University of Cincinnati

Date: 10/21/2015

I, Lirong Tan, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Computer Science & Engineering.

It is entitled:
Identification of Disease Biomarkers from Brain fMRI Data using Machine Learning Techniques: Applications in Sensorineural Hearing Loss and Attention Deficit Hyperactivity Disorder

Student's name: Lirong Tan

This work and its defense approved by:

Committee chair: Long Lu, Ph.D.
Committee member: Raj Bhatnagar, Ph.D.
Committee member: Scott Holland, Ph.D.
Committee member: Jaroslaw Meller, Ph.D.
Committee member: Ali Minai, Ph.D.
Identification of Disease Biomarkers from Brain fMRI Data using Machine Learning Techniques: Applications in Sensorineural Hearing Loss and Attention Deficit Hyperactivity Disorder

A dissertation submitted to the Graduate School of University of Cincinnati in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science of the College of Engineering and Applied Science

by

Lirong Tan

M.Phil., Computer Science, City University of Hong Kong, Hong Kong, 2010
B.S., Life Science, Wuhan University, Wuhan, China, 2008

Advisor: Dr. Jason L. Lu
October, 2015
Abstract

Machine learning techniques have received intense attention in the neuroimaging field during the past decade. Nevertheless, machine learning analysis of Magnetic Resonance Imaging (MRI) data is still in its early stage, especially for functional MRI (fMRI) images. The research in this thesis was focused on the automatic diagnosis and prognosis of brain disorders using fMRI data. We had two major research goals: (1) we attempted to design novel machine learning models for the automatic diagnosis of brain disorders as well as for the prognosis of disease treatment; (2) we would like to derive disease biomarkers, which might help us to better understand the disorder and brain function.

The first brain disorder that was investigated in this thesis was sensorineural hearing loss (SNHL). Children with SNHL receive little or no benefits from hearing aids and face challenges in developing language abilities. Cochlear implant (CI) is an electronic device that can be inserted into the cochlea for direct stimulation of the auditory nerve and has been demonstrated to be effective in restoring hearing in patients suffering from SNHL. However, variability in speech and language outcomes for CI remains high and individual outcomes may be difficult to predict. To tackle this problem, our collaborators have acquired fMRI images from a group of SNHL patients before the surgery as well as a group of age-matched controls with normal hearing. Using this dataset, we first developed a model for the classification of normal hearing infants vs. hearing impaired infants with SNHL, which provided a proof of principle that it is possible to accurately determine the functional, developmental status of the central auditory system in congenitally hearing impaired children based on MR images alone by utilizing machine learning techniques. Second, we trained a machine learning model to predict CI outcomes, which has validated the hypothesis that pre-implant cortical activation patterns
revealed by fMRI during infancy correlate with language performance two years after cochlear implantation. This result might encourage further studies in this field to develop a reliable predictive model based on a larger training set that can eventually be applied in the clinical setting to provide specific prognostic information to patients considering cochlear implantation.

The other disease that was studied in this dissertation was Attention Deficit Hyperactivity Disorder (ADHD). The dataset for this study came from the ADHD-200 global competition, which made hundreds of anatomical and functional images publicly available. For this dataset, we proposed a novel feature type called functional volume, which quantified the volume of brain tissue that was actually active during fMRI imaging. Functional volume was a counterpart of anatomical volume, which was measured from structural MR images. We applied them to the automatic diagnosis of ADHD. It turned out that functional volume performed much better than anatomical volume. Besides, our model based on functional volume also outperformed other relevant models that were published previously for the ADHD-200 dataset. The brain regions highlighted by our model might provide some new perspectives about the brain regions affected by ADHD.
Acknowledgement

First of all, I would like to express my sincere appreciation to Dr. Jason Lu, who is the supervisor of my PhD study. He supported my dissertation research, both intellectually and financially. Without his support, I would not be able to complete my PhD study. We discussed research progress every week, and I could always get help from him whenever I needed it. Most importantly, he introduced me to the field of Data Science and helped me to find my intense interests in this field. Dr. Lu is open-minded and always willing to listen to opinions that are different from his. I enjoyed working with him very much.

I would also like to convey my heartfelt gratitude to Dr. Scott Holland. Without his guidance, I would not be able to find a dissertation topic so quickly. He spent countless hours discussing the research work with me. He provided precious guidance to my work, and offered significant help in drafting papers and addressing review comments. His knowledge and insights were extremely valuable and greatly enriched my work. It was a great opportunity and a unique experience to work with him. Dr. Holland was the Principal Investigator of the original NIH grant (R01-DC07186) that produced the brain imaging data from infants in the cochlear implant study and the normal hearing controls. Dr. Holland also provided financial support for my graduate work.

My appreciation also goes to the other committee members, Dr. Jarek Meller, Dr. Raj Bhatnagar and Dr. Ali Minai, for their guidance and comments on my research work. They were always so supportive for the committee meetings, although their schedules were usually very tight. Besides, Dr. Meller and Dr. Bhatnagar are also my academic advisors. They offered great help to me when I encountered problems in courses and tuitions. Without their help, I would not be able to focus on my research work.
Finally, I am also thankful to my supportive friends and loving family. I thank them all for their continuous encouragement and constant support throughout my study and life.
Table of Contents

CHAPTER 1   INTRODUCTION ................................................................................................................. 9
  1.1 Overview ........................................................................................................................................ 9
  1.2 Contributions ................................................................................................................................. 10
  1.3 Organization of this dissertation.................................................................................................. 13

CHAPTER 2   BACKGROUND .................................................................................................................. 15
  2.1 BOLD fMRI .................................................................................................................................... 15
  2.2 Preprocessing for fMRI images ..................................................................................................... 15
  2.3 Challenges in machine learning analysis of fMRI data ............................................................... 16
  2.4 A brief literature review for published feature extraction algorithms .................................. 17
    2.4.1 Activation maps for task fMRI ............................................................................................ 17
    2.4.2 Activity maps for resting-state fMRI .................................................................................. 19
    2.4.3 Functional Connectivity ....................................................................................................... 20
    2.4.4 Network Topological Properties ......................................................................................... 21
  2.5 Model learning .............................................................................................................................. 22

CHAPTER 3   COMBINED ANALYSIS OF SMRI AND FMRI IMAGING DATA
  PROVIDES ACCURATE DISEASE MARKERS FOR HEARING IMPAIRMENT ........................................ 24
  3.1 Introduction ................................................................................................................................. 24
  3.2 Materials and Methods ................................................................................................................ 28
    3.2.1 Data acquisition and preprocessing ................................................................................... 28
    3.2.2 Feature extraction and model learning based on structural MR images ......................... 33
    3.2.3 Feature extraction and model learning based on functional MR images ....................... 39
    3.2.4 Integrated model ................................................................................................................ 44
    3.2.5 Validation of the classifier ................................................................................................ 45
  3.3 Results .......................................................................................................................................... 46
    3.3.1 Classifier performance .......................................................................................................... 46
    3.3.2 Feature selection in sMRI analysis ....................................................................................... 47
    3.3.3 Stability of feature selection in fMRI analysis .................................................................... 49
    3.3.4 Discriminative brain regions ............................................................................................. 50
  3.4 Discussion ...................................................................................................................................... 53
  3.5 Conclusion .................................................................................................................................... 58

CHAPTER 4   A COMPREHENSIVE COMPARISON OF BOW APPROACH
  WITH TWO TRADITIONAL FEATURE EXTRACTION APPROACHES .............................................. 59
  4.1 Overview ....................................................................................................................................... 59
  4.2 Data .............................................................................................................................................. 59
CHAPTER 5 A SEMI-SUPERVISED SVM MODEL FOR PREDICTING THE LANGUAGE OUTCOMES FOLLOWING COCHLEAR IMPLANTATION BASED ON PRE-IMPLANT BRAIN FMRI IMAGING

5.1 Introduction ........................................................................................................... 70

