onset of stuttering.

The remainder of the paper is organized as follows: In Section 2, we present the methodology, including participant and data collection details, fNIRS data feature extraction and structured sparse feature selection models. In Section 3, we present the results of the pattern discovery of biomarkers as well as performance consistency on the novel test-set of data from recovered children. In section 4, we discuss the selected features and their interpretations in terms of brain regions of interest. Finally, we conclude the study in section 5.

II. Method

A. Participants, fNIRS Data Collection & Pre-processing

In experiment (1), fNIRS data from the 32 children who participated in the Walsh et al. (2017) study [13] was analyzed; 16 children who stutter (13 males) and 16 age- and socioeconomic status-matched controls (11 males). The participants were between the ages of 7-11 years (M = 9 years). Stuttering diagnosis and exclusionary criteria are provided in [13].

In experiment (2), a group of 14 children (10 males) between the ages of 8-16 years (M = 12 years) who recovered from stuttering was analyzed as an additional test group. All of the children completed a picture description experiment in which they described aloud different picture scenes (talk trials) that randomly alternated with null trials in which they watched a fixation point on the monitor. In order to compare hemodynamic responses among the groups of children, only fluent speech trials were considered in the analyses.

For each experiment, we recorded hemodynamic responses with a continuous wave system (CW6; TechEn, Inc.) that uses near-infrared lasers at 690 and 830 nm as light sources, and avalanche photodiodes (APDs) as detectors for measuring intensity changes in the diffused light at a 25-Hz sampling rate. Each source/detector pair is referred to as a channel. The fNIRS system acquired signals from 18 channels (9 over the left hemisphere and 9 over homologous right hemisphere regions) that were placed over ROIs relying on 10-20 system coordinates Figure (1).

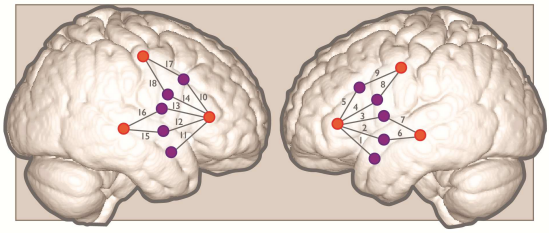


Fig. 1: Approximate positions of emitters (orange circles) and detectors (purple circles) are shown on a standard brain atlas (ICBM 152). The probes were placed symmetrically over the left and right hemisphere, with channels 1-5 spanning inferior frontal gyrus, channels 6-7 over superior temporal gyrus, and channels 8-9 over precentral gyrus/premotor cortex.

Data analysis is detailed in Walsh et al. [13]. Briefly, the fNIRS data was preprocessed using Homer2 software [20]. Usable channels of raw data were low-pass filtered at 0.5 Hz and high-pass filtered at 0.03 Hz. Concentration changes in Oxy-Hb and Deoxy-Hb were then calculated and a correlation-based signal improvement approach applied to the concentration data to reduce motion artifacts [21]. Finally, we derived each child's Oxy-Hb and Deoxy-Hb event-related hemodynamic responses from all channels from stimulus onset to the end of the trial. We then subtracted the average hemodynamic response associated with the null trials from the average hemodynamic response from the talk trials to derive a differential hemodynamic response for each channel [22]. The average Oxy-Hb and Deoxy-Hb hemodynamic response averaged over all 18 channels is plotted as a function of time for each child in Figure (2) and (3).

B. Feature Extraction

As shown in Figure (4), each experimental trial was partitioned into three phases: perception or the see-phase (0-2s, the

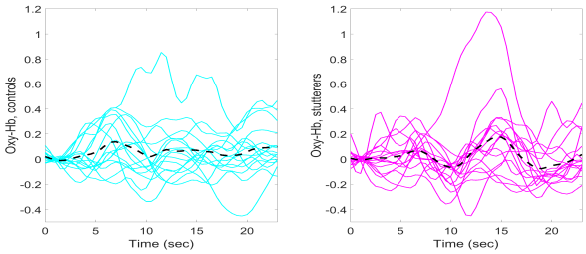


Fig. 2: Oxy-Hb hemodynamic responses averaged over all 18 channels for each subject. Controls are plotted on the left (cyan curves) and stutterers on the right (magenta curves). The grand average hemodynamic response across all channels and subjects is represented by the black dashed curve.

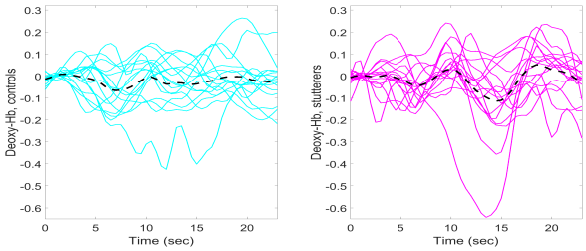


Fig. 3: Deoxy-Hb hemodynamic responses averaged over all 18 channels for each subject. Controls are plotted on the left (cyan curves) and children who stutter on the right (magenta curves). The grand average hemodynamic response across all channels and subjects is represented by the black dashed curve.

children saw a picture on the monitor), the talk-phase (3-8s, the children described aloud the picture), and the recovery-phase (9-23s, the hemodynamic response returned to baseline) for measurements of Oxy-Hb and Deoxy-Hb. We extracted 21 features from each channel; $21 = 4 + 3 + 3 + 1 + (5 \times 2(\text{for 1 and 2 sec of delay}))$. These delays were implemented to account for correlation of the signal to its lagged values. The names of the feature group and subgroups are shown in Figure (4). Therefore, for each subject with 18 channels of fNIRS data, there were 378 extracted features from Oxy-Hb and Deoxy-hb measurements in each phase.

The extracted groups of features are summarized in the following.

- Statistical features capture descriptive information of the signals.
- Morphological features comprised the number of peaks and zero crossings and measures of curve length.
- Hjorth parameters capture signal variation over time expressed as activity, mobility, and complexity.. The three features are defined as: $activity = Var(y(t))$, $mobility = \sqrt{\frac{var(y(t)dy/dt}{var(y(t))}}$, $complexity = \frac{mobility(y(t)dy/dt)}{mobility(y(t))}$.
- Normalized Area Under the Signal (NAUS) calculates the sum of values which have been subtracted from a defined baseline divided by the sum of the absolute values for the fNIRS signal.
- Autocorrelation captured the linear relationship of the signal with its historical values considering 1 and 2 s delays Kendall, partial, Spearman and Pearson are four ways to compute autocorrelation.
- Bicorrelation computes the bicorrelation on the time series X_v for given delays in τ_v. Bicorrelation is an extension of the autocorrelation to the third order moments, where the two delays are selected so that the second delay is twice the original , (i.e. $x(t)x(t-\tau)x(t-2\tau)$). Given a delay of τ and the standardized time series X_v with length n, denoted as Y_v, the $bicorr(\tau)$ can be calculated as:

$$\frac{\sum_{j=1}^{n-2\tau} Y_v(j)Y_v(\tau+j)Y_v(2\tau+j)}{n-(2\times\tau)} \quad (1)$$

1) Personalized Feature Normalization: As illustrated in Figures (2) and (3) fNIRS signals vary dynamically across subjects, imposing a challenge to biomedical research. Because of inter-individual variability in signal features, it is difficult to build a robust diagnostic model to accurately discriminate between groups of participants. Outliers can further distort the trained model, thus impeding generalization. To tackle these issues, we applied a personalized feature normalization approach to standardize the extracted feature values of each subject onto the same scale to enhance feature interpretability across subjects.

To accomplish this, we calculated the upper and lower limits for each extracted feature using the formula V_l= max(minimum feature value, lower quartile + 1.5 $\times$ interquartile range) for the lower limit, and V_u= min(maximum feature value, upper quartile + 1.5 $\times$ interquartile range) for the upper limit. Feature values outside of this defined interval were considered to be outliers and mapped to 0 or 1. More details can be found in study [23]. Assuming the raw feature value was F_{raw}, the scaled feature value F_{scaled} was obtained by:

$$F_{scaled} = \frac{F_{raw} - V_l}{V_u - V_l}. \quad (2)$$

C. Integrated Structured Sparse Feature Selection using Mutual Information

Feature selection techniques are widely used to improve model performance and promote generalization in order to gain a deeper insight into the underlying processes or problem. This is accomplished by identifying the most important decision variables, while avoiding overfitting a model. Most feature selection techniques classify into three categories: embedded methods, wrapper methods, and filter methods [24]. Both embedded and wrapper methods seek to optimize the performance of a classifier or model. Thus, the feature selection performance is highly limited to the embedded classification models. Filter feature selection techniques assess the relevance of features by measuring their intrinsic properties. Widely used models include correlation-based feature selection [25], fast correlation-based feature selection [26], minimum redundancy maximum relevance (mRMR) [27] and information-theoretic-based feature selection methods [28].

Sparse modeling-based feature selection methods have gained attention owed to their well-grounded mathematical

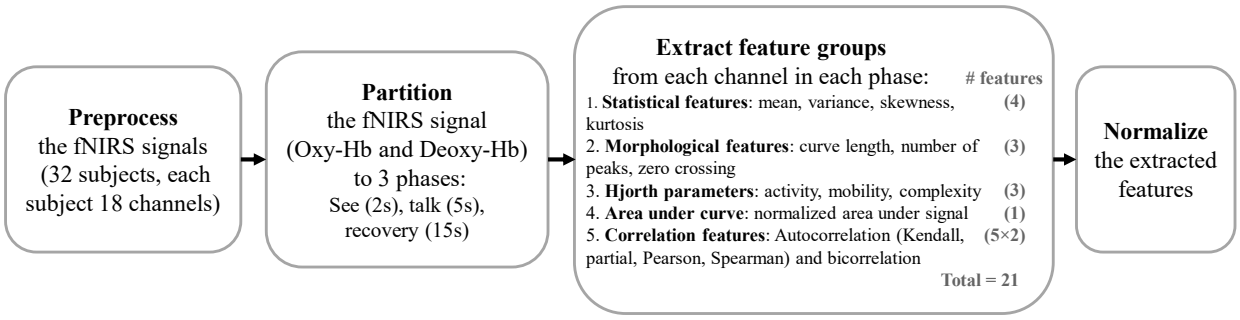


Fig. 4: The process of feature engineering: pre-process input data, features extraction, post-process the features

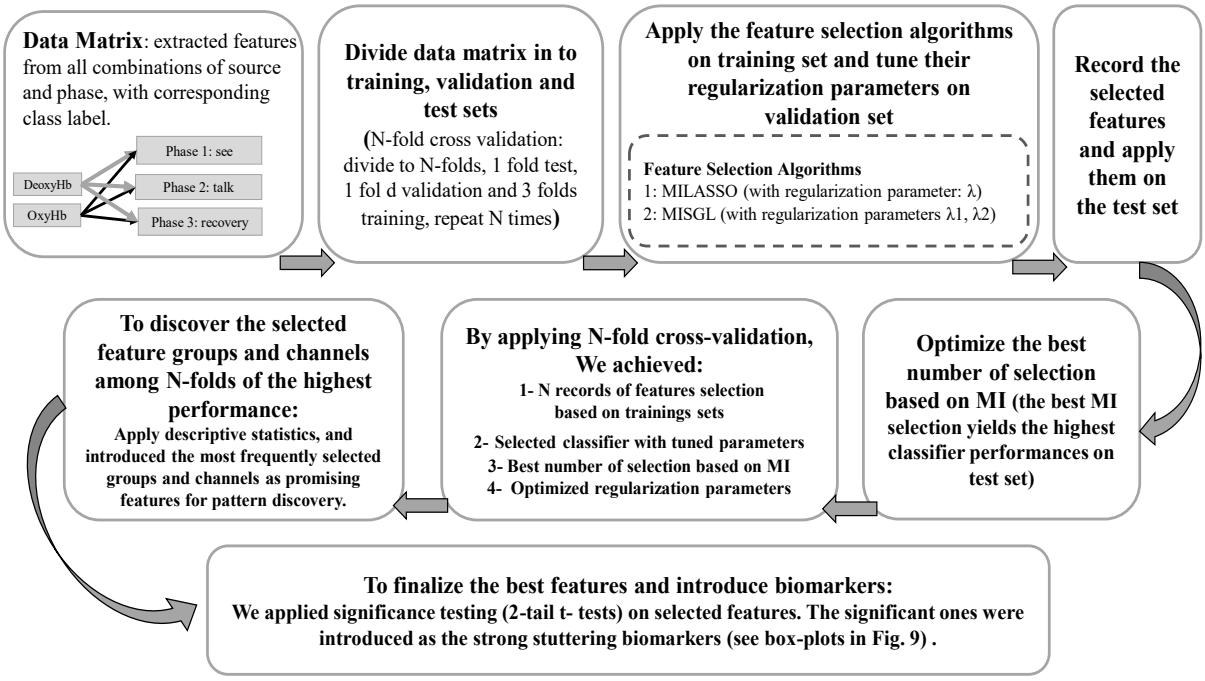


Fig. 5: Feature selection and tuning the regularization parameters via N-fold cross-validation in order to introduce the promising features (biomarkers).

theories and optimization analysis. These feature selection algorithms employ sparsity-inducing regularization techniques, such as L_1-norm constraint or sparse-inducing penalty terms, for variable selection. To construct more interpretable models, structured sparse modeling algorithms that consider feature structures have recently been proposed and show promising results in many practical applications including brain informatics, gene expression, medical imaging analysis, etc. [29], [30], [31], [32]. However, most of the current structured sparse modeling algorithms only consider linear relationships between response variables and predictor variables (features) in the analysis and may miss complex nonlinear relationships between features and response variables that may be present. On the other hand, although some filter or wrapper methods have the capability to capture nonlinear relationships between features and response variables, feature structures may not be optimally identified in the feature selection procedure. Constructing interpretable learning models with efficient feature selection remains an open and active research area in the machine learning community. Zhongxin et al. [33] proposed a feature selection algorithm based on mutual information (MI) and least absolute shrinkage and selection operator (LASSO) using L_1 regularization with application to microarray data produced by gene expression. In our previous study, we also proposed a MI-based sparse feature selection model for EEG feature selection and applied it to epilepsy diagnosis [34]. However, feature structures were not considered during feature selection in both [33] and [34].

To consider both linear and nonlinear relationships between features and response variables, while acknowledging feature structures in feature selection, we propose a novel feature selection framework that integrates information theory-based feature filtering and structured sparse learning models to effectively capture feature dependencies and identify the most informative feature subset. There are two differences with respect to earlier studies [33] and [34]: (1) we did not use regularization techniques like LASSO as the second rank filtering; rather, we used sparse-inducing regularization to reveal the second-level feature-response relationships; (2) we applied structured feature learning by penalizing the

feature groups. We implemented the proposed information-theory-based structured sparse learning framework to identify the optimal feature subset as discriminant neurophysiological biomarkers of stuttering.

1) Mutual Information for Feature Selection: MI is an index of mutual dependency between two random variables that quantifies the amount of information obtained about one random variable from the other random variable [35]. MI effectively captures nonlinear dependency among random variables and can be applied to rank features in feature selection problems [27]. The fundamental objective of MI-based filtering methods is to retain the most informative features (i.e., with higher MI) while removing the redundant or less-relevant features (i.e., with low MI). The mutual information of two random variables X and Y, denoted by $I(X,Y)$, is determined by the probabilistic density functions p(x), p(y), and $p(X,Y)$:

$$I(X;\,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right), \qquad (3)$$

2) Structured Sparse Feature Selection: A sparse model generates a sparse (or parsimonious) solution using the smallest number of variables with non-zero weights among all the variables in the model. One basic sparse model is LASSO regression, which employs L_1 penalty-based regularization techniques for feature selection [36]. The LASSO objective function is formulated as follows:

$$\min\left\{\|\mathbf{Ax}-\mathbf{Y}\| + \lambda_1\|\mathbf{x}\|_1\right\}, \qquad (4)$$

where A is the feature matrix, Y is the response variable, λ_1 is a regularization parameter and x is the weight vector to be estimated. The L_1 regularization term produces sparse solutions such that only a few values in the vector x are non-zero. The corresponding variables with non-zero weights are the selected features to predict the response variable Y.

Structured Features (Sparse Group LASSO (SGL)) The basic LASSO model, and many L_1 regularized models, assume that features are independent and overlook the feature structures. However, in most practical applications, features contain intrinsic structural information, (e.g., disjoint groups, overlapping groups, tree-structured groups, and graph networks) [32]. The feature structures can be incorporated into models to help identify the most critical features and enhance model performance.

As outlined in section 2.2, the features we extracted from the raw fNIRS data are disparate; thus they can be categorized into disjoint groups. The sparse group LASSO regularization algorithm promotes sparsity at both the within- and between-group levels and is formulated as:

$$\min\left\{\|\mathbf{Ax}-\mathbf{Y}\| + \lambda_1\|\mathbf{x}\|_1 + \lambda_2\sum_{i=1}^{g}\omega_i^g\|\mathbf{x}_{G_i}\|_2\right\} \qquad (5)$$
$$A\in\mathbf{R}^{m\times n},\quad y\in\mathbf{R}^{m\times 1},\quad x\in\mathbf{R}^{n\times 1},$$

where the weight vector $\mathbf{x}$ is divided by g non-overlapping groups: $\{x_{G_1}, x_{G_2}, \ldots, x_{G_g}\}$, and $\omega_i^g$is the weight i for group g. The parameter λ_1 is the penalty for sparse feature selection, and the parameter λ_2 is the penalty for sparse group selection (i.e. the weights of some feature groups will be all zeroes). In cases where feature groups overlap, the sparse overlapping group LASSO regularization can be used [37].

3) Integrated MI-Sparse Feature Selection Framework: The objective of our approach is to consider structured feature dependency while keeping the search process computationally efficient. To accomplish this, we employed the MI-guided feature selection framework outlined in Algorithm (1). Given a number of features k, the subset of top k features ranked by MI is denoted by S, and the subset of the remaining features is denoted by W. From S, the optimal feature subset is selected by exploring the k_1 high-MI features which includes the iterative process of removal of highly-correlated features with 0.96 threshold. From W, the k_2 sparse-model selected low-MI features. The final selected features subset is the set of $(k_1 + k_2)$ features which are evaluated based on the cross-validation classification performance. Enumeration of k_1 starts from 1 and ascends until reaching the stopping criteria (i.e., when the cross-validation accuracy converges and cannot be further improved). MISS Algorithm (1) can be applied in two ways: (1) without group structure, which is a combination of mutual information and LASSO namely (MILASSO), (2) with group structure, which is a combination of mutual information and SGL namely (MISGL).

Algorithm 1 Mutual Information Sparse Feature Selection (MISS)

```
Rank all features based on mutual information
repeat
    k1 = k1 + 1
    repeat
        Divide sorted features to high-MI and low-MI
        S ←high-MI
        Remove redundant features from S
    until k1 features remain after reduction
    W ←low-MI
    Apply sparsity learning to W
    k2 ← number of selected features by SGL or LASSO
    Build classifier model with k1 + k2 selected features
until classifier performance converges
```

D. Machine Learning Algorithm Selection & Evaluation

We applied established ML algorithms [38] (i.e., support vector machine (SVM), k-nearest neighbor (kNN), decision tree, ensemble, and linear discriminant) to assess whether cerebral hemodynamic features could accurately differentiate the group of children who stutter from controls. An overview of the steps involved in feature extraction and model evaluation is provided in Figure (6).

1) Support vector machines: SVM is considered to be a popular and promising approach among classification studies [39]. It has been used in a variety of biomedical applications; for example, to detect patterns in gene sequences or to classify patients according to their genetic profiles, with EEG signals in brain-computer interface systems, and to discriminate hemodynamic responses during visuomotor tasks [40], [17], [41], [42].

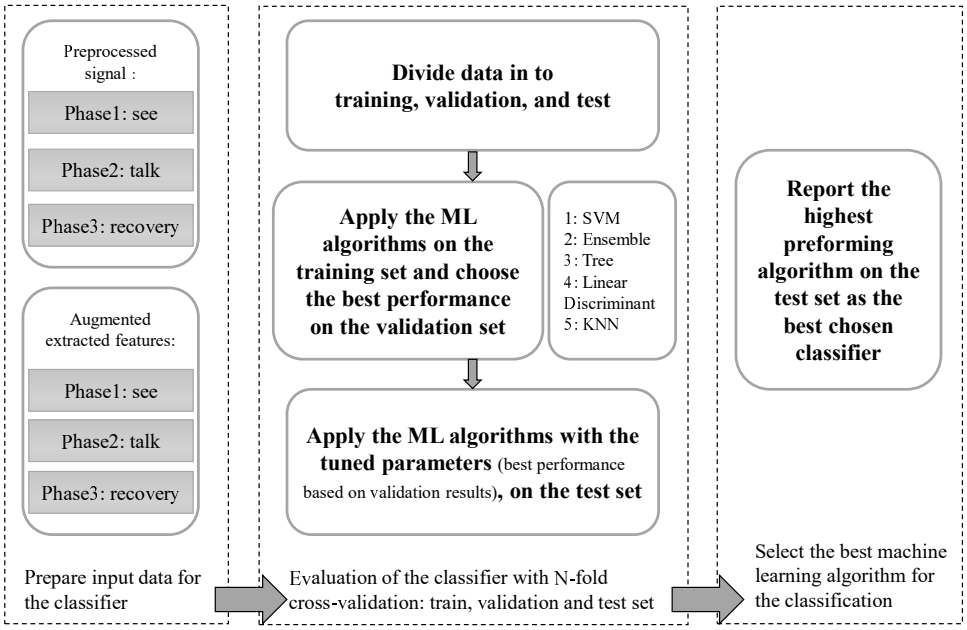


Fig. 6: The process of choosing the most accurate ML classification algorithm with N-fold cross-validation and parameter tuning

In this study we applied Gaussian radial basis function (RBF) as the kernel which maps input data x to higher dimensional space.

2) Bayesian Parameter Optimization: Parameters in each classifier significantly affect its performance. We applied Bayesian optimization, part of the Statistics and Machine Learning Toolbox in Matlab, to optimize hyper-parameters of classification algorithm [43]. By applying Bayesian optimization algorithm, we want to minimize a scalar objective function $(f(x) =$ cross-validation classification loss$)$ for the classifier parameters in a bounded domain.

3) N-fold Cross-Validation: We applied N-fold cross-validation (N=5) for training and testing. First, we selected the features and optimized the parameters of the classification algorithm on the training set then applied the tuned model on the testing set, see Figure (6). Accuracy is defined as the ratio of correctly classified test subjects to the total number of subjects. Sensitivity is the ratio of children in the stuttering group correctly identified as stuttering to all of the children in the stuttering group. Specificity is the ratio of children correctly identified as controls to the total number of children in the control group. In this study, we used the average sensitivity and specificity values to measure binary classification accuracy for each ML model.

III. Results

Classifier performance is reported for experiment (1) based on the outcome of the N-fold cross-validation procedure on the test-set, see Table (I). For experiment (2) classification performance was established on a novel test-set of 14 children who had recovered from stuttering, see Table (IV).

A. Experiment (1): Choosing the best ML Algorithm

The most accurate ML algorithm on the raw fNIRS data was the tree classifier with 77.5 % accuracy. The highest accuracy obtained after feature extraction and application of feature selection (MILASSO) was SVM (with RBF kernel) that achieved 87.5 % accuracy, Table (I). The phase of the fNIRS trial that distinguished the groups of children was the talk interval and the source was Oxy-Hb. However in some cases performance using features derived from Deoxy-Hb measurements reached comparable accuracy as those from Oxy-Hb.

B. Experiment (1): Comparing Feature Selection Algorithms

In Table (II), we compared the performance of the proposed feature selection algorithm (MISS) with the popular existing MI-based method like mRMR and linear regularized methods like LASSO and SGL. MISS approach outperformed mRMR in feature selection by yielding higher SVM classification performance with the same number of selected features, (14 and 11 for measurement source of deoxy-Hb and oxy-Hb), approximately 7.5 and 27.5 % respectively. MISS approach outperformed LASSO and SGL in feature selection yielding higher classification accuracy approximately 2.5 to 12.5 %.

C. Experiment (1): Selected Features

From an extended set of features, a subset that provided the highest classification accuracy was identified by MISGL and MILASSO in the SVM(RBF) model. This subset of features, shown in Figure (7), comprises statistical, NAUS, Hjorth parameters, autocorrelation and bicorrelation features. Channels that provided the highest discriminative power to differentiate between children who stutter and controls were localized to the left hemisphere; specifically, channels 1, 4, and 5 over left IFG.