5.2 Materials and Methods .......................................................................................... 74
  5.2.1 Participants ......................................................................................................... 74
  5.2.2 Cochlear implantation outcomes ......................................................................... 74
  5.2.3 MRI data acquisition & preprocessing ................................................................. 78
  5.2.4 Feature extraction .............................................................................................. 78
  5.2.5 Supervised SVM model ..................................................................................... 79
  5.2.6 Semi-supervised SVM model ............................................................................ 79
  5.2.7 Feature selection .............................................................................................. 80
  5.2.8 Predicting new subjects .................................................................................... 80
  5.2.9 Model evaluation .............................................................................................. 81

5.3 Results .................................................................................................................... 85
  5.3.1 Supervised model .............................................................................................. 85
  5.3.2 Semi-supervised model with automatic parameter selection ............................. 85
  5.3.3 Semi-supervised model with fixed parameter r .................................................. 87
  5.3.4 Why did semi-supervised model perform better? ............................................. 88
  5.3.5 Blind test on normal hearing infants .................................................................. 95
  5.3.6 Semi-supervised model by adding the unlabeled samples one by one .............. 95
  5.3.7 Semi-supervised model with artificial unlabeled samples ................................ 97
  5.3.8 Model with feature selection ............................................................................ 99
  5.3.9 Validate the feature selection algorithm ............................................................ 100
  5.3.10 Stability of feature selection .......................................................................... 101
  5.3.11 Discriminative brain regions .......................................................................... 103

5.4 Discussion .............................................................................................................. 106

5.5 Conclusion .............................................................................................................. 117
CHAPTER 6 FUNCTIONAL VOLUME OUTPERFORMS ANATOMICAL VOLUME FOR AUTOMATIC DIAGNOSIS OF ATTENTION DEFICIT HYPERACTIVITY DISORDER ............................................. 119

6.1 Introduction ........................................................................................................ 120

6.2 Materials and Methods ......................................................................................... 126

6.2.1 Participants ........................................................................................................ 126
6.2.2 Preprocessing of images .................................................................................... 126
6.2.3 Features ............................................................................................................ 127
6.2.4 Feature selection ................................................................................................ 131
6.2.5 Classification model learning ............................................................................ 132
6.2.6 Model evaluation ............................................................................................... 132
6.2.7 Identification of Important Features ................................................................. 133
6.2.8 Linear Regression Analysis .............................................................................. 134

6.3 Results .................................................................................................................. 135

6.3.1 Linear Regression Analysis Results .................................................................. 135
6.3.2 Classification performance without feature selection ....................................... 138
6.3.3 Classification performance with feature selection .............................................. 140
6.3.4 Important features ............................................................................................ 141

6.4 Discussion ............................................................................................................ 144

6.5 Conclusion ............................................................................................................ 149

CHAPTER 7 CONCLUSION & FUTURE DIRECTIONS .............................................. 151
Chapter 1 Introduction

1.1 Overview

Machine learning is the science of making computers recognize patterns in data and then apply the pattern rules to new unseen data. Machine learning tasks can be generally classified into three major categories, namely classification, regression, and clustering. Classification has received a lot of interests in neuroimaging field for the automatic diagnosis of brain disorders and cognitive state decoding. The basic idea of classification is to learn a function \( f \) to map the data points in the feature space to the class label space. The function \( f \) is also known as classifier. Take the binary classification for example. The input for the analysis is a training set \( D = \{ (X_1, y_1), \ldots, (X_M, y_M) \} \), where \( M \) is the number of training samples, \( X_k = (x_1, \ldots, x_N) \) and \( y_k \in \{-1, +1\} \) are the feature vector and class label for the \( k \)-th sample respectively, \( N \) is the number of features. The output of the analysis is the classifier \( f \). The process to derive the classifier \( f \) is called model training/learning. With the trained/learned classifier \( f \), a new sample with feature vector \( X_{\text{new}} \) whose class label is unknown will be predicted to have class label \( \hat{y} \) (\( \hat{y} = f(X_{\text{new}}) \)).

Feature extraction, the construction of feature vector \( (X) \), is a key step before model learning. As the saying goes, garbage in, garbage out. The goodness of the features has a significant influence on the performance of the classifier. Normally, we would consider feature extraction as part of the machine learning analysis.

Functional Magnetic Resonance Imaging (fMRI) is a noninvasive imaging technology to measure the brain activities over a period of time. fMRI data are 4-dimensional, with the first three dimensions representing the human brain space, and the last dimension representing time.
In the recent decade, machine learning techniques have been widely used for the analysis of fMRI data. The complex format of fMRI data makes the feature extraction even more critical for the machine learning analysis. In this dissertation, we investigated the problem of feature extraction from fMRI data under the background of three classification problems with two datasets. Our first fMRI dataset was collected by our collaborators from a group of infants for the study of sensorineural hearing loss. Using this dataset, we proposed a Bag-of-Words algorithm for feature extraction, and applied it to two classification problems: classification of normal hearing infants from hearing impaired infants, and classification of the outcomes from the cochlear implantation surgery. The other dataset we used in this dissertation was ADHD-200, which was a publicly available dataset for the study of Attention Deficit Hyperactivity Disorder (ADHD). For this dataset, we have proposed a novel type of feature called functional volume for fMRI images, and applied it to the automatic diagnosis of ADHD.

1.2 Contributions

The main contributions of this thesis include the following:

(1) Inspired by the idea of Bag-of-Words (BoW), we proposed a novel approach for extracting features from fMRI data. Although the BoW approach has been widely used for document classification, it is not straightforward to move from documents to fMRI data, given their very different data formats. Therefore, we consider our approach as a new approach rather than a simple application or modification of the original BoW approach. Compared with two traditional approaches for feature extraction from fMRI data, our novel approach exhibited various advantages, such as less sensitive to registration errors, lower feature-to-sample ratio, and better characterization of the activation patterns.
(2) We designed a novel two-layer model to integrate structural MRI (sMRI) data with fMRI data. Integration of different types of data, e.g. data from multiple modalities, has been demonstrated to be more powerful for classification of different groups of individuals (Fan et al., 2007, Fan et al., 2008, Tosun et al., 2010, Wang et al., 2012b). However, how to implement such integrations in the best way remains to be explored (Orru et al., 2012). Traditionally, features from different types of data were concatenated and a single classifier was trained. Clearly, this traditional approach would make the feature-to-sample ratio even higher. By contrast, our two-layer model was much more flexible. It can integrate as many types of data as possible, without worrying about the high dimensionality of the feature space.

(3) We developed a machine learning model for the classification of infants with normal hearing versus infants with hearing loss. Our classification model highlighted some discriminative brain regions as biomarkers for sensorineural hearing loss. Some of the highlighted brain regions were aligned with our original hypothesis, and the others may provide a new understanding of the brain function and of the disorder itself.

(4) We built a machine learning model to predict whether or not a cochlear implant (CI) candidate will develop effective language skills within two years after the CI surgery by using the pre-implant brain fMRI data from the candidate. Such kind of prognostic information is extremely useful but currently not available to clinicians from any other means, especially given the high costs and risks of the CI surgery. The work in this dissertation validated the hypothesis that pre-implant cortical activation patterns revealed by fMRI during infancy correlate with language performance two years after cochlear implantation, while our previous simple linear regression analysis failed to do so. Results in this dissertation may encourage further studies in this direction, and promotes a predictive model that can be ultimately used in clinical. This study
also highlighted the discriminative power of the brain activation pattern stimulated by human speech in comparison with noise. This is actually a highly significant discovery from the machine that humans have not been able to make. This discovery may provide some clues for the future design of fMRI protocol.