TABLE I: Comparison among performance of various ML classifiers (before & after) feature extraction and application of feature selection

Input data: fNIRS signal, from phase : Talk							
		Source: Oxy-Hb			Source: Deoxy-Hb		
	Classifier	avg	sen	spe	avg	sen	spe
	SVM	0.75	0.75	0.75	0.725	0.6	0.85
	Ensemble	0.7	0.65	0.75	0.65	0.7	0.6
	KNN	0.7	0.7	0.7	0.675	0.5	0.85
	L Discr	0.75	0.75	0.75	0.725	0.6	0.85
	Tree	0.775	0.8	0.75	0.575	0.8	0.35

Input data: extracted features from signal in phase (Talk) with application of MISS for feature selection										
	Source: Oxy-Hb					Source: Deoxy-Hb				
Classifier	MI num	Tot num	avg	sen	spe	MI num	Tot num	avg	sen	spe
SVM	**2**	**11**	**0.875**	**0.85**	**0.9**	10	14	0.825	0.8	0.85
Ensemble	2	11	0.825	0.85	0.8	9	32	0.85	0.8	0.9
KNN	2	11	0.825	0.75	0.9	3	7	0.85	0.85	0.85
L Discr	1	10	0.825	0.8	0.85	10	33	0.775	0.7	0.85
Tree	3	13	0.675	0.8	0.55	6	29	0.75	0.8	0.7

sen: sensitivity, spe: specificity , avg: average of sen and spe, L Discr: linear discriminant
MI num: number of selected features based on MI
Tot num: total number of selection (based on MI and based on SGL or LASSO)

TABLE II: Comparison among performance of various feature selection algorithms via SVM classification accuracy on the selected features with each approach

Feature selection Method	Deoxy-Hb		Oxy-Hb	
	Tot num feat	Avg(sen, spe)	Tot num feat	Avg(sen, spe)
mRMR	14	75	11	60
LASSO	$\sim$ 6.4 $*$	80	$\sim$ 4.8 $*$	75
SGL	$\sim$ 6.4 $*$	77.5	$\sim$ 7 $*$	78
MISS(MILASSO, MISGL)	14	82.5	11	87.5

Tot num feat: total number of selected features
$*$ indicates the average number of selected features among N-fold for LASSO and SGL methods
Avg(sen, spe)= average of sensitivity and specificity (%)

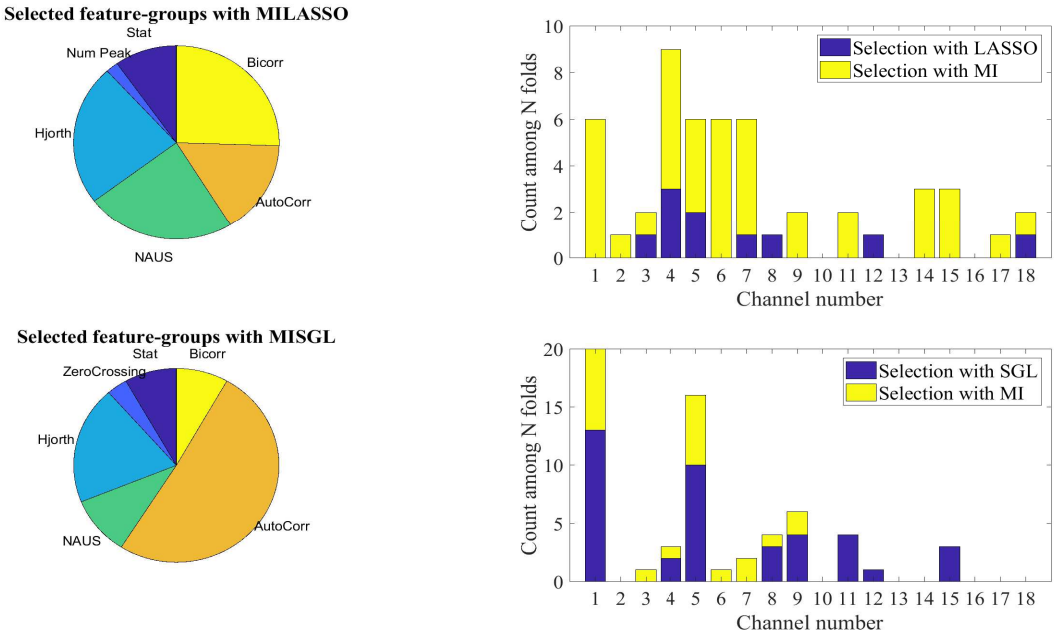


Fig. 7: Statistical summary of the selected feature groups and channels with MILASSO and MISGL in N-fold cross validation. In each fold, there was 11 to 14 selected features, from different channels and feature-groups. The pie charts illustrate the group that selected features most frequently came from. The histograms summarize the channel selection with MISGL and MILASSO. For example, from approximately 60 total features selected from 5 folds, 6 features were selected from channel 1, and 9 features from channel 4 (either based on MI ranking (yellow bar) or LASSO coefficients (blue bar) which are stacked for each channel).

The top 14 features from the entire feature set are listed in Table (III). These features, (2 based on MI and 12 based on LASSO), were extracted from the talk-phase with source Oxy-Hb. We performed 2-tailed t-tests on these features. *p*-values ≤ 0.05, confirm a significant statistical difference between children who stutter and controls for a given feature.

1) Feature Selection Optimization: The number of features selected by MILASSO or MISGL affects the performance of the classifier; a more sparse selection enhances model performance, promotes generalization, and facilitates the interpretation of results. During the enumeration process for MI selection, we learned that with less than 10 MI features (total features $\leq 15-22$), the average classifier performance was approximately 80 %; with 15 to 30 MI features ($25-35 \leq$ total features ≤ 40) , performance was approximately 75 %; with more than 30 MI features, (total features ≥ 42), the accuracy decreased to 70 %. The highest accuracy with the least number of features came from 11 total features with the MILASSO approach, 2 MI and 9 LASSO and 12 total features with the MISGL approach, 8 MI and 4 SGL.

2) Biomarkers: The features in Table (III) that showed significant differences between children who stutter and controls are recognized as biomarkers. Box-plots of these features for the children who stutter and controls are plotted on a common scale in Figure (8). The discriminative features we detected in Figure (8) comprised significantly lower values of NAUS and slightly higher values of Hjorth mobility and bicorrelation with 2 sec of delay for children who stutter compared to controls.

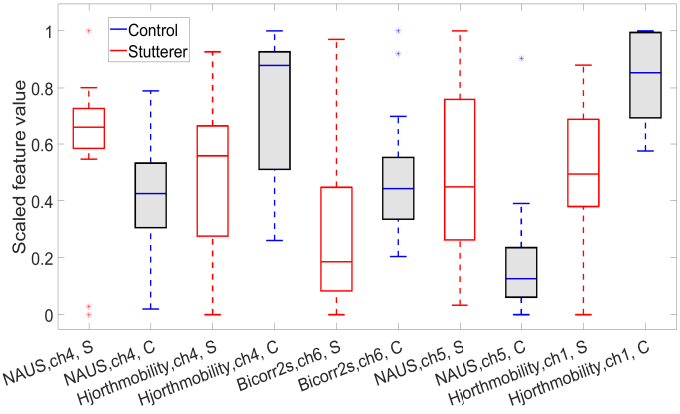


Fig. 8: Box-plot of top 5 significant features from talk-phase and source Oxy-Hb, ch: channel, (S: stutterer , C: control).

D. Experiment (2): Stuttering Recovery Assessment with Selected Features

In this section we report the performance of the classifier on the additional test-set (data from 14 children who recovered from stuttering), shown in Table (IV). We applied the best-performing algorithm based on the results from experiment (1): SVM with tuned parameters sigma = 1 and penalty = 0.001 on the entire dataset. We documented that 71.43 %, or approximately 10 out of 14 children who had recovered from stuttering, classified into the control group based on features derived from fNIRS signals derived from the talk-phase of the experiment. The same degree of stuttering recovery assessment (SRA) was achieved with both Oxy-Hb and Deoxy-Hb sources Table (IV).

IV. Discussion

In experiment (1), we applied structured sparse feature learning models to previously collected speech-evoked fNIRS data from Walsh et al. [13] to explore whether neurophysiological biomarkers could accurately classify hemodynamic patterns from children who do and do not stutter. Following feature extraction and feature selection with MISS, the SVM achieved the highest classification accuracy of 87.5 %. With this model, classification performance was improved by 10 % using feature extraction and sparse MI-based features selection. This degree of accuracy was reached using features extracted during the talk interval of the trial from the source, Oxy-Hb (although features extracted from Deoxy-Hb reached comparable accuracy). A feature set comprising statistics, NAUS, Hjorth parameters, autocorrelation and bicorrelation features provided the highest discriminative power. Notably, nearly all of these features were extracted from channels localized to the left hemisphere (i.e. channels 1-9). The selected features may not be significant individually as shown in Table (III), thus they can be ignored or missed in basic statistical analyses used by many feature selection algorithms. The MISS approach is valuable to reveal clear discriminative patterns among features in a higher dimensional space, and to discover relevant multivariate biomarkers.

Features from channels 1, 4 and 5, which span left IFG, were identified as neurophysiological biomarkers that distinguished hemodynamic characteristics of children who stutter from controls. These included significantly reduced NAUS in left IFG channels 4 and 5 and increased Hjorth mobility parameters, denoting increased variability, in left IFG channels 1 and 4 in children who stutter.

In our earlier study [13], we found significantly reduced Oxy-Hb and increased Deoxy-Hb concentrations during the talk interval in channels over left IFG in the group of children who stutter. The left IFG comprising Broca's area is integral to speech production and may develop atypically in children who stutter. Neuroanatomical studies reveal aberrant developmental trajectories of white and gray matter of left IFG in children who stutter compared to controls [44], [45]. Moreover, there is evidence of reduced activation of IFG/Broca's area during speech production from fMRI studies with adults who stutter [46], [47]. In our earlier study [13], we hypothesized that this finding may represent a shift in blood flow to regions outside of our recording area to compensate for functional deficits in left IFG. An alternative possibility is a disruption in cortical-subcortical loops resulting in a net inhibition of this region. This is the first study to elucidate group-level differences by classifying individual children as either stuttering or not stuttering using features derived from their speech-evoked brain hemodynamics. Based on the sensitivity index from the final model, three children who stutter classified as controls (i.e., false negatives). Interestingly, two of these three children were considered to be mild stutterers when they participated and

TABLE III: Top 14 features selected with MISS along with *p*-value (0.05 threshold for statistically significant *t*-test). With top 11 features, 87.5 % accuracy was achieved in N-fold cross-validation

Feature rank	Feature name	*p*-value	Feature rank	Feature name	*p*-value
1	NAUS, ch 4	**0.0001**	8	Hjorth mobility, ch 1	**0.0014**
2	Hjorth mobility, ch 4	**0.0022**	9	NAUS, ch 8	0.1095
3	Hjorth activity, ch 1	0.2800	10	AC partial 2s, ch 14	0.1745
4	Bicorrelation 2s, ch 6	**0.0225**	11	AC Spearman 1s, ch 6	0.9238
5	NAUS, ch 5	**0.0003**	12	Hjorth activity, ch 4	0.1792
6	Variance, ch 9	0.5319	13	Variance, ch 4	0.0605
7	Bicorrelation 1s, ch 14	0.6252	14	AC Spearman 1s, ch 7	0.6277

AC: autocorrelation, ZC: zero crossing, CL: curve length
NAUS: normalized area under signal, ch: channel
1 or 2s: 1 or 2 second of delay

TABLE IV: The best SVM performance on the additional test-set (recovered samples)

phase	source	Fsel M	Tot num	SRA
talk	Deoxy-Hb	MILASSO	14	71.43
talk	Oxy-Hb	MISGL	11	71.43

Fsel M: feature selection method, SRA: stuttering recovery assessment
Tot num: total number of selected features with MISS

have since recovered from stuttering (determined via a follow-up visit or through parental report). It is tempting to speculate that the recovery process had already begun for these children when we recorded their hemodynamic responses during the initial study. However, longitudinal studies in younger children (i.e., near the onset of stuttering) are necessary to track the developmental trajectories of their hemodynamic responses as they either recover from or persist in stuttering to empirically assess this assumption.

Finally, we compared the consistency of the best-performing SVM classifier using N-fold cross-validation from experiment (1) with results achieved using the SVM classifier on a novel test-set of data from 14 children who had recovered from episodes of early childhood stuttering in experiment (2). We found that the majority of the recovered children, or 71.43 %, classified as controls, rather than children who stutter. This suggests that left-hemisphere stuttering biomarkers that dissociated stuttering children's speech-evoked hemodynamic patterns from controls, may indicate chronic stuttering, while recovery from stuttering in many of these children was associated with hemodynamic responses similar to those from children who never stuttered. Stuttering recovery may thus be supported, in part, by functional reorganization of regions such as left IFG that corrects anomalous brain activity patterns. Although this speculation warrants further study and replication, an fMRI study with adults who recovered from stuttering identified the left IFG as a pivotal region associated with optimal stuttering recovery [48].

A final point to consider is that although most of the recovered children had hemodynamic patterns similar to controls, four of these children classified into the stuttering group. Given that stuttering is highly heterogeneous, with multiple factors implicated in the onset and chronicity of the disorder [2], it is not surprising to find evidence suggesting that recovery processes may be different for some children. More research is clearly needed to substantiate the neural reorganization that accompanies both spontaneous and therapy-assisted recovery from stuttering.

V. Conclusion

In this final section, we present several suggestions regarding data preprocessing, feature selection and ML training and evaluation to guide future investigations in this line of research.

First, the personalized feature scaling approach facilitated the discovery of discriminative patterns by removing data outliers and reducing the variability in each feature. This was a critical step in our approach to address inherent inter-individual differences in the physiological signals.

Second, the MISS approach yielded a final feature space that was both parsimonious and interpretable. In particular, MISGL, that considers feature group structures in sparse feature learning, and achieved the best classification performance with the least number of selected features. We compared our result from the MISS approach with commonly used feature selection techniques in Table (II), and the results proved that MISS outperformed the methods which solely applied either MI or regularized linear regression significantly. More importantly, MISS pinpointed specific left hemisphere channels that classified children as stuttering/nonstuttering with higher accuracy and corroborated findings from our earlier experiment [13].

In summary, the proposed MI-based structured sparse feature learning method demonstrates its effectiveness to discover the most discriminative features in a high dimensional feature space with a limited number of training samples, a common challenge for health care and medical data mining approaches. Compared to other methods, the proposed MISS approach offers a promising, interpretable solution to facilitate data-driven advances in clinical and experimental research applications.