(5) Our machine learning model highlighted two brain regions as biomarkers for the prediction of CI surgery outcomes. One of these two brain regions is aligned with our original hypothesis, while previous univariate correlation analysis or multivariate regression analysis failed to provide persuasive predictive value by looking at a region of interest in this part of the brain alone. The other brain region, whose predictive power was underestimated according to our original hypothesis, appeared to be surprising at the first thought. Through a deeper investigation, we found substantial number of scientific publications supporting this discovery and achieved a better understanding of the brain function. Our work demonstrated the advantage of machine learning techniques, which can automatically detect the predictive features and draw our attention to features that are important but beyond our prior knowledge.

(6) We have initially proposed the concept of functional volume, which is a counterpart of anatomical volume. Functional volume measures the volume of brain tissue that is actually active during fMRI imaging, and demonstrated higher discriminative power for the classification of ADHD patients vs. healthy controls, when compared with anatomical volume. Our classification model using functional volume also outperformed other relevant models published previously. The brain regions highlighted by our classifier might provide some new insights into the brain regions affected by ADHD.

(7) We designed a two-step feature selection algorithm, which has integrated the advantages of both variance threshold algorithm and Recursive Feature Elimination (RFE)
algorithm. This two-step algorithm helped us significantly improved the accuracy for the automatic diagnosis of ADHD.

(8) Our study pointed to the weakness of current algorithm for the spatial normalization of fMRI images. We also call special attentions for the comparison of (fractional) amplitude of low frequency fluctuations (ALFF/fALFF) between different groups of individuals, especially given there is brain shrinkage for one group of individuals.

(9) This dissertation is problem oriented rather than algorithm oriented. We did not study a specific algorithm, and there was no complicated algorithm either. We designed new algorithms to solve specific problems, and our proposed algorithms were shown to outperform relevant algorithms in literature. For the project of automatic diagnosis of ADHD, we started out with existing feature extraction approaches and applied advanced model learning algorithms, such as a hierarchical model based on the deep learning idea, but failed to develop an accurate classification model. On the other hand, the combination of our functional volume features with the ordinary linear SVM model achieved a significant improvement in the classification accuracy. This result highlighted the importance of feature extraction for the success of machine learning analysis, and might provide some hints to those who are new to machine learning.

1.3 Organization of this dissertation

This dissertation is organized as follows:

Chapter 2 provides the background information about fMRI and a brief literature review about the approaches for feature extraction from fMRI data.

Chapter 3 investigates the classification of normal hearing infants vs. hearing impaired infants.
Chapter 4 is devoted to the comparison of our Bag-of-Words algorithm with two traditional feature extraction algorithms.

Chapter 5 presents the project of predicting the cochlear implantation outcomes.

Chapter 6 describes the project of automatic diagnosis of attention deficit hyperactivity disorder (ADHD).

Finally, Chapter 7 provides a conclusion to the whole dissertation.
Chapter 2 Background

2.1 BOLD fMRI

fMRI is a noninvasive imaging technology to measure brain activities. The primary form of fMRI uses the blood-oxygen-level dependent (BOLD) contrast. When a neuron is activated, it requires extra oxygen to perform the tasks, e.g., processing a visual input. The required oxygen is provided by the blood cells in the capillaries around the neuron. As a result, the blood flow in this active brain area will increase and this brain area switches from oxygen-poor to oxygen-rich. Furthermore, oxygen-rich blood and oxygen-poor blood have opposite magnetization properties, and therefore the change in cerebral blood flow can be detected by the magnetic resonance imaging techniques. Thus, an active brain area with high blood flow has high imaging signal, while an inactive brain area with low blood flow has low imaging signal. That’s how the fMRI technology is used to map the neural activities.

While the structural MRI (sMRI) only acquires one image for the brain anatomy, functional MRI usually monitors the brain activities in a period of time, e.g., a brain volume is acquired to map the brain activities at each time point. There are two types of fMRI, namely resting-state fMRI and task fMRI. Resting-state fMRI data are acquired while the candidate is under rest without performing any particular task. Task fMRI imaging requires the candidate to perform some tasks, e.g., listening to a short story, during the acquisition of fMRI images, although there could also be rest periods for contrast purposes.

2.2 Preprocessing for fMRI images

The fMRI image preprocessing is a complex process including two major steps, namely temporal processing and spatial processing. The toolbox Statistical Parametric Mapping (SPM,
http://www.fil.ion.ucl.ac.uk/spm/) (Friston et al., 2006) or Automated of Functional NeuroImages (AFNI, http://cnl.web.arizona.edu/afni.htm) (Cox, 1996) is the program used for the processing of fMRI data. The first step for preprocessing is slice timing correction. Since the brain is scanned slice by slice during the image acquisition, there will be time differences among different slices. Slice timing correction is applied to correct the time differences of each slice and to correct the data from each slice for intensity variations. The spatial processing mainly includes motion correction and spatial normalization. Due to the motion of the candidate during fMRI image acquisition, the brain volumes might not be aligned with each other. Therefore, “realign” is performed to register all the brain volumes to the first volume in the fMRI time series. Spatial normalization is used to register individual images to a standard template brain. The normalization is usually achieved through the normalization of sMRI image. The functional images are initially registered to the sMRI image. The sMRI image is then mapped to a standard template to produce the spatial transformation parameters, which are subsequently applied to the functional images. In this way, both the sMRI and fMRI images are aligned to the standard template. Since the images of all the subjects are aligned to the standard template, the images are also aligned among different subjects, which facilitate subsequent group analysis and machine learning analysis. For a more detail description of the preprocessing steps, please refer to the SPM or AFNI manual.

2.3 Challenges in machine learning analysis of fMRI data

The machine learning analysis for fMRI data has two major applications. One is the cognitive state classification, which is also known as “mind-reading”. For example, a machine learning model could be trained for lie detection or decoding the stimuli that were given to the participant during imaging. The other major application is automatic disease diagnosis and biomarker
detection. The machine learning analysis has already been applied to various diseases. Our study in this thesis also falls into this category.

There are a couple of challenges for the machine learning analysis of fMRI data. First, the preprocessing procedure is complicated, which might prevent the machine learning experts from jumping into this field. Besides, the fMRI dataset usually has very limited sample size and is highly noisy, which makes it difficult to train an accurate model, especially given the very high dimensionality of fMRI data, usually with ~20,000 voxels (3D pixel) and ~100 time points. How to extract discriminative features from this huge space, how to reduce the feature-to-sample ratio and increase the signal-to-noise ratio are questions waiting to be investigated. As a start in this thesis, we attempted to explore those questions by improving the feature extraction approaches.

2.4 A brief literature review for published feature extraction algorithms

As we mentioned above, the fMRI data is 4-dimensional. Feature extraction from such a complex data format is the major challenge for the machine learning analysis. The popular feature extraction algorithms in existing literature were reviewed as below.

2.4.1 Activation maps for task fMRI

Brain activations were widely used as features for task fMRI data. There are two main ways to calculate brain activation. For mind-reading tasks, one or a couple of continuous images in the fMRI time series corresponds to a sample in the machine learning analysis, in which each sample was actually a stimulus or a brain status. For example, a letter “A” might be shown on the screen as a stimulus to the participant during imaging. This letter might last for 5 seconds, during which two fMRI images were acquired. In such cases, the image itself (one image corresponding to a sample) or the collapsed image by averaging multiple images (multiple continuous images corresponding to a sample) is used as the activation map for the sample. Optionally, the percent
signal change relative to a baseline condition can be calculated to replace the original images (Shinkareva et al., 2008, Chang et al., 2011, Andersen et al., 2012).

For classification problems to separate patients from healthy controls, however, each participant is treated as a sample. Therefore, the whole fMRI time series corresponds to a single sample, which differs from the above mind-reading tasks. For patient classification problems, the general linear model (GLM) is usually employed to generate contrast maps (Worsley et al., 2002), which are the activation maps used in subsequent machine learning analysis. The contrast maps could be a beta map, T map or Z map.

With the activation maps, the intensities of the image voxels are concatenated to form a feature vector for each sample, with each single voxel becoming a feature. To reduce the dimensionality of feature space, feature selection is usually performed before model learning. Different studies employed different approaches to reduce the dimensionality of feature space. (Mitchell et al., 2004, De Martino et al., 2008, Kuncheva and Rodriguez, 2010) employed two-sample t-test or Wilcoxon rank-sum test to select the voxels most discriminative between different classes, e.g., patients vs. healthy controls. (De Martino et al., 2008, Kuncheva and Rodriguez, 2010, Ryali et al., 2010) selected the most activated voxels because those voxels were hypothesized to represent the actual activation pattern with high signal-to-noise ratio. (Shinkareva et al., 2008, Chang et al., 2011) selected the voxels with stable responses to the external stimuli. The aforementioned univariate filtering approaches could be used as an initial step for dimensionality reduction, the multivariate feature selection algorithm recursive feature elimination (RFE) might be optionally used as a second step to further reduce the dimensionality (De Martino et al., 2008, Kuncheva and Rodriguez, 2010). In addition to the approaches for selecting a subset of features, Principal Component Analysis (PCA) might be used as an
alternative way to reduce the dimensionality of feature space (Ford et al., 2003, Kuhn et al., 2010).

While the above studies used each voxel as a feature, a couple of studies tried to cluster the voxels into regions of interest (ROIs). Davatzikos et al. divided the brain into 560 cubes, each having a size of 16 mm × 16 mm × 16 mm. The mean activation intensity within each cube was calculated and used as features (Davatzikos et al., 2005). This ROI-based approach has significantly reduced the dimensionality of feature space, but was rarely adopted by previous studies. This might be because there was not a good way to define the ROIs.

### 2.4.2 Activity maps for resting-state fMRI

In recent years, resting-state fMRI has become highly popular in the neuroimaging field. Spontaneous low frequency fluctuations in blood oxygen level dependent (BOLD) activity are a fundamental feature of the brain at rest. The relative magnitude of these fluctuations is usually measured by the amplitude of low frequency fluctuations (ALFF) (Zang et al., 2007) or fractional amplitude of low frequency fluctuations (fALFF) (Zou et al., 2008). ALFF is a voxel-by-voxel calculation of the power spectrum of the BOLD fMRI time series. fALFF is the ratio of power spectrum of low-frequency (0.009–0.08 Hz) to that of the entire frequency range. Generally, ALFF/fALFF maps are considered as activity maps for resting-state fMRI, and widely used for the automatic diagnosis of various diseases.

Feature extraction approaches from ALFF/fALFF maps are generally similar to the approaches used for activation maps of task fMRI as described in last section. There are two major categories. The first category uses each single voxel as a feature, and the value of the feature is measured as the ALFF/fALFF coefficient for the corresponding voxel (Cheng et al., 2012, Long et al., 2014, Savio et al., 2014, Valli and Jiji, 2014, Chyzhyk et al., 2015). The other
category is based on ROIs. Each ROI becomes a feature, and the value of the feature is calculated as the mean ALFF/fALFF within this ROI. For the definition of ROIs, the Automated Anatomical Labeling (AAL) atlas is used most frequently (Dai et al., 2012b, Long et al., 2012, Liu et al., 2015). The AAL atlas segments the cerebrum into 90 regions (45 in each hemisphere) and segments the cerebellum into 26 regions (9 in each cerebellar hemisphere and 8 in the vermis), based on the anatomy of the brain. Recently, Craddock et al. proposed a spatially constraint normalized cut algorithm to segment the brain into spatially contiguous and functionally coherent ROIs (Craddock et al., 2012). This clustering software has been made publicly available at: [http://www.nitrc.org/projects/cluster_roi/](http://www.nitrc.org/projects/cluster_roi/). It could be applied to generate a functional atlas, which may be subsequently used for ALFF/fALFF maps (Sato et al., 2012b).

### 2.4.3 Functional Connectivity

The functional connectivity features were especially popular for resting-state fMRI data. Generally, the whole-brain was first parcellated into a set of brain regions, namely ROIs. The mean time course for each ROI was calculated and a connectivity matrix was generated by calculating the pair-wise Pearson’s correlation coefficient between the mean time courses of the ROIs. Optionally, the regional mean time courses may be decomposed into multiple sub-bands with different frequencies using the discrete wavelet transform or fast Fourier transform (Richiardi et al., 2011, Wee et al., 2012), partial correlation coefficient may be calculated by removing some influencing effects, e.g., age effect in the identification of patients with Alzheimer’s disease, from the original Pearson’s correlation coefficient (Chen et al., 2011), and the correlation coefficients may be further converted to z-scores by applying the Fisher transformation to improve the normality of the correlation coefficients (Anderson et al., 2011, Shirer et al., 2012, Wee et al., 2012). Then, the correlation coefficients in the up-right triangle of
the connectivity matrix were concatenated to form the feature vector, which had a dimensionality of $N \times (N - 1)/2$ where $N$ was the number of ROIs.

Different studies employed different approaches to define the ROIs. Several studies (Shen et al., 2010, Chen et al., 2011, Richiardi et al., 2011, Wee et al., 2012, Zeng et al., 2012) applied the AAL atlas to define the ROIs. Some of the studies used the 90 cerebral regions only, while others used all the 116 regions. One drawback of the AAL atlas is that it does not consider the function of the brain. Voxels within a single AAL region may have different functions with inconsistent time courses. As a result, the mean time course of a region may not represent any of the constituent time courses. Anderson et al. defined 7266 ROIs to form a lattice covering the entire grey matter and calculated the functional connectivity measures for the automatic diagnosis of autism (Anderson et al., 2011, Nielsen et al., 2013). Shirer et al. employed the group independent component analysis (ICA) to define ROIs involved in the intrinsic connectivity networks (Shirer et al., 2012). Compared with the AAL atlas and ROI-lattice, the approach based on ICA requires some manual intervention and fine-tuning. As in (Shirer et al., 2012), they selected only the independent components involved in the intrinsic networks based on their prior knowledge and then thresholded the component maps to generate distinct moderately sized ROIs in the cortex and subcortical grey matter. Finally, the functional atlas generated from Craddock’s clustering algorithm (Craddock et al., 2012) was another option to define ROIs. Cheng et al. applied this functional atlas for the automatic diagnosis of attention deficit hyperactivity disorder (Cheng et al., 2012).

2.4.4 Network Topological Properties

The connectivity matrix can be further thresholded to construct a binary brain network. The topological properties, such as small-worldliness and clustering coefficient, were repeatedly
reported to distinguish between patient group and healthy control group using univariate analysis such as two-sample t-test (Stam et al., 2007, Liu et al., 2008, Wang et al., 2009, Liao et al., 2010, Nomura et al., 2010, Yu et al., 2011, Zhang et al., 2011a, Achard et al., 2012, Zhao et al., 2012). A complete list of network topological measures can be found in (Rubinov and Sporns, 2010). While most studies employed univariate analysis, some studies used the network topological properties as features to train machine learning classifiers to separate patients from healthy controls (Zhang et al., 2011b, Lord et al., 2012, Fekete et al., 2013).

2.5 Model learning

After the feature extraction and optional feature selection, machine learning models were trained using the training samples. Among the various machine learning models, Support Vector Machine (SVM) with a linear kernel received the greatest popularity probably due to its good abilities in handling the high feature-to-sample ratio, which was obviously shared by fMRI datasets. The basic idea of a linear SVM model was provided below.