Acknowledgment

Dr. Walsh was supported by NIH grant (NIH/NIDCD R03 DC013402) [13]

References

[1] O. Bloodstein and N. B. Ratner. *A handbook on stuttering*. Cengage Learning, Clifton Park, NY, 6 edition edition, oct 2008.
[2] A. Smith and C. Weber. How Stuttering Develops: The Multifactorial Dynamic Pathways Theory. *Journal of Speech, Language, and Hearing Research*, pages 1–23, August 2017.
[3] A. Smith. Stuttering: a unified approach to a multifactorial, dynamic disorder. In *Stuttering research and practice: bridging the gap*. Psychology Press, 1999.

[4] E. Yairi and N.G. Ambrose. *Early childhood stuttering for clinicians by clinicians*. Pro ed, Austin, Tex, 1 edition edition, November 2005.

[5] H. Mnsson. Childhood stuttering. *Journal of Fluency Disorders*, 25(1):47–57, March 2000.

[6] B. Guitar. *Stuttering: An integrated approach to its nature and treatment*. LWW, Baltimore, third edition edition, October 2006.

[7] E. Blumgart, Y. Tran, and A. Craig. Social anxiety disorder in adults who stutter. *Depression and Anxiety*, 27(7):687–692, July 2010.

[8] J. F. Klein and S. B. Hood. The impact of stuttering on employment opportunities and job performance. *Journal of Fluency Disorders*, 29(4):255–273, January 2004.

[9] S. OBrian, M. Jones, A. Packman, R. Menzies, and M. Onslow. Stuttering severity and educational attainment. *Journal of Fluency Disorders*, 36(2):86–92, June 2011.

[10] L. Iverach and R. M. Rapee. Social anxiety disorder and stuttering: Current status and future directions. *Journal of Fluency Disorders*, 40:69 – 82, 2014. Anxiety and stuttering.

[11] E. Usler, A. Smith, and C. Weber. A lag in speech motor coordination during sentence production is associated with stuttering persistence in young children. *Journal of Speech, Language, and Hearing*, 60(1):51 – 61, 2017.

[12] R. Mohan and WeberFox C. Neural systems mediating the processing of sound units of language distinguish recovery versus persistence in stuttering. *Journal of Neurodevelopmental Disorders*, 7(1), 2015.

[13] B. Walsh, F. Tian, J. A. Tourville, M. A. Ycel, T. Kuczek, and A. J. Bostian. Hemodynamics of speech production: An fNIRS investigation of children who stutter. *Scientific Reports*, 7, June 2017.

[14] A. Villringer and B. Chance. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neurosciences*, 20(10):435–442, October 1997.

[15] F. Homae. A brain of two halves: insights into interhemispheric organization provided by near-infrared spectroscopy. *NeuroImage*, 85:354–362, January 2014.

[16] W. A. Chaovalitwongse, R. S. Pottenger, S. Wang, Y. Fan, and L.D. Iasemidis. Pattern-and network-based classification techniques for multichannel medical data signals to improve brain diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(5):977–988, 2011.

[17] S. Wang, Y. Zhang, C. Wu, F. Darvas, and W.A. Chaovalitwongse. Online prediction of driver distraction based on brain activity patterns. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):136–150, 2015.

[18] K. Kam, J. Schaeffer, S. Wang, and H. Park. A comprehensive feature and data mining study on musician memory processing using eeg signals. In *International Conference on Brain and Health Informatics*, pages 138–148. Springer, 2016.

[19] K. Puk, K.C. Gandy, S. Wang, and H. Park. Pattern classification and analysis of memory processing in depression using eeg signals. In *International Conference on Brain and Health Informatics*, pages 124–137. Springer, 2016.

[20] T. J. Huppert, R. D. Hoge, S. G. Diamond, M. A. Franceschini, and D. A. Boas. A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *NeuroImage*, 29(2):368–382, January 2006.

[21] X. Cui, S. Bray, and A. L. Reiss. Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *NeuroImage*, 49(4):3039–3046, February 2010.

[22] M. M. Plichta, S. Heinzel, A.-C. Ehlis, P. Pauli, and A. J. Fallgatter. Model-based analysis of rapid event-related functional near-infrared spectroscopy (NIRS) data: a parametric validation study. *NeuroImage*, 35(2):625–634, April 2007.

[23] S. Wang, J. Gwizdka, and W.A. Chovalitwongse. Using wireless eeg signals to assess memory workload in the n-back task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435, June 2016.

[24] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[25] M. Hall. *Correlation-based feature selection for machine learning*. PhD Thesis. New Zealand: Department of Computer Science, Waikato University, 1999.

[26] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Maching Learning Research*, pages 1205–1224, 2004.

[27] H. Peng, C. Ding, and L. Fulmi. Feature selection based on mutual information; criteria of max- dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[28] J. R. Vergara and P. A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, 2014.

[29] F. Liu, S. Wang, J. Rosenberger, J. Su, and H. Liu. A sparse dictionary learning framework to discover discriminative source activations in eeg brain mapping. In *AAAI*, pages 1431–1437, 2017.

[30] C. Xiao, S. Wang, L. Zheng, X. Zhang, and W.A. Chaovalitwongse. A patient-specific model for predicting tibia soft tissue insertions from bony outlines using a spatial structure supervised learning framework. *IEEE Transactions on Human-Machine Systems*, 46(5):638–646, 2016.

[31] C. Xiao, J. Bledsoe, S. Wang, W. A. Chaovalitwongse, S. Mehta, M. Semrud-Clikeman, and T. Grabowski. An integrated feature ranking and selection framework for adhd characterization. *Brain informatics*, 3(3):145–155, 2016.

[32] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE transactions on neural networks and learning systems*, 2017.

[33] W. Zhongxin, S. Gang, Z. Jing, and Z.J. Jia. Feature selection algorithm based on mutual information and lasso for microarray data. volume 10, pages 278 – 286, 2016.

[34] S. Wang, C. Xiao, J Tsai, W. Chaovalitwongse, and T. J. Grabowski. A novel mutual-information-guided sparse feature selection approach for epilepsy diagnosis using interictal eeg signals. In *International Conference on Brain and Health Informatics*, pages 274–284. Springer, 2016.

[35] T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2001.

[36] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[37] J. Liu and J. Ye. Moreau-yosida regularization for grouped tree structure learning. In *Advances in Neural Information Processing Systems*, pages 1459–1467, 2010.

[38] X. Wu, V. Kumar, J. Ross Q, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37, December 2007.

[39] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, December 2001.

[40] S. Wang, C.J. Lin, C. Wu, and W.A. Chaovalitwongse. Early detection of numerical typing errors using data mining techniques. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(6):1199–1212, 2011.

[41] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugne, T. Furey, M. Ares, , and D. Haussler. Knowledge-base analysis of microarray gene expressiondata by using support vector machines. *PNAS*, 97(1):262267, 2000.

[42] W.S. Noble. *Support vector machine applications in computational biology*, chapter 3. Computational molecular biology. MIT Press, 2004.

[43] M.A. Gelbart, J. Snoek, and R.P. Adams. Bayesian Optimization with Unknown Constraints. *ArXiv e-prints*, March 2014.

[44] D. S. Beal, J. P. Lerch, B. Cameron, R. Henderson, V. L. Gracco, and L. F. De Nil. The trajectory of gray matter development in brocas area is abnormal in people who stutter. *Frontiers in Human Neuroscience*, 9:89, 2015.

[45] S.E. Chang, D. C. Zhu, A. L. Choo, and M. Angstadt. White matter neuroanatomical differences in young children who stutter. *Brain*, 138(3):694 711, 2015.

[46] S. Chang, M. Kenney, T.M.J. Loucks, and C.L. Ludlow. Brain activation abnormalities during speech and non-speech in stuttering speakers. *NeuroImage*, 46(1):201 – 212, 2009.

[47] N.E. Neef, C. Btfering, A. Anwander, A.D. Friederici, W. Paulus, and M. Sommer. Left posterior-dorsal area 44 couples with parietal areas to promote speech fluency, while right area 44 activity promotes the stopping of motor responses. *NeuroImage*, 142(Supplement C):628 – 644, 2016.

[48] C. A. Kell, K. Neumann, K. von Kriegstein, C. Posenenske, A. W. von Gudenberg, H. Euler, and A.L. Giraud. How the brain repairs stuttering. *Brain*, 132(10):2747–2760, 2009.

CHAPTER 4

Biomarker Identification of Post-traumatic Stress Disorder from Functional Near-Infrared Spectroscopy using a Novel Mutual Information-Guided Sparse Feature Learning Approach

Submitted to IEEE Journal of Biomedical and Health Informatics, 2018.

Biomarker Identification of Post-traumatic Stress Disorder from Functional Near-Infrared Spectroscopy using a Novel Mutual Information-Guided Sparse Feature Learning Approach

Rahilsadat Hosseini, Fenghua Tian, Hanli Liu and Shouyi Wang, *Member, IEEE*

***Abstract*—Post-traumatic stress disorder (PTSD) is a common mental disorder that can develop after a person is exposed to a traumatic event. In clinical practice, it is still a challenging problem to identify PTSD related brain patterns and also assess how a PTSD patient effectively responses to certain treatment. Functional near infrared spectroscopy (fNIRS) is a neuroimaging technique with an excellent temporal resolution for brain activity monitoring. In this study, we made extensive feature extraction of fNIRS brain imaging data and applied the mutual information integrated structured sparse (MISS) feature learning framework that we had previously proposed in our previous study to identify a set of robust fNIRS pattern features as biomarkers to discriminate brain activity patterns of veterans with PTSD and healthy control subjects in a working memory test experimental setting. The feature extraction and the MISS framework discovered three top features as a PTSD biomarkers that can discriminate PTSD patients from the control subjects with a N-fold-cross-validation classification accuracy of 100%. Compared with other popular feature selection methods, MISS framework identified a robust feature subset with the highest discriminative power using the minimum number of features. MISS provides great potentials to facilitate effective personalized PTSD assessment and treatment in the future. Moreover, MISS method is a general feature selection framework for multivariate time series that consider feature structures and take account linear and nonlinear relationships between response and predictor variables.**

***Index Terms*—extensive feature extraction, fNIRS, mutual information, sparse group lasso, feature selection**

I. Introduction

Post-traumatic stress disorder (PTSD) is a common mental disorder that can develop after a person is exposed to a traumatic event, such as warfare, traffic collisions, assault, or others life threats [1]. In the United States, about 9% of people may develop PTSD at some point in their life [2]. PTSD causes neural circuits changes in the brain, and a patient with PTSD may present cognitive dysfunctions, such as memory impairments, attention deficits, and dysexecutive syndromes. In order to diagnose and characterize brain responses of the patients with PTSD, neuroimaging techniques like functional magnetic resonance imaging (fMRI) [3] , [4] , and functional near infrared spectroscopy (fNIRS) [5], [6], and functional imaging techniques of single-photon emission computed tomography (SPECT), positron emission tomography (PET) [7] have been applied.

Functional near infrared spectroscopy (fNIRS) is a non-invasive, portable, and low-cost neuroimaging technology for brain activity monitoring with excellent temporal resolution that monitors hemodynamic changes in the concentration of oxygenated (HbO2) and deoxygenated (Hb) hemoglobin molecules in the blood, which can be used to assess cerebral brain activity on the basis that neural activation and vascular response are tightly coupled [8]. Through decades of development, fNIRS has become a valuable neuroimaging technique for its portability, and reliability. The application of fNIRS in cerebral functioning studies has been validated by other neuroimaging techniques, which showed that the fNIRS signal maintains a strong correlation with the fMRI Blood Oxygen Level Dependent (BOLD) signal [9], [10], [11], as well as the PET measures of changes in regional cerebral blood flow (rCBF) [12]. FNIRS has been growing rapidly in clinical settings and research and has been used in many studies of brain functions and brain disorders, e.g., attention deficit hyperactivity disorder (ADHD) and Autism [13], depression [14], etc.

The main treatments for patients with PTSD are counseling and medication and various types of treatment and interventions have been proposed and utilized in the past decades, such as trauma-focused cognitive behavioral therapy (CBT), cognitive processing therapy(CPT), and prolonged exposure (PE) [15]. However, there is considerably less attention given to the accurate assessment of treatment effectiveness using brain imaging biomarkers via portable and reliable neuroimaging techniques like fNIRS. Most of the existing studies on PTSD pattern recognition are based on structural MRI data [16]. It is desired to discover informative biomarkers (e.g., brain activity and patterns) to evaluate whether an individual responds to a treatment well and moves towards the right direction to the measures of the healthy control population. In this study, we aimed to develop an effective data-driven method to discover important fNIRS features from sparse number of voxels (channels) as biomarkers that are highly discriminative between the PTSD group and the control group.

In our previous study, Tian et al. [5] used a 36 channel fNIRS setup to image the prefrontal activations in a group of veterans diagnosed with PTSD and a group of age/gender-

matched healthy controls during two working memory tasks, namely a digit forward task and a digit backward task. Both tasks required serial encoding, maintaining and recall of a string of 6 digits presented on the computer screen. In the digit forward task, the subjects recalled the digits in the same order as they were presented; in the digit backward task, the subjects recalled the digits in the reverse order. The healthy controls showed robust hemodynamic activations during the encoding and retrieval processes. In contrast, the veterans with PTSD were found to have activation during the encoding process, but followed by distinct deactivation during the retrieval process. This deactivation was more pronounced in the right dorsolateral prefrontal cortex (DLPFC). It appeared that veterans with PTSD suppressed prefrontal activity during memory retrieval, which could be a useful biomarker to evaluate the cognitive dysfunction associated with PTSD.

In this study, we made extensive feature analysis on the fNIRS data obtained by Tian et al. [5] and applied the effective mutual-information-based sparse feature learning approach (MISS) which had been proposed in our previous study [17], to discover the most critical features that discriminate patients with PTSD and the control group. All of the computations are conducted in Matlab. In particular, we extracted eight groups of features from fNIRS signals. We explored distinguishing patterns among all of the possible combination of experimental design factors, including forward or backward, the phase of the process (encode, maintain and recall) and the source of fNIRS measurements (oxygenated hemoglobin, HbO2 and deoxygenated hemoglobin, Hb). We discovered that the third phase of the experiment, namely the recall phase, was the most significant phase period, and that the most discriminative feature groups were statistical measures, autocorrelation, Hjorth parameters and SVD. With only top three features selected by MISS from autocorrelation group, we achieved 100% accuracy in classification of PTSD from controls. MISS outperformed 2 popular feature selection techniques namely minimum redundancy and maximum relevancy (mRMR) [18] and sparse group LASSO (SGL) [19] by capturing both linearly and non-linearly related features to the response variable from sparse number of voxels (channels) that can be considered as the region of interest (ROI) in this study. RIO discovery can improve the interpretation and classification accuracy [13]. Defining voxels (channels) as the non-overlapping feature groups results in the selection belonging to the same channel. Sometimes it is a necessity to choose features in a group and sometimes it just provides higher interpretability and repeatability [20].

The rest of the paper is organized as follows. In Section 2, we present the proposed methods, mainly including feature extraction of fNIRS data and feature selection techniques presenting information theory-integrated structured sparse feature learning models. In Section 3, we show results of biomarker pattern discovery and performance comparisons of several popular feature selection techniques. In section 4, we have conclusions.

II. METHOD

A. Participants, fNIRS data acquisition and preprocessing

A total of 16 war-zone veterans diagnosed with PTSD and 16 healthy controls were the two groups of participants that matched in age ($age = 29.4 \pm 9.6 years$) and gender (all males) [5]. As shown in Figure 1, the participants were instructed to complete a session of digit forward task (eight trials) and a session of digit backward task (eight trials) sequentially while their brain activities were scanned by a high-performance fNIRS brain imager (Cephalogics LLC., Boston, MA). The fNIRS system acquired data from 36 source-detector pairs (channels) placed on the forehead. The location of the fNIRS channels is shown in Figure 2. The sampling rate was 10.8 Hz for the fNIRS signals. The channel-wise fNIRS data from each task session was preprocessed using a standard toolbox, Homer [21] to remove significant motion artifacts. The data in optical density were then low-pass filtered at 0.2 Hz and high-pass filtered at 0.01 Hz. Then the changes of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) concentrations were calculated for each channel. At last, for each task, event-related HbO_2 and Hb changes were averaged over the good trials to obtain averaged hemodynamic responses, which were the data used in this study.

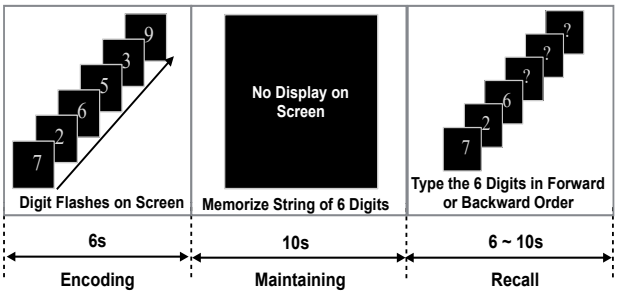


Fig. 1. A session for digit forward or backward task with definition of three phases: encode, maintain and recall

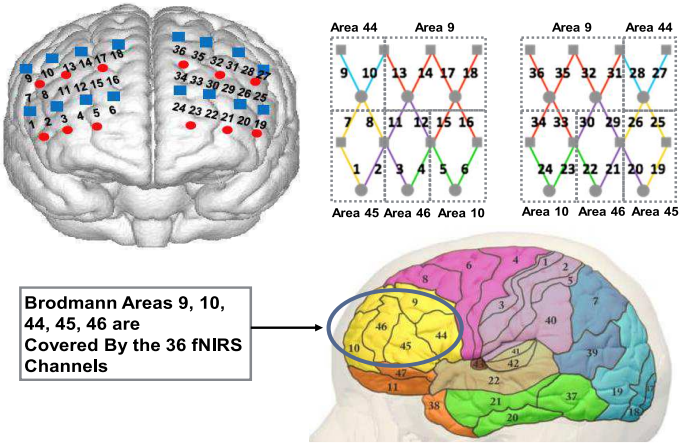


Fig. 2. Location of channels in Brodmann area on the brain

B. Feature Extraction

As shown in Figure 1, a trial of fNIRS data can be segmented to three phases: the encoding phase (0-6s), the

maintenance phase (6-16s), and the recall phase (16-trial end). In feature extraction, we also consider two experimental factors: 1) two recall tasks in forwarding or backward order of 6 digits, 2) data measurements on the frontal activation in oxygenated or deoxygenated hemoglobin changes respectively $\Delta[HbO_2]$, $\Delta[Hb]$ as estimates of Hemoglobin Response (HR) [22].

The averaged fNIRS signal (in terms of Hb and HbO_2) of each subject in PTSD group and control group under backward and forward tasks are shown in Figures 3, 4, 5 and 6. Magenta lines are the average of the fNIRS signal over 36 channels for each veteran with PTSD, and the blue lines depict average of the fNIRS signal over 36 channels for each healthy control subject. In each of these figures, the dotted lines show the group averages of all participants in each group. From the figures, one can observe that although the group average of PTSD and control group show different temporal hemodynamic response patterns, the cross-individual variability of the hemodynamic response patterns are very high. Thus, we were aimed to discover a set of robust discriminative features hidden in the highly variable fNIRS response patterns and can identify PTSD signature patterns accurately for each individual subject. For a data-driven approach, we made an extensive feature extraction investigation on the Hb and HbO_2 response patterns in the three phases under the experimental settings.

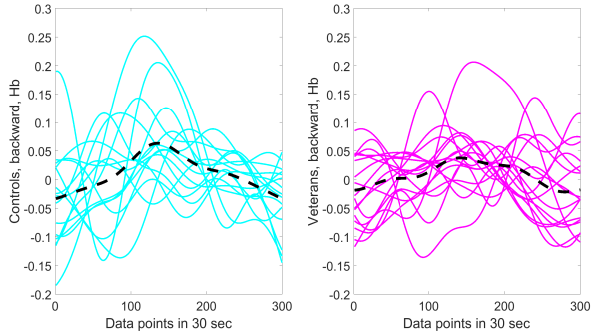


Fig. 3. The averaged Hb concentration values of each subject in the backward task: the thin blue solid lines on the left subplot represent healthy control subjects and the magenta lines on the right subplot represent veterans with PTSD. The thicker dotted lines are group averages.

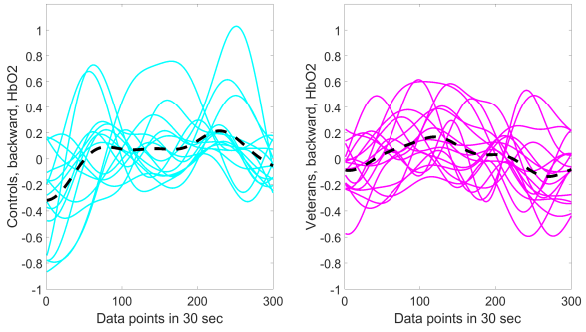


Fig. 4. The averaged HbO_2 concentration values of each subject in the backward task: the thin blue solid lines on the left subplot represent healthy control subjects and the magenta lines on the right subplot represent veterans with PTSD. The thicker dotted lines are group averages.

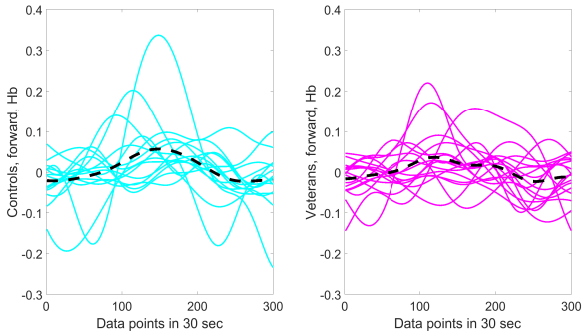


Fig. 5. The averaged Hb concentration values of each subject in the forward task: the thin blue solid lines on the left subplot represent healthy control subjects and the magenta lines on the right subplot represent veterans with PTSD. The thicker dotted lines are group averages.

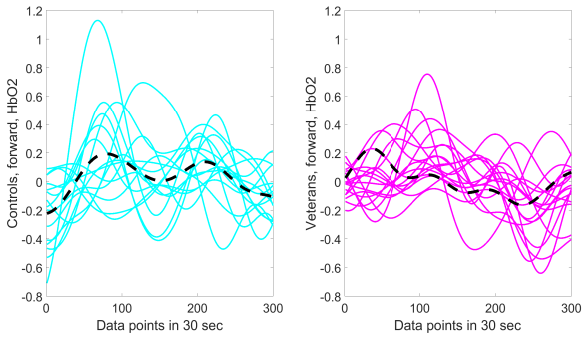


Fig. 6. The averaged HbO_2 concentration values of each subject in the forward task: the thin blue solid lines on the left subplot represent healthy control subjects and the magenta lines on the right subplot represent veterans with PTSD. The thicker dotted lines are group averages.

We extracted eight groups of features summarized in the following.

- Statistical features capture descriptive information of the signals.
- Number of peaks and zero crossing capture morphological features. Zero Crossings is the number of times the value of the feature cross the zero line.
- Hjorth parameters capture activity, mobility, and complexity of a signal's variation in time. The three features are defined as: $activity = Var(y(t))$, $mobility = \sqrt{\frac{var(y(t)dy/dt}{var(y(t))}}$, $complexity = \frac{mobility(y(t)dy/dt)}{mobility(y(t))}$.
- Normalized Area Under Signal (NAUS) calculates the sum of values which have been subtracted from a baseline (first value in each phase), divided by the sum of the absolute values for the fNIRS signal.
- Autocorrelation captures linear relationship of the signal with its historical values considering different delays, Kendal, partial, Spearman and Pearson are four ways of calculation. In this study, eight delays of 0.28, 0.56, 0.74, 1.02, 1.30, 1.57, 1.76, 2.04 seconds were employed for autocorrelation features.
- Bicorrelation computes the bicorrelation on the time series X_v for given delays in τ_v. Bicorrelation is an ex-

tension of the autocorrelation to the third order moments, where the two delays are selected so that the second delay is twice the original , (i.e. $x(t)x(t-\tau)x(t-2\tau)$). Given a delay of τ and the standardized time series X_v with length n, denoted as Y_v, the $bicorr(\tau)$ can be calculated as:

$$\frac{\sum_{j=1}^{n-2\tau} Y_v(j)Y_v(\tau+j)Y_v(2\tau+j)}{n-(2\times\tau)} \quad (1)$$

- Singular Value Decomposition (SVD) derived features: SVD is a generalized form of eigen decomposition of positive semi-definite normal matrix. In particular, a $m \times n$ matrix $\mathbf{M}$ can be decomposed to three terms: $M_{m\times n} = \mathbf{U}_{m\times m}\Sigma_{m\times n}\mathbf{V}^*_{n\times n}$, where $\mathbf{U}, \Sigma, \mathbf{V}$ are unitary matrix, rectangular diagonal matrix and real or complex unitary matrix, respectively. Columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal bases. Singular values in this study is calculated based on the row wise channels, meaning after decomposition of the feature matrix consisting of 36 rows (channels) and 300 columns as data points, we had diagonal values of Σ known as σ_i of matrix $\mathbf{M}$ (features matrix). We employed the 36 singular values, logarithm of the singular values, and the range of the singular values as the SVD features of fNIRS signals.

C. Personalized Feature Normalization & Processing

A challenge of many biomedical studies is high inter-individual variability. As one can observe in Figure 3-6, the collected fNIRS signals vary dynamically across subjects. Thus, the corresponding signal features can vary largely and it is difficult to build a robust diagnostic model to discriminate PTSD subject accurately. In addition, due to various artifacts existing in the collected signals, there are inevitable outliers in the extracted signal features which can also distort model training and deteriorate model generalization performance. To tackle these issues, we applied a personalized feature normalization approach described in study [23] to standardize the extracted feature values to increase feature interpretability across subjects. The personalized feature scaling reduces inter-individual variability that may be caused by signal drift and baseline changes and eliminates feature outliers.