SVM is originally designed for binary classification. The input for the supervised SVM model training is a training set \( D = \{(X_1, y_1), \ldots, (X_M, y_M)\} \), where \( M \) is the number of training samples, \( X_k = (x_1, \ldots, x_N) \) and \( y_k \in \{-1, +1\} \) are the feature vector and class label for the \( k \)-th sample respectively, \( N \) is the number of features. Our learning objective is to estimate the model \( \hat{y} = wX + b \), where \( w = (w_1, \ldots, w_N) \) is the weight vector and \( b \) is the bias by minimizing the objective function in Equation (2.1).

\[
\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{M} \xi_i
\]

s.t. \( \forall i \in [M]: y_i (wX_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \)
where $C$ is a regularization parameter controlling the trade-off between margin and training errors, $\xi$ is the slack variable. According to previous publications and our experiences, the classification performance is not sensitive to parameter $C$ once $C$ is larger than a certain threshold. We usually set $C$ to be 1, unless we apply a grid search to optimize this parameter.
Chapter 3 Combined Analysis of sMRI and fMRI Imaging Data

Provides Accurate Disease Markers for Hearing Impairment

In this chapter, we developed a robust two-layer classifier that can accurately classify normal hearing (NH) from hearing impaired (HI) infants with congenital sensori-neural hearing loss (SNHL) based on their Magnetic Resonance (MR) images. Unlike traditional methods that examine the intensity of each single voxel, we extracted high-level features to characterize the structural MR images (sMRI) and functional MR images (fMRI). The Scale Invariant Feature Transform (SIFT) algorithm was employed to detect and describe the local features in sMRI. For fMRI, we constructed contrast maps and detected the most activated/de-activated regions in each individual. Based on those salient regions occurring across individuals, the bag-of-words strategy was introduced to vectorize the contrast maps. We then used a two-layer model to integrate these two types of features together. With the leave-one-out cross-validation approach, this integrated model achieved an AUC score of 0.90. Additionally, our algorithm highlighted several important brain regions that differentiated between NH and HI children. Some of these regions, e.g. planum temporale and angular gyrus, were well known auditory and visual language association regions. Others, e.g. the anterior cingulate cortex (ACC), were not necessarily expected to play a role in differentiating HI from NH children and provided a new understanding of brain function and of the disorder itself.

3.1 Introduction

It has been estimated that approximately 1 to 6 infants per 1,000 are born with severe to profound congenital sensori-neural hearing loss (SNHL) (Northern, 1994, Bachmann and Arvedson, 1998, Kemper and Downs, 2000, Cunningham and Cox, 2003). Those children
receive little or no benefit from hearing aids and face challenges in developing language abilities due to their inability to detect acoustic-phonetic signals, which are essential for hearing-dependent learning. Cochlear implantation is a surgical procedure that inserts an electronic device into the cochlea for direct stimulation of the auditory nerve and has been demonstrated to be effective in restoring hearing in patients suffering from SNHL. Statistical data from National Institute on Deafness and Other Communication Disorders (NIDCD) indicate that approximately 28,400 children in United States have received a cochlear implant (CI) as of December 2010. While many congenitally deaf CI recipients achieve a high degree of accuracy in speech perception and develop near-normal language skills, about 30% of the recipients do not derive any benefit from the CI (Niparko et al., 2010). A deeper understanding of hearing loss and better characterization of the brain regions affected by hearing loss will help reduce the high variance in CI outcomes and result in a more effective treatment of children with hearing loss.

In recent years, Magnetic Resonance (MR) images have been used to study neurological disorders and brain development in children, such as reading and attention problems, traumatic brain injury, hearing impairment, perinatal stroke and other conditions (Tillema et al., 2008, Leach and Holland, 2010, Smith et al., 2011, Tlustos et al., 2011, Horowitz-Kraus and Holland, 2012). Brain MRI scans have revealed significant differences between Hearing Impaired (HI) and Normal Hearing (NH) children. Jonas et al. reviewed a total number of 162 patients’ structural MRI scans, and detected 51 abnormalities in 49 patients. Those abnormalities included white matter changes, structural or anatomical abnormalities, neoplasms, gray matter changes, vasculitis and neuro-metabolic changes (Jonas et al., 2012). Similar studies have showed consistent results (Lapointe et al., 2006, Trimble et al., 2007, Smith et al., 2011). Furthermore, functional MRI studies have demonstrated that the activation pattern of HI is different from that
of NH during certain scanning tasks (Scheffler et al., 1998, Bilecen et al., 2000, Tschopp et al., 2000, Patel et al., 2007, Propst et al., 2010). For example, Propst and colleagues studied the activation pattern of HI with narrowband noise and speech-in-noise tasks (Propst et al., 2010). In the narrowband noise task, they found that HI children had weaker activation in the auditory areas when compared to NH children. Meanwhile, NH also activated auditory association areas and attention networks, which were not detected in HI children. In the speech-in-noise task, HI children activated the secondary auditory processing areas only in the left hemisphere, rather than bilaterally as is typical of NH. Recently, we have tried to use the activation in the primary auditory cortex (A1) to predict CI outcomes. A strong correlation (linear regression coefficient, R=0.88) was detected between the improvement in post-CI hearing threshold and the amount of activation in the A1 region before CI (Patel et al., 2007). Despite these recent advances, it remains unclear whether these structural and functional abnormalities are sufficient to distinguish HI from NH individuals.

In this study, we set out to investigate whether we can accurately classify HI from NH individuals based on MR images alone by utilizing machine learning techniques. We have trained three classifiers, one based on structural MR (sMRI) images, another based on functional MR (fMRI) images, and a third that integrates sMRI and fMRI images. While traditional methods utilize voxel-based morphometric (VBM) features, in which each single voxel serves as an independent feature, we extracted high-level features to characterize the 3D images. Specifically, we employed the Scale Invariant Feature Transform (SIFT) algorithm to detect and describe local features in sMRI and extracted region-level features to represent the functional contrast maps. Based upon the extracted features, SVM classifiers were trained to separate HI from NH.
The SIFT algorithm was first proposed by Lowe for object recognition (Lowe, 1999). Since then, it has been widely used in the computer vision field. Basically, the SIFT algorithm detects blob-like image components and calculates a vector to describe each of these components. Each vector becomes a SIFT feature. The set of SIFT features extracted from an image contain important characteristics of this image and can be used for subsequent analysis, e.g. object recognition, gesture recognition and etc. In this study, we employed the SIFT algorithm to extract SIFT features from brain structural MR images, and devised an approach for the automatic classification of NH vs. HI based on the SIFT features.

There are three levels of significance for this study. First of all, we convincingly demonstrate that hearing loss can be accurately diagnosed based on MR images alone. Secondly, brain regions identified by the classifiers enable us to better understand hearing loss, and may serve as valuable indicators for the CI outcome and facilitate follow-up treatment post-CI (Jonas et al., 2012). Finally, our algorithm can be easily extended to assist in diagnosing other disorders affecting children’s brains, e.g., speech sound disorders of childhood, leading to a path for improving child health.

The organization of this chapter is as follows. In Materials and Methods, we describe in sequence the data sources and the preprocessing procedures, the methods of analyzing sMRI and fMRI images, the integrative model that combines these two methods, and the validation of our classifiers. In Results, we compare the classification performance of the sMRI classifier, fMRI classifier and the combined classifier, and assess the stability of feature selection as well as the discriminative power of features. Finally, in Discussion, we summarize the present work, highlight the significance of our approach, and discuss the limitations and envisioned future
improvements. We also examine the predictive brain regions our classifiers identified and discuss their relevance in the context of hearing loss.

### 3.2 Materials and Methods

#### 3.2.1 Data acquisition and preprocessing

**A. Participants**

Thirty-nine infants and toddlers participated in a clinically indicated MRI brain study under sedation. This study was conducted with approval from Cincinnati Children’s Hospital Medical Center Institutional Review Board (IRB). Eighteen of the participants had SNHL (10 females, average age = 14 months, range = 8-24 months). All hearing impaired participants were referred by the Division of Otolaryngology for MRI as part of the cochlear implant staging process and consented to participate in our adjoining fMRI protocol. They had documented bilateral severe to profound hearing loss with average hearing thresholds in the range of 90 dB or greater. Nine of these subjects had no measureable hearing response in either ear at the maximum level of our audiometry equipment, at 120dB and can be considered deaf. The remaining 21 participants were normal hearing controls (15 females, average age =12 months, range = 8-17 months). These children received clinical MRI scans with sedation for non-hearing related indications. They were recruited for the control group if they met inclusion criteria: gestational age of at least 36 weeks, normal otoacoustic emissions hearing, and normal neuroanatomy determined by the neuroradiologist. Informed consent of parent or guardian was obtained prior to the study protocol, and the parent agreed to additional hearing tests at a separate visit. The child’s reason for referral for brain MRI was not related to hearing. Exclusions included head circumference <5 percentile or >95 percentile, orthodontic or metallic implants that interfere with the MRI, abnormal brain pathology in the central auditory pathways. Examples of indications for scanning in this group
were, “odd body positioning-rule out chiari malformation”, “recent onset irritable behavior-rule out brain tumor”. All participants were screened for hearing loss using otoacoustic emission (OAE) prior to the MRI scan. Failed OAE at the time of scan was also an exclusion criterion for the normal control group. All of these brain scans of both hearing impaired group and control group were reviewed by a pediatric neuroradiologist and assessed as having no anatomical findings of significance. One of the challenges of research in pediatric neuroimaging is that it is unethical to expose children to more than minimal risk for the purposes of research. This principle is dictated by our conscience as well as by the IRB at most institutions. Consequently, one of the fine points in the design of the present study is that we were required to select our control population among infants who were referred for an MRI scan with sedation because of a clinical indication. With the precautions described above and other procedures we took to insure normal auditory function and brain anatomy, this is perhaps the best control group that could be obtained for this age group in an ethical fashion. However, it is important to note that the controls were not randomly sampled from the general population. For a complete description of the clinical population selected for this study and earlier analysis of the same data sets the reader is referred to (Smith et al., 2011).

B. MRI/fMRI acquisition

Anatomical images for this study were acquired using a 3.0 Tesla Siemens Trio MRI scanner in the clinical Department of Radiology. Isotropic images of the brain were acquired using an inversion recovery prepared rapid gradient-echo 3D method (MP-RAGE) covering the entire brain at a spatial resolution of $1 \times 1 \times 1$ mm in an axial orientation. 3D MP-RAGE acquisition parameters were as follows: TI/TR/TE = 1100/1900/4.1 ms, FOV = 25.6 $\times$ 20.8 cm, matrix =
256 × 208, scan time = 3 minutes and 50 seconds. These high resolution 3D-T1 weighted images were used for co-registration of fMRI scans which were also acquired during this scheduled MRI.

Functional MRI scans were performed using a silent background fMRI acquisition technique that allowed auditory stimuli to be presented during a silent gradient interval of the scan, followed by an acquisition interval that captured the peak BOLD response of relevant brain regions (Schmithorst and Holland, 2004). Using the scanner described above we acquired BOLD fMRI scans in an axial plane (4 × 4 mm resolution), using the manufacturers standard gradient echo, EPI sequence covering the same FOV as the 3D T1 images (see paragraph above), with the following parameters: TR/TE = 2000/23 msec, flip angle = 90°, matrix = 64 × 64 and 25 axial slices with thickness = 5 mm. In the present study, all stimulus and control intervals were of equal duration (5 sec) in a three-phase auditory paradigm consisting of speech, silence, and narrow band noise tones interleaved with acquisition periods of 6 seconds during which 3 image volumes were obtained covering the whole brain. A timing diagram for the fMRI data acquisition and stimulation paradigm is shown in Figure 3-1.

![Figure 3-1. Timing diagram for fMRI paradigm](image_url)
The speech stimulus consisted of sentences read in a female voice. Altogether 36 sentences were read in 18 segments of 5 seconds duration and comprising 2 sentences each. This condition was followed by a 6 second data acquisition and then a 5 second interval of silence as a control condition. After another 6 second control interval acquisition, a second auditory control condition was played. This condition consisted of Narrow Band Noise (NBN) tones patterned after standard audiology evaluations for detection of hearing thresholds. Five NBN tones of 1 second duration with center frequencies of 250, 500, 1,000, 2,000 and 4,000 Hz and bandwidth of 50% were played in random order during this control condition, for a total of 5 seconds during a silent interval of the scanner. An additional interval of 1 second of silence followed each acquisition to provide an acoustic demarcation prior to the stimulus onset of each stimulus condition. This resulted in the fMRI acquisition time of approximately 11 minutes. See Figure 3-1 for a detailed schematic of the task and timing. Auditory stimuli were administered through calibrated MR compatible headphones at a sound level of 10-15 dB greater than the individual participant’s Pure Tone Average (PTA) hearing threshold. Each hearing impaired participant in the study had a recent audiogram, which was used to determine the sound level for fMRI. Our MR compatible audio system was modified to allow for an output through the headphones measuring up to 130 dB.

C. Data analysis - preprocessing

fMRI data were initially analyzed on a voxel-by-voxel basis to identify the activated brain regions using a standard pre-processing pipeline implemented in Cincinnati Children’s Hospital Image Processing Software (CCHIPS) (Schmithorst et al., 2010) written in IDL computer language. In this chapter, we use voxel for 3-dimensional images and pixel for 2-dimensional images.
Since the subjects were sedated, we assumed that the anatomical image was naturally aligned with the functional images for each individual. Therefore, alignments between anatomical images and functional images were not needed in preprocessing. In this case, it does not matter if we apply the normalization transformation before or after contrast determination. To generate both normalized contrast maps used in the current study as well as contrast maps in native space for other uses, we first generated contrast maps in each individual’s native space and then normalized the contrast maps to standard space. The raw EPI images were simultaneously corrected for Nyquist ghosting and geometrical distortion (due to B0 field inhomogeneity) (Schmithorst et al., 2001). EPI functional MR time-series images were corrected on a voxel-by-voxel basis for drift using a quadratic baseline correction. Motion artifacts were corrected using a pyramid iterative co-registration algorithm (Thevenaz et al., 1998). During this stage, infant brain images were transformed to the AC-PC plane. Finally, the individual image volumes (1,2,3) in the event-related fMRI acquisition were separated and submitted to a final pre-processing step using the General Linear Model (Worsley et al., 2002) to construct individual Z-maps for each volume and contrast condition (speech vs. silence, speech vs. noise and noise vs. silence). Z-maps showing activation for each condition for each participant were then computed by averaging the Z-maps from the individual volumes for each contrast condition (Schmithorst and Holland, 2004, Patel et al., 2007). These Z-maps, in each individual’s native space were used by the radiologists and neurotologists for clinical interpretation of findings. The neuroradiologist reviewed both functional and anatomical MRI scans for each participant and completed a standardized report indicating whether brain abnormalities or brain activities were detected in primary auditory areas, language areas or other brain regions. After that, we performed spatial normalization using SPM8 with a T1 template
constructed from a control group of age-matched subjects selected specifically for this infant cohort (Altaye et al., 2008). The normalized anatomical images and functional Z-maps were then submitted to the next stages of analysis.

3.2.2 Feature extraction and model learning based on structural MR images

For sMRI images, we used SIFT features to represent the brain images and developed an algorithm to analyze the SIFT features. We have previously applied this method to Alzheimer’s disease, Parkinson’s disease and bipolar disease, and demonstrated promising classification performance (Chen et al., 2013).