In summary, as shown in Figure 7, there are 12 combinations of experimental settings in feature selection: forward/backward tasks, encoding/maintaining/recall phases, and Hb/HbO_2 signals; and for each setting, eight groups of 115 features were extracted. Thus, a total number of $12 \times 115 = 1380$ features were extracted for each subject. In the following, we will present a novel structured sparse feature selection framework that integrates information theory with structured sparse modeling to discover the fNIRS pattern signatures to access PTSD brain activity in memory tasks.

D. Integrated Structured Sparse Feature Selection using Mutual Information

Feature selection techniques have been widely used to identify most important decision variables, to avoid overfitting and improve model performance, and to gain a deeper insight into the underlying processes or problem. Most feature selection techniques generally can be categorized into three categories: embedded, wrapper, and filter methods [24]. Both embedded and wrapper methods rely on an employed classifier or model therefore, the feature selection performance is specific and limited to the embedded classification/prediction models. Typical such approaches include Pudi's floating search [25] and stepwise selection [26]. Filter feature selection techniques assess the relevance of features by looking only at the intrinsic properties of the feature values. Some popular examples include correlation-based feature selection [27], fast correlation-based feature selection [28], and minimum redundancy maximum relevance (mRMR) [18], information-theoretic-based feature selection methods [29]. In addition, sparse modeling-based feature selection methods have gained increasing research interests due to well-grounded mathematical analysis and optimization theories. These feature selection algorithms employ sparsity-inducing regularization techniques, such as L_1-norm constraint or sparse-inducing penalty terms, for variable selection. Recently, to construct more interpretable models, structured sparse modeling algorithms that consider feature structures have been proposed and shown promising results in many practical applications including computer vision, gene expression, medical imaging analysis, etc. [30], [31]. However, most of the current structured sparse modeling algorithms only consider linear relationships between response variables and predictor variables (features), some complex nonlinear relationships could be missed in the linear function modeling procedure. On the other hand, some filter-based methods and wrapper methods could capture nonlinear relationships between features and response variables, but the feature structure usually cannot be well considered in the feature selection procedure. To make interpretable learning models with efficient feature selection is still an open and active research area in machine learning community. To consider both linear and nonlinear relationships between features and response variable, and to consider feature structure (in this study defined as voxels / channels) in feature selection, we apply the novel feature selection framework that integrates information theory-based feature filtering and structured sparse learning models. This method is explained in details and has been applied on a different set of FNIRS data in study [17]. However we briefly describe the components of the MISS.

1) Structured Sparse Feature Selection: A sparse model generates a sparse solution with a small number of variables with non-zero weights among all the variables in the model. The most basic sparse model is least absolute shrinkage and selection operator (LASSO) regression, which employ L_1 penalty-based regularization techniques for feature selection [32]. The LASSO model and various L_1 regularized models assume that features are independent and do not consider structures of features. However, in most practical applications, features follow some essential structures, such as disjoint groups, overlapping groups, tree-structured groups, and graph networks [33]. The feature structures can be greatly useful to guild the optimization procedure and help identify the important features with better interpretability. In the MISS method, we improved the sparse learning from basic LASSO

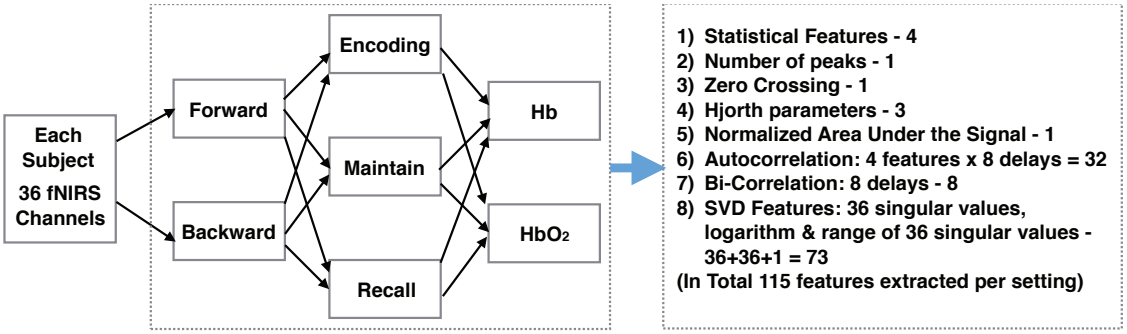


Fig. 7. The feature extraction structure for each subject and each channel.

model to structured learning by applying SGL, and used voxels (channels) to define non-overlapping groups of extracted features. The sparse group sparsity is designed to produce a solution with simultaneous between- and within group sparsity. The SGL regularization is formulated as:

$$\min\left\{\|\mathbf{Ax}-\mathbf{Y}\|+\lambda_1\|\mathbf{x}\|_1+\lambda_2\sum_{i=1}^{g}\omega_i^g\|\mathbf{x}_{G_i}\|_2\right\} \qquad (2)$$
$$A\in\mathbf{R}^{m\times n},\quad y\in\mathbf{R}^{m\times 1},\quad x\in\mathbf{R}^{n\times 1},$$

where the weight vector $\mathbf{x}$ is divided by g non-overlapping groups: $\{x_{G_1}, x_{G_2}, \ldots, x_{G_g}\}$, and $\omega_i^g$is the weight i for group g. The parameter λ_1 is the penalty for sparse feature selection, i.e. weights of some features in non-zero groups can be zero, and the parameter λ_2 is the penalty for sparse group selection, i.e. the weights of some feature groups will be all zeros. In this study, λ_2 decides the selection of the voxels (channels) which can be interpreted as ROI and λ_1, decides the features of non-zero (selected) voxels.

2) Mutual Information for Feature Selection: In information theory, mutual information (MI) is a measure of inherent dependence between two independent variables [34]. MI measures how much information a feature contains about the class without making any assumptions about the nature of their underlying relationships, moreover it captures nonlinear relationship between random variables and is invariant under transformation of the features [35]. The mutual information of two variables X and Y, denoted by $I(X,Y)$, is calculated by:

$$I(X;\,Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right), \qquad (3)$$

where p(x) and p(y) are the marginal probability and $p(X,Y)$ is the joint probability distribution of the variable X and Y. MI can be applied to rank features and has been frequently used for feature selection [18], [36], [37]. The basic idea of most MI-based feature filtering methods is to keep the more informative features (with higher MI) and remove the redundant or less-relevant features (with low MI) in a filtering procedure, a very popular example is mRMR [18]. However, mRMR is not global because of greedy search and yields selection from all over the feature space [20].

Although these approaches can work well in many cases, they are subject to issues of missing some important features by just excluding low MI-ranked features and important feature structures. On the other hand, most of the current structured sparse modeling algorithms can only handle linear relationships between response and predictor variables. Based on this consideration, we applied MISS for feature selection.

3) Integrated MI-Sparse Feature Selection Framework: The key idea of the proposed approach is to take into account structured feature dependency while keeping the searching process highly computationally efficient. The proposed mutual-information-guided feature selection framework is built on the three steps: MI-based feature ranking for high-MI features, structured sparse feature learning on low MI-ranked features, and integration of the selected high- and low-ranked features in an enumeration procedure. In the feature ranking step, we use MI to rank features and identify a subset of high MI features that have the best informative power individually to class labels. Among those features, the highly correlated features are considered as redundant features and removed in a way similar to the mRMR approach. Given a number of features k, the subset of top k features ranked by MI is denoted by S, and the subset of the remaining features is denoted by W. In the second step, we employ the structured sparse learning algorithms based on feature structure of the studied problem. A structured sparse model (as described above) is employed to select important feature subset with combined discriminative power from the low-ranked features set W. Assume k_2 features are selected by the structured sparse learning algorithm. The next step is the only difference between the applied MISS in this study and application of MISS in study [17]. In study [17], we used all the nonzero coefficients selected by the sparse learning method. But in this study, the third step is to further reduce the feature subset and discover the optimal feature subset by exploring the k_1 high-MI features and the k_2 sparse-model selected low-MI features. Within a small set of $(k_1 + k_2)$ features, it becomes possible to enumerate all the combinations of the selected feature subsets with a small feature pool. Feature subset evaluation is based on the cross-validation classification performance. In particular, we propose to evaluate feature subset in an ascending order of feature set size. It starts with one feature, then combinations of 2, 3, and .... The subset evaluation stops (optimal feature subset is reached) when the cross-validation accuracy converges and cannot be further improved. The applied mutual-information-guided structured sparse feature selection (MISS) framework

is shown in Figure 8, if the sparse learning is based on LASSO, it is referred as MILASSO, and if based on SGL, as MISGL. Iterative feature integration framework is illustrated in Figure 9.

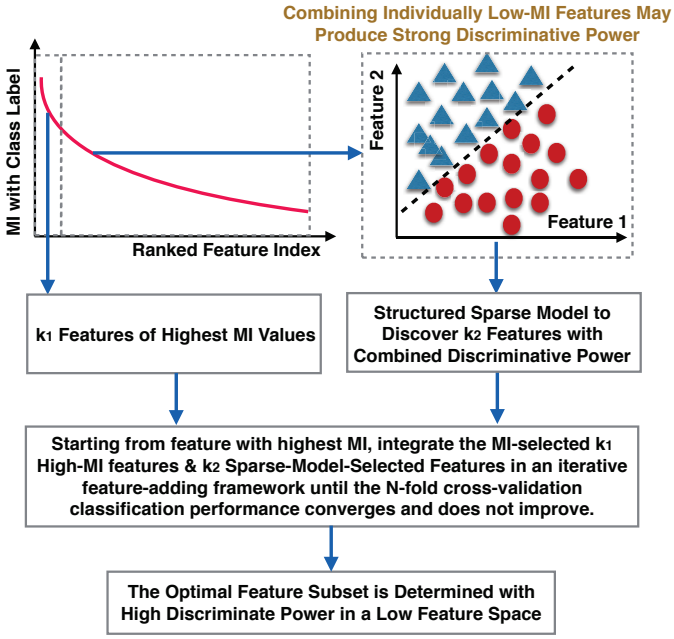


Fig. 8. The framework of the mutual information-guided feature selection approach.

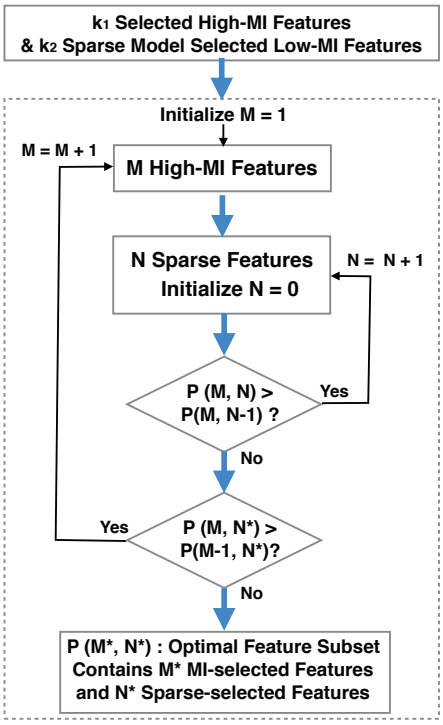


Fig. 9. The optimal feature subset selection by exploring the MI-selected features and sparse-model selected features. The P(M,N) in the flowchart denotes the best N-fold cross-validation classification accuracy using M MI-selected features (after removing the highly correlated ones, with .96 threshold) and N sparse-model selected features.

4) Classification Method: In this study, we employed Proximal Support Vector Machines (PSVM) [38] as the classifier to access the discriminative performance of feature subset. PSVM is a robust fast alternative for SVM but it assigns the points to the closest two parallel planes instead of two disjoint half-spaces. SVM has been a successful classification model in many brain related studies [23], [39], [40], [41]. Although there is application of other popular classifiers in fNIRS studies like GLM [42] and LDA [43], SVM is commonly used [44] and in some cases comparison proved its efficacy [14]. In general, a SVM model finds a hyperplane that divides two classes (PTSD class as 1 and control class as -1) with the least error and maximum distance of the closest sample in each class to the separating hyperplane using the extracted features' patterns [45]. In this study, the linear PSVM model was employed to discriminate features of PTSD and control group.

E. Training and Evaluation

Feature selection was conducted inside N-fold cross validation loop, on the training set. Therefore, all the available parameters of the selection algorithms namely, the regularization parameters in SGL formulation (λ_1 and λ_2) and number of features with high MI values are optimized/tuned in training loop. The reported results pertain to the highest performance for the best combination in the optimization. After iterations for feature selection on N fold ends, we had N (number of folds) groups of K selected features, meaning, we eventually had an ensemble of feature rankings that needed to be integrated by application of consensus ranking. The one-consensus ranking was defined based on the frequency of selection over the N-fold procedure. In particular, the ranks for features were defined from the highest to the lowest frequency of selection in N-fold cross validation procedure, and then the first K features were picked. If there was same frequency for some features, the decision was based on the priority of that feature in each loop of selection. Priority of each feature was the average of its rank in all folds of selection.