A. Obtaining 2D slices from 3D brain images

Due to the high density of SIFT features in the brain images and the pair-wise comparison among SIFT features required in a later step, analyzing the 3D brain image as a whole is computationally infeasible. Thus, the spatially normalized 3D brain (157×189×136) was divided into 560 20×20×20 cubes. Since the dimensions of brain image were not divisible by 20, the cubes at the end of dimensions only contained the remaining volume of the brain image and therefore had a size smaller than 20×20×20. The number 20 was determined based on our experience from the application of this algorithm to several other diseases. The cube size mainly affects the computation speed and accuracy of the likelihood scores as described in the section Feature evaluation below. A larger size leads to much longer computation time, while a smaller size decreases the accuracy of likelihood scores and subsequently leads to lower classification accuracy. According to our experimental results, the cube size 20×20×20 provides a good balance between speed and accuracy. Every cube was sliced along three different orientations to obtain 3 sets of 20 2D brain images. We analyzed every cube and every set of 2D brain images individually. The analysis results were combined together in the last step.
B. Extracting SIFT features

The SIFT algorithm for analyzing 2D images was implemented in several stable software packages (Lowe, Vedaldi and Fulkerson, 2010). In this study, we used the SIFT algorithm provided in a publicly available computer vision software package vlFeat (Vedaldi and Fulkerson, 2010). The SIFT features are described by center locations, scales, orientations and appearance matrices. An example of SIFT features is shown in Figure 3-2. The SIFT features are shown as circles in Figure 3-2(a). Each circle represents a SIFT feature. The center and radius of the circle represent the center location and the scale of the SIFT feature. The existence of a SIFT feature suggests there is a blob-like image component at the center location of the SIFT feature and the scale of the feature represents the radius of the blob-like component. The image intensity distribution around the blob-like component is further characterized by an orientation and an appearance matrix. The orientation, as shown by the line starting from the center of the circle, represents the general direction of change in image intensity. The appearance matrix represents the detailed change in image intensity. An example of an appearance matrix is shown in Figure 3-2(b). The square centered at the center location of a SIFT feature is divided into 16 subsquares. There are 8 lines starting from the center of each subsquare along 8 different directions. The length of a line represents the number of pixels which have a gradient direction the same as the line, and some of the lines may have a length of zero. For example, many of the pixels in the lower left corner subsquare, as shown in Figure 3-2(b), have a gradient direction pointing to the lower side of the image; therefore the length of the line starting from the center of this subsquare and pointing to the lower side is long. The center location, scale, direction and appearance matrix of a SIFT feature can be organized as a vector of 133 numbers: the center location includes 3 numbers representing its coordinates in the 3D volume of the brain image;
the scale and orientation is represented as one number respectively; the appearance matrix is represented by 128 numbers, 8 numbers for each of the 16 subsquares. This vector form is used in the computation; while the isomorphic graph representation, as shown in Figure 3-2, is used as a user friendly way of representing the SIFT features.

![Figure 3-2. SIFT features](image)

C. Feature evaluation

The extracted SIFT features were identified as one of the three feature types, namely patient feature, healthy feature and noise feature. The features were evaluated based on their frequencies of occurrence in patient brains and healthy brains.

There were two steps to evaluate the features, and each SIFT feature was evaluated separately. The first step was to find all the other features that were similar to the feature that was being analyzed. The similarity between two features was measured by four criteria: the distance between the center locations $\Delta x(i, j)$, the scale difference $\Delta \sigma(i, j)$, the orientation difference $\Delta \alpha(i, j)$ and the difference between their appearance matrix $\Delta a(i, j)$. They were defined as follows:
\[ \Delta_x(i,j) = \frac{\|x_i - x_j\|_2}{\sigma_i} \quad (3.1) \]
\[ \Delta_\sigma(i,j) = \left| \ln \frac{\sigma_j}{\sigma_i} \right| \]
\[ \Delta_o(i,j) = \min\left( |o_i - o_j|, 2\pi - |o_i - o_j| \right) \quad (3.3) \]
\[ \Delta_a(i,j) = \|a_i - a_j\|_2 \quad (3.4) \]

where \( x_i \) was the center location of feature \( i \), \( \sigma_i \) was the scale of feature \( i \), \( o_i \) was the orientation angle of feature \( i \) and \( a_i \) was the appearance matrix of feature \( i \). If all the four differences were less than their corresponding threshold, two features were considered to be similar. All the features that were similar to feature \( i \) constituted the similar feature set for feature \( i \):

\[ S_i = \{ f_j : \Delta_x(i,j) < \epsilon_x \land \Delta_\sigma(i,j) < \epsilon_\sigma \land \Delta_o(i,j) < \epsilon_o \land \Delta_a(i,j) < \epsilon_a \} \quad (3.5) \]

where \( \epsilon_x, \epsilon_\sigma, \epsilon_o \) and \( \epsilon_a \) were similarity thresholds for center locations, scales, orientations and appearance matrices, respectively. According to (Toews et al., 2010), the thresholds \( \epsilon_x \) and \( \epsilon_\sigma \) were set to 0.5 and 2/3 respectively. The thresholds \( \epsilon_o \) and \( \epsilon_a \) were set to \( \pi/2 \) and 0.45 respectively based on a grid search (Chang and Lin, 2011). Grid search is an efficient way to find the best parameter combinations, when there are multiple parameters in a model and the parameters are continuous variables. First, we discretized the continuous parameters. Parameter \( \epsilon_o \) was discretized into three discrete values \([\pi/4, 2\pi/4, 3\pi/4]\), and parameter \( \epsilon_a \) was discretized into five discrete values \([0.3, 0.35, 0.4, 0.45, 0.5]\). Then all the combinations of these discrete values, 15 combinations in total, were tried and the parameter combination with the highest classification accuracy was chosen as the best parameter setting.

The second step for feature evaluation was to assign likelihood scores to the SIFT features. The likelihood score was defined as follows:
\[ L_i = \begin{cases} 
\ln \frac{|S_i \cap P|/N_p}{|S_i \cap C|/N_C}, & |S_i| \geq N_p + N_C \\
0, & \text{otherwise} 
\end{cases} \]  

(3.6)

where \( S_i \) was the similar feature set for SIFT feature \( i \), \( P \) was the patient feature set which included all the SIFT features extracted from all patient brains in the training set, \( C \) was the healthy feature set including all the SIFT features from all healthy brains in the training set, \( N_p \) and \( N_c \) was the number of patient brains and the number of healthy control brains in the training set, respectively.

A SIFT feature was identified as a patient feature if \( L_i \) was larger than a threshold \( \epsilon_i \); it was a healthy feature if \( L_i \) was smaller than \(-\epsilon_i\); it was a noise feature otherwise. Formally, the class labels of the features were determined as follows:

\[ C_i = \begin{cases} 
1, & L_i > \epsilon_i \\
0, & |L_i| \leq \epsilon_i \\
-1, & L_i < -\epsilon_i 
\end{cases} \]  

(3.7)

where \( \epsilon_i \) was the threshold for likelihood scores. We used grid search to determine the best parameter setting. For the threshold, the value from 0.1 to 1.2 with a step size of 0.1 was searched. After the grid search, \( \epsilon_i \) was set to be 0.9.

According to the above feature evaluation process, we need to find the similar feature set for every feature (Equation (3.5)), which requires to compare this feature with all other features. For more than \( 10^5 \) features in 39 brains, it would require \( 10^{10} \) pair-wise distance calculations, which is a very slow process. Upon those observations, we divided the whole brain volume into small cubes. For the evaluation of a feature, we only calculated its distance to the other features in the same cube. In this way, the computation time is significantly reduced, but the classification accuracy may be adversely affected. For example, a feature close to cube boundaries may have some of its similar features (Equation (3.5)) in adjacent cubes. Ignoring those similar features in
adjacent cubes could lead to an inaccurate likelihood score (Equation (3.6)) for this feature. This issue is especially serious when the number of training samples is limited as in our project. On the other hand, a larger cube size would have fewer features close to cube boundaries, and would result in more accurate likelihood scores and hence higher classification accuracy. According to our previous experience from the application of this algorithm to the classification of several other diseases, such as Parkinson’s disease, Alzheimer’s disease and bipolar disorder, 20×20×20 was considered to be an appropriate cube size. This cube size 20×20×20 determined based on adult-sized brains in our previous studies was used directly for the infant brains in the present study, since our infant brains were normalized using the infant template and the infant template was enlarged to the size very close to that of adult brains (Altaye et al., 2008).