We applied 10-fold cross-validation for training and testing. We selected the features and optimized the parameters of the classification algorithm on the training set and then applied the results on the testing set. Accuracy is the ratio of correctly classified test subjects to the total number of test subjects. Sensitivity is true positive rate, i.e. the accuracy of PTSD group, meaning ratio of correctly predicted of PTSD subjects to all PTSD subjects. Specificity is true negative rate, i.e. the accuracy of the control group. In this study, we used the average of sensitivity and specificity as an unbiased accuracy measure to the binary classification performance.

III. Experimental Results

The classification results of raw fNIRS signals are shown in Table I. The average classification performance is between 53 to 69% for different task conditions and data settings.

The poor classification performance confirmed the existing of the inevitable high cross-individual variability of hemodynamic response patterns in fNIRS data as shown in Figures 3 - 6. Thus, we applied the proposed extensive feature extraction of raw fNIRS signal and feature selection (MISS) from 12 possible combinations of experimental settings to discover a

TABLE I
CLASSIFICATION PERFORMANCE OF LINEAR PSVM FOR RAW FNIRS SIGNALS BETWEEN VETERANS WITH PTSD AND HEALTHY CONTROLS

Description	Task	Source	Accuracy	Sen	Spe
Only Fw	Fw	both	68.8	64.7	73.3
Only Bw	Bw	both	59.4	64.7	53.3
Only Hb	both	Hb	53.1	52.9	53.3
Only HbO2	both	HbO2	59.4	64.7	53.3
All	both	both	62.5	64.7	60

Fw: Forward, Bw: Backward
Sen: sensitivity, Spe: specificity

set of robust subject-invariant discriminative features as PTSD pattern signatures.

A. Statistics of the Extracted Features & Phase's Significance

By visualizing the statistics of the extracted features, it is discovered that the most discriminative phase of the experiment is the last one, namely recall phase, when the participants are asked to recall the digits. In Figure 10, 11, 12 and 13 the significance of the top rank feature in each outperforming group including statistics, Hjorth parameters, autocorrelation and SVD respectively is shown. Boxplot of the values of the top feature in each group indicates that in the last phase meaning recall, values of the feature becomes significantly differentiating between the classes. Based on the box plots, channel-wise mean and autocorrelation in controls group have higher values than PTSD group, this means that there is more active potential and higher correlation with historical values for fNIRS measurements in healthy control group. On the other hand, singular values after SVD and Hjorth parameter mobility in veterans with PTSD have higher values, which means there is more variation in fNIRS measurement for this class. All these quantitative conclusions are consistent with observed differences between two classes in Figures 3, 4, 5 and 6.

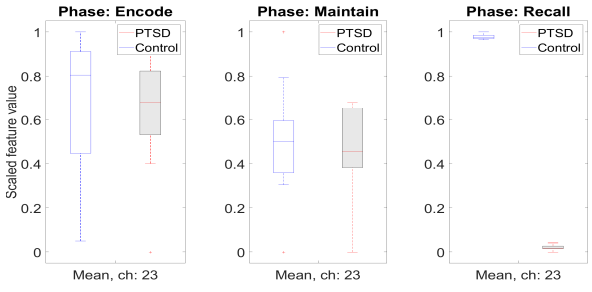


Fig. 10. Boxplot of the values of top statistics feature: mean, among 3 phases of the experiment: encode, maintain and recall

B. Evaluation of Feature Selection Methods

Without application of any feature selection techniques, the classification results using all the extracted eight groups of features are shown in Table II. The best performance (highlighted in bold) was achieved with an accuracy of 92.31% (the average of sensitivity and specificity) at the settings of the backward task, recall phase, and HbO_2 while in other combination of settings ranged from 48 to 71%. From Table II, we observe that the features in recall phase demonstrate significant discriminative patterns for the two groups. Thus,

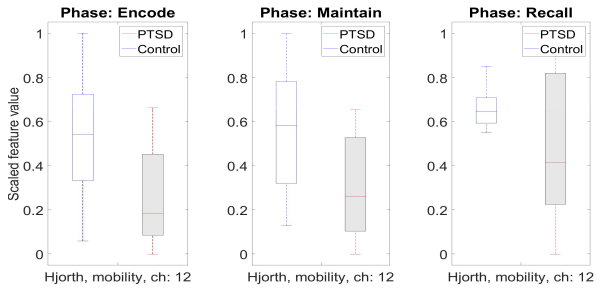


Fig. 11. Boxplot of the values of top Hjorth parameters feature: mobility, among 3 phases of the experiment: encode, maintain and recall

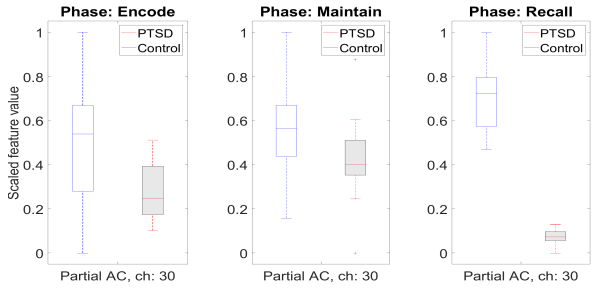


Fig. 12. Boxplot of the values of top autocorrelation feature: partial, among 3 phases of the experiment: encode, maintain and recall

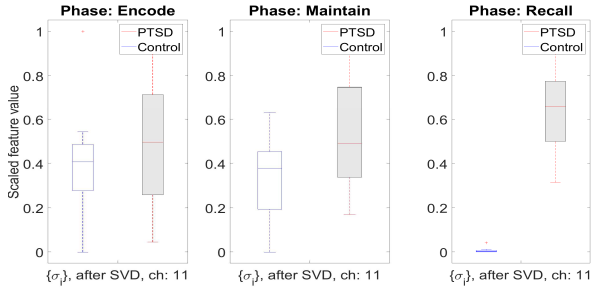


Fig. 13. Boxplot of the values of top SVD feature: singular values, among 3 phases of the experiment: encode, maintain and recall

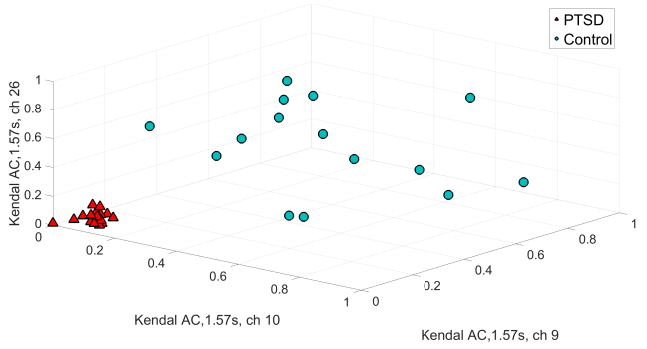


Fig. 14. The three selected features by the proposed $MISS_{SGL}$ feature selection framework. The three selected features combined together can achieve an accuracy of 100% to discriminate the PTSD patients from the control subjects. Compared with other feature selection methods, the proposed MISS feature selection framework identified a feature subset with the highest discriminative power and the minimum number of features.

TABLE II
CLASSIFICATION PERFORMANCE OF LINEAR PSVM FOR ALL EXTRACTED FEATURES FROM FNIRS SIGNALS WITHOUT APPLICATION OF FEATURE SELECTION TECHNIQUES BETWEEN VETERANS WITH PTSD AND HEALTHY CONTROLS, EACH ROW SHOWS DIFFERENT SETTING FOR EPOCH DEFINITION: MEANING DIFFERENT COMBINATION OF TASK, PHASE AND SOURCE OF THE MEASUREMENT

Task	Phase	Source	Accuracy	Sen	Spe
Fw	Encode	Hb	48.08	50.0	46.2
Fw	Encode	HbO2	58.17	62.5	53.8
Fw	Maintain	Hb	49.20	29.2	69.2
Fw	Maintain	HbO2	49.20	79.2	19.2
Fw	Recall	Hb	65.87	62.5	69.2
Fw	Recall	HbO2	54.97	79.2	30.8
Fw	All trial	Hb	52.56	16.7	88.5
Fw	All trial	HbO2	60.26	66.7	53.8
Bw	Encode	Hb	53.85	50.0	57.7
Bw	Encode	HbO2	63.62	54.2	73.1
Bw	Maintain	Hb	55.77	50.0	61.5
Bw	Maintain	HbO2	47.76	41.7	53.8
Bw	Recall	Hb	71.15	50.0	92.3
Bw	*Recall*	*HbO2*	**92.31**	**100.0**	**84.6**
Bw	All trial	Hb	39.90	37.5	42.3
Bw	All trial	HbO2	55.93	54.2	57.7
Fw& Bw	Encode	Hb & HbO2	69.55	58.3	80.8
Fw& Bw	Maintain	Hb & HbO2	41.83	37.5	46.2
Fw& Bw	Recall	Hb & HbO2	82.69	100.0	65.4
Fw& Bw	All trial	Hb & HbO2	53.69	45.8	61.5

Fw: Forward, Bw: Backward, Sen: sensitivity, Spe: specificity

TABLE III
COMPARING PERFORMANCE FEATURE SELECTION TECHNIQUES, MI GUIDED SPARSE SELECTION METHODS OUTPERFORMS MRMR AND SGL

Selection method	Number of selected feats	Best setting for epoch definition	Accuracy	Sen	Spe
mRMR	3	Fw task, recall phase, Hb	84.29	91.67	76.92
	10	Fw task, recall phase, Hb & HbO2	92.15	95.83	88.46
	45	Fw & Bw task, recall phase, Hb	98.07	100.00	96.15
SGL	3	Fw & Bw task, recall phase, Hb & HbO2	94.00	93.75	87.50
	10	Fw & Bw task, recall phase, Hb & HbO2	96.00	95.83	91.67
	19	Fw & Bw task, recall phase, Hb & HbO2	100.00	100.00	100.00
MILASSO	3 : 1 MI, 2 LASSO	Fw task, recall phase, Hb & HbO2	90.22	95.83	84.62
	10: 1 MI, 9 LASSO	Fw task, recall phase, Hb	98.08	100.00	96.15
	16: 8 MI, 8 LASSO	Fw task, recall phase, Hb	100.00	100.00	100.00
MISGL	3: 1 MI, 2 SGL	Fw & Bw task, recall phase, HbO2	100.00	100.00	100.00

feat: features, Sen: sensitivity, Spe: specificity, Fw: forward, Bw: backward
MI: number of selection based on mutual information, SGL: based on sparse group LASSO
In each section of the feature selection method, first row shows the result for 3 selection,row 2, 10 features and last row shows the highest performance possible to chieve with more features

we applied feature selection on the recall phase using the MILASSO, MISGL, SGL, mRMR feature selection methods. Table III summarizes the feature selection and classification performance of the four feature selection approaches.

For mRMR method, as a filtering method, the classification accuracies of the top 3, 10, and 45 features are shown in the Table II, ranging from 84 to 98% and there was no improvement with more than 45 features. For SGL method with the optimized values for λ_1 and λ_2, the accuracy ranged from 94 with 3 features to 100% with 19 features. For MILASSO and MISGL, implementation of MISS yielded higher accuracy with less number of features. As shown in Figure 14, the three selected features have strong combined discriminative power and can be used as the robust and subject-invariant bio-signature to discriminate PTSD patients from the control subjects.

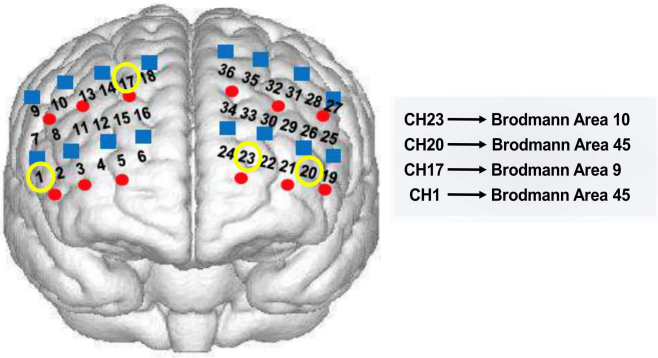


Fig. 15. Selected discriminative channel locations; channels 23, 20, 17 and 1 are frequently selected with the four applied feature selection methods

IV. CONCLUSIONS

The key value of this study was application of a novel effective feature learning method which has been proposed in

our previous research [17] and to discover brain biomarkers of PTSD using fNIRS imaging data. With a set of robust biomarkers, the hypothesis on the patient effectively responding to the treatment can be tested.

The first contribution of this study was an extensive feature extraction and pattern study for fNIRS imaging data and explored the feature groups that demonstrate discriminative patterns between veterans with PTSD and healthy control subjects. We presented g a new feature extraction method which was derived from singular value decomposition and was found to be individually very promising in the discrete format for classification. The most distinguishing feature groups were statistics, bicorrelation, autocorrelation, singular values after SVD and Hjorth parameters. The second contribution was application of the MISS framework which integrates information theory with structured sparse learning theory to achieve efficient feature selection. Compared with the current popular feature selection techniques, namely mRMR which achieved the highest accuracy of 98% with 45 features, SGL which obtained the highest accuracy of 100% using 19 features, MISS outperformed and achieved the highest performance of 100% with 16 features with MILASSO, and reached the highest performance of 100% with only three features with MISGL. Therefore, MISS structured feature selection approach is capable of finding the most sparse feature subset with the highest discriminative power and a minimum number of feature size.

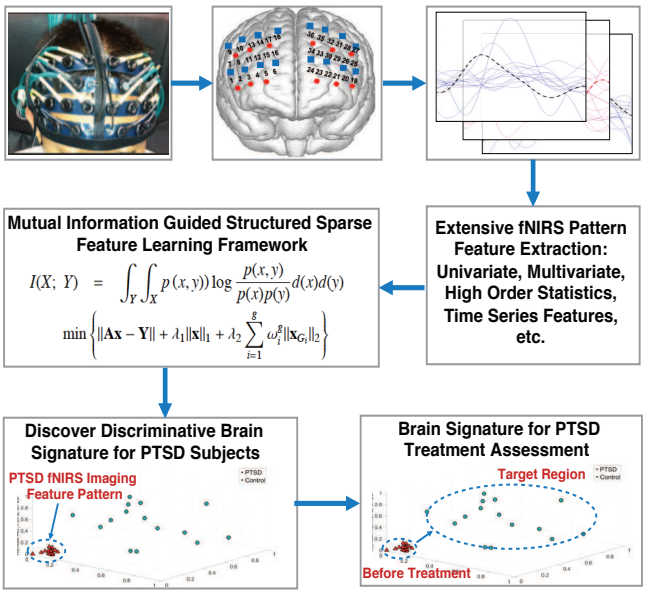


Fig. 16. The potential application of the identified biomarkers to make an assessment of PTSD treatment methods and help doctors/physicians determine the most effective treatment plan for each individual patient. An effective treatment is assumed to move the feature pattern of a PTSD patient toward the target feature pattern region of health control group. The proposed feature extraction and MISS feature selection framework provide great potentials to facilitate effective personalized PTSD assessment and treatment in the future.

The proposed feature extraction and effective feature learning method discovered top three features that can be used for perfect separation of PTSD patients from healthy controls. These top three features are Kendal auto-correlation values for 1.57 sec delay of fNIRS signals from three channels which can be used as the pre-treatment biomarkers predicting PTSD. The most frequently selected channels were located in Brodmann areas 9, 10 and 45 which could be recognized as the ROI in this study. Most importantly, this study can generate high impact on fNIRS brain imaging analysis with potentially important applications to PTSD treatment assessment. In the current clinical practice, it is still a challenging problem to quantify and accurately assess how a PTSD patient effectively responses to a certain treatment plan. The existing brain imaging analysis tools cannot handle this problem well. The proposed feature extraction and the MISS feature selection framework provide a novel data-driven approach to discover a set of robust biomarkers to discriminate PTSD patients from the control subjects. As shown in Figure 16, the identified biomarkers can be used to make an assessment of PTSD treatment methods and help doctors/physicians determine the most effective treatment plan for each individual patient. An effective treatment is assumed to move the feature pattern of a PTSD patient toward the identified target feature region of health control group. The MISS selected top fNIRS imaging pattern features in this study provide one possible solution to achieve this goal and facilitate effective personalized PTSD assessment and treatment in the future. More importantly, the applied information-theory-guided structured sparse feature selection MISS framework is a general framework which can be applied in the analysis and learning of any multivariate time-series.

ACKNOWLEDGMENT

Assistance provided by Alexa Smith-Osborne was greatly appreciated.

REFERENCES

[1] J. I. Bisson, S. Cosgrove, C. Lewis, and N. P. Roberts. Post-traumatic stress disorder. *BMJ*, 351, 2015.

[2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. American Psychiatric Publishing, Arlington, VA, 5 edition, 2013. An optional note.

[3] C. Hou, J. Liu, K. Wang, L. Li, M. Liang, Z. He, Y. Liu, Y. Zhang, W. Li, and Jiang T. Brain responses to symptom provocation and trauma-related short-term memory recall in coal mining accident survivors with acute severe ptsd. *Brain Research*, 1144:165 – 174, 2007.

[4] Jasmeet Hayes, Michael VanElzakker, and Lisa Shin. Emotion and cognition interactions in ptsd: a review of neurocognitive and neuroimaging studies. *Frontiers in Integrative Neuroscience*, 6:89, 2012.

[5] F. Tian, A. Yennu, A. Smith-Osborne, F. Gonzalez-Lima, C. S. North, and H. Liu. Prefrontal responses to digit span memory phases in patuients with post-traumatic stress disorder (ptsd): A functional near infrared spectroscopy study. *NeuroImage: Clinical*, 4:808–819, 2014.

[6] K. Matsuo, K. Taneichi, A. Matsumoto, T. Ohtani, H. Yamasue, Y. Sakano, T. Sasaki, M. Sadamatsu, K. Kasai, A. Iwanami, N. Asukai, N. Kato, and T. Kato. Hypoactivation of the prefrontal cortex during verbal fluency test in ptsd: a near-infrared spectroscopy study. *Psychiatry Research: Neuroimaging*, 124(1):1 – 10, 2003.

[7] V. Francati, E. Vermetten, and J.D. Bremner. Functional neuroimaging studies in posttraumatic stress disorder: review of current methods and findings. *Depression and Anxiety*, 24(3):202–218, 2007.

[8] OJ. Arthurs and SJ. Boniface. What aspect of the fmri bold signal best reflects the underlying electrophysiology in human somatosensory cortex? *Clinical Neurophysiology*, 114(7):1203 – 1209, July 2003.

[9] B.J. MacIntosh, L.M. Klassen, and R.S. Menon. Transient hemodynamics during a breath hold challenge in a two part functional imaging study with simultaneous nearinfrared spectroscopy in adult humans. *NeuroImage*, 20(2):1246 – 52, Oct 2003.

[10] T.J. Huppert, R.D. Hoge, S.G. Diamond, M.A. Franceschini, and D.A. Boas. A temporal comparison of bold, asl, and nirs hemodynamic responses to motor stimuli in adult humans. *NeuroImage*, 29(2):368–82, Jan 2006.

[11] X. Cui, S. Bray, D. M. Bryant, G. H. Glover, and A. L. Reiss. Fast algorithms for mining association rules. *NeuroImage*, 54(4):2808–2821, 2001.

[12] C. Hock, K. Villringer, F. Mller-Spahn, R. Wenzel, H. Heekeren, S. Schuh-Hofer, M. Hofmann, S. Minoshima, M. Schwaiger, U. Dirnagl, and A. Villringer. Decrease in parietal cerebral hemoglobin oxygenation during performance of a verbal fluency task in patients with alzheimer's disease monitored by means of near-infrared spectroscopy (nirs)–correlation with simultaneous rcbf-pet measurements. *Brain research*, 755(2):293–303, May 1997.

[13] H. Ichikawa, J. Kitazono, K. Nagata, A. Manda, K. Shimamura, R. Sakuta, M. Okada, and M.K. Yamaguchi. Novel method to classify hemodynamic response obtained using multi-channel fnirs measurements in to two groups: exploring the combinations of channels. *Frontiers in Human Neuroscience*, 8, July 2014.

[14] H. Song, W. Du, X. Yu, W. Dong, W. Quan, W. Dang, H. Zhang, J. Tian, and T. Zhou. Automatic depression discrimination on fnirs by using general linear model and svm. In *2014 7th International Conference on Biomedical Engineering and Informatics (BMEI)*, volume 8, pages 278–282, Oct 2014.

[15] The Management of Post-Traumatic Stress Working Group. *VA/DoD Clinical Guideline, 2010. Management of Post-Traumatic Stress*. Department of Veterans Affairs - Department of Defense, VA, Washington, DC, 2010. An optional note.

[16] T. Wolfers, J.K. Buitelaar, C.F. Beckmann, B. Franke, and Marquand A.F. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*, 57:328 – 349, 2015.

[17] R. Hosseini, B. Walsh, F. Tian, and S. Wang. An fnirs-based feature learning and classification framework to distinguish hemodynamic patterns in children who stutter. *IEEE Transactions on Neural Systems & Rehabilitation Engineering*, 26(6):1254 – 1263, April 2018.

[18] H. Peng, C. Ding, and L. Fulmi. Feature selection based on mutual information; criteria of max- dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[19] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Center for Evolutionary Medicine and Informatics The Biodesign Institute Arizona State University, 2011.

[20] D. Wang, F. Nie, and H. Huang. Feature selection via global redundancy minimization. *IEEE Transactions on Knowledge and Data Engineering*, 27(10):2743–2755, Oct 2015.

[21] T. J. Huppert, S. G. Diamond, M. A. Franceschini, and D. A. Boas. Homer: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Appl. Opt.*, 48(10):D280–D298, Apr 2009.

[22] S. Brigadoi, F. Scarpa, S. Cutini, P. Scatturin, R. Dell'Acqua, M. Zorzi, and G. Sparacino. Hemodynamic response estimation from fnirs signal through a modeling approach exploiting the reference channel. In *2012 6th International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, volume 3, pages 661–664, May 2012.

[23] S. Wang, J. Gwizdka, and W.A. Chovalitwongse. Using wireless eeg signals to assess memory workload in the n-back task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435, June 2016.

[24] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[25] P. Pudil, J. Novoviov, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119 – 1125, 1994.

[26] N.R. Draper and H. Smith. *Applied Regression Analysis*, pages 307–312. Hoboken, NJ, Wiley-Interscience, 2nd edition edition, 1998.

[27] Hall M. *Correlation-based feature selection for machine learning*. PhD Thesis. New Zealand: Department of Computer Science, Waikato University, 1999.

[28] Yu L. and Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Maching Learning Research*, pages 1205–1224, 2004.

[29] Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, 2014.

[30] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[31] Jie Gui, Zhenan Sun, Shuiwang Ji, Dacheng Tao, and Tieniu Tan. Feature selection based on structured sparsity: A comprehensive study. *IEEE transactions on neural networks and learning systems*, 2017.

[32] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[33] Lin-bo Qiao, Bo-feng Zhang, Jin-shu Su, and Xi-cheng Lu. A systematic review of structured sparse learning. *Frontiers of Information Technology & Electronic Engineering*, 18(4):445–463, 2017.

[34] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2001.

[35] J. R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, Jan 2014.

[36] N. Kwak and C.H. Choi. Input feature selection by mutual information based on parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(12):1667–1671, 2002.

[37] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, Feb 2009.

[38] G. Fung and O. L. Mangasarian. Proximal support vector machine classifiers. In F. Provost and R. Srikant, editors, *Proceedings KDD-2001: Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, CA*, pages 77–86, New York, 2001. Asscociation for Computing Machinery. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/01-02.ps.

[39] B. Blankertz, .G Curio, and K. Muller. Classifying single trial eeg: towards brain computer interfacing. *Advances in Neural Information Processing Systems*, 14(2):157–164, 2002.

[40] D. Garrett, D.C. Peterson, Anderson, and M. Thau. Comparison of linear, nonlinear, and feature selection methods for eeg signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):141–144, 2003.

[41] M. Kaper, P. Meinicke, U Grossekathoefer, T. Lingner, and H. Ritter. Bci competition 2003-data set iib: support vector machines for the p300 speller paradigm. *IEEE Transactions on Biomedical Engeneering*, 51(6):1073–1076, 2004.

[42] M. Uga, I. Dan, T. Sano, H. Dan, and E. Watanabe. Optimizing the general linear model for functional near-infrared spectroscopy: an adaptive hemodynamic response function approach. *Neurophotonics*, 1, 2014.

[43] M. R. Bhutta and K. S. Hong. Classification of fnirs signals for deception decoding using lda and svm. In *2013 13th International Conference on Control, Automation and Systems (ICCAS)*, volume 8, pages 1776–1780, Oct 2013.

[44] B. Xu, Y. Fu, L. Miao, Z. Wang, and H. Li. Classification of fnirs data using wavelets and support vector machine during speed and force imagination. In *2011 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1224–1229, Dec 2011.

[45] R. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Wiley-Interscience, 2000.

CHAPTER 5

Conclusions

In this section we draw some conclusions based on the application of the proposed network construction method and feature learning framework. More details can be found in each article in the previous chapters.

The advantage of structure/network learning from the MVTS based on MST is detecting the strongest connections as a unique sub-graph from the underlying network with unique edge-weights which are estimated based on the linear conditional Gaussian and maximum likelihood. The methods is more effective in smaller networks and longer durations. Under some circumstances, it outperforms the lag-based methods like Granger causality. However, based on the nature of the MST algorithm, it performs poorly when estimating the network from a ground-truth structure with cyclic, backward, many and shared-input connections. Another advantage is that the weights can be used as extracted features from the MVTS and can be applied for machine learning purposes like classification.

MISS was significantly promising as a feature selection technique, it outperformed mRMR, LASSO, SGL and was capable of finding the most sparse set of discriminative features. Moreover, the proposed method facilitated finding the region of interest (ROI) on the brain for a specific brain disease with a data-driven approach for the cases that we do not have access to any prior-knowledge-based (ROI). The sparsely selected features known as biomarkers were used to detect the discriminative patterns between control and non-control class. The sparse set of biomarkers makes

the final model be more interpretable and generalizable, moreover, the model could facilitate diagnostics and tracking of the patients' treatment.