D. Training SVM classifiers

We trained a linear SVM for every set of 2D slices in every cube to classify the set of SIFT features extracted from this set of 2D slices across subjects into 3 categories. As we described in section 2.5, the SVM classifier was originally designed for binary classification. How to extend the binary classifier for multi-class situations is still a research issue waiting for further investigations. The LIBSVM package (Chang and Lin, 2011) used in this project took the one-against-one strategy (Hsu and Lin, 2002) to extend the two-class SVM classifier to be a multi-class classifier. For a new SIFT feature from a brain image whose class-label is unknown, the corresponding SVM is expected to be able to predict the class label of this new SIFT feature without finding its similar feature set in the huge amount of SIFT features extracted from the brain images used for training.
E. **Predicting new subjects**

To predict a new subject to be NH or HI, the subject’s sMRI scan was first normalized to the standard space using SPM8 with the infant T1 template (Altaye et al., 2008). The normalized brain was divided into cubes and sliced along three orientations as described above. SIFT features were extracted and then classified using the SVM model that was trained for the same cube and same slice orientation. After all the SIFT features were classified, we counted the number of features of the three types. The total number of noise features was not used in the final decision process. The new subject was classified according to the following equation:

\[
\text{Class Label} = \begin{cases} 
    \text{HI}, & \text{if } C_{\text{sum}} > \epsilon_s \\
    \text{NH}, & \text{otherwise}
\end{cases}
\]  

(3.8)

where \( C_{\text{sum}} = \sum_i \hat{C}_i \), \( \hat{C}_i \) is the predicted class label of the \( i \)-th SIFT feature as shown in Equation (3.8), \( \epsilon_s \) is a threshold for the final classification of sMRI and its value is determined based on the method described in section 3.2.5 **Validation of the classifier.**

3.2.3 **Feature extraction and model learning based on functional MR images**

For fMRI images, we constructed contrast maps using the general linear model (GLM) (Worsley et al., 2002) as described in the **Data preprocessing** section. Contrast values were estimated from the difference in image intensity for each voxel between two conditions. A positive contrast value indicated that the brain activation was higher in the first condition when compared to the second condition, while a negative contrast value suggested a lower activation in the first condition. We generated region-level features and proposed a novel approach to vectorize the contrast maps utilizing the “bag-of-words” strategy (Sivic and Zisserman, 2009).

A. **Feature generation from contrast maps**

The Bag-of-Words (BoW) approach was originally developed in document classification for assigning a document to two or more classes. All the words occurring across all of the...
documents constitute a dictionary. Suppose we have N words in the dictionary. Based on this dictionary, a document can be represented into an N-dimensional feature vector, with each word becoming a feature. The value of a feature is measured as the occurrence frequency of the corresponding word in the document. We have introduced this idea to the contrast maps. By analogy, each contrast map represents a document, each significantly activated/deactivated region is a word or a region of interest (ROI), and all the ROIs occurring across all of the contrast maps constitute a dictionary. The frequency of each word was measured by the mean contrast value within a ROI.

In specific, the normalized Z-maps were first thresholded to select voxels with extreme contrast values for subsequent analysis. Among the selected voxels, we connected the voxels which were adjacent to each other in a 3D neighborhood, in which each voxel had 26 neighbors if it was not on the border. As a result, the selected voxels were merged into a set of disjoint regions, each of which was defined as a characteristic contrast region or ROI (Dykstra, 1994, Pokrajac et al., 2005). To prevent mixing positive voxels and negative voxels in a single ROI, which could negate the signal, we considered these two categories of voxels separately. Positive voxels were ranked decreasingly whereas negative voxels were ranked increasingly according to their activation magnitudes. Only the top 5% of each category were selected. The cutoff of 5% was chosen because it outperformed other cutoffs, 1% and 10%, with respect to the classification performance. In this way, a number of ROIs were delineated to characterize the pattern of a contrast map. Due to individual differences and random noise, however, the set of ROIs delineated from different subjects varied significantly. To address this problem, we delineated a set of ROIs based on each subject, and applied all ROIs derived from all subjects to each single subject to form a long vector for each subject, with each dimension representing the mean
contrast value over all voxels within the corresponding ROI. Finally, we concatenated the vectors from the three contrast maps, and obtained a 1,474-dimensional vector for each subject. An intuitive view of the contrast map vectorization process is shown in Figure 3.3.

![Figure 3.3. Vectorization of contrast maps](image)

Since we performed ROI detection on each contrast map and then concatenated all the ROIs together, ROIs that were consistent among subjects were detected more than once. To merge those similar ROIs into one single feature, we performed a hierarchical clustering with average linkage (Johnson, 1967). The original space was represented as:

\[
\begin{align*}
S_{(1,1)} & \quad \cdots \quad S_{(1,1474)} \\
\vdots & \quad \ddots \quad \vdots \\
S_{(N,1)} & \quad \cdots \quad S_{(N,1474)} 
\end{align*}
\]

where each row represents a training sample and each column represents a ROI, \(S_{(i,j)}\) is the mean contrast value of ROI \(j\) for subject \(i\), \(N\) is the total number of subjects. The distance between two ROIs was calculated as the Euclidean distance:
We cut the hierarchical tree with the inconsistency coefficient of 0.01, and calculated the mean value of the ROIs that were clustered together as the value of the joint feature. The cutoff of 0.01 was easily determined since the cluster results did not change in the cutoff range from 0.01 to 0.7. After hierarchical clustering, the dimensionality was reduced to 969.

\[ \text{dist}(ROI_i, ROI_j) = \sqrt{\sum_{k=1}^{N} (S_{(k,l)} - S_{(k,j)})^2} \]  \hspace{1cm} (3.10)

B. Sedation method

Subjects were sedated with three different sedation methods during the MRI scanning. Different sedation methods were expected to affect the activation pattern differently (DiFrancesco et al., 2013). Therefore, we added sedation method as an additional feature, which was represented as a 3D binary vector.

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]  \hspace{1cm} (3.11)

As shown in the matrix defined in Equation (3.11), each row of the matrix represented one of the three sedation methods. In this way, we represented each subject as a 972-dimension feature vector, including 969 features from the contrast maps after hierarchical clustering and 3 binary features from sedation method. Therefore, our dataset was represented as \( \mathcal{D} \) defined in Equation (3.12):

\[
\mathcal{D} = \{(x^{(1)}, y^{(1)}) \cdots, (x^{(i)}, y^{(i)}) \cdots, (x^{(39)}, y^{(39)}) \mid x^{(i)} \in \mathbb{R}^{972}\} 
\]  \hspace{1cm} (3.12)

where \( x^{(i)} \) and \( y^{(i)} \) was the feature vector and group label (NH or HI) for the \( i \)-th subject, respectively. This dataset \( \mathcal{D} \) was used for subsequent feature selection and model learning.
C. Feature selection and model learning

The WEKA software package was utilized to select a subset of features that were highly correlated with class labels and uncorrelated with each other (Hall, 1999). The merit of a subset of features was measured as:

\[
M_S = \frac{k r_{cf}}{\sqrt{k + k(k-1)r_{ff}}}
\]  

(3.13)

where \( r_{cf} \) was the mean correlation between class label and selected features, \( r_{ff} \) was the mean correlation between two features, \( k \) was the number of features in subset \( S \). Greedy hill-climbing augmented with a backtracking facility was applied to search through the space of feature subsets (Dechter and Pearl, 1985). For explanation purposes, we can imagine that there was a rooted tree, which had included all possible feature subsets. In this tree, each node was a feature subset, which was represented as a 972 dimensional binary vector, with 1(0) indicating the corresponding feature was (not) selected. Each node had 972 successors/children, each of which was generated by flipping one of the 972 dimensions of the current node. Our goal was to step through this tree to find a node with relatively high \( M_s \). In practice, the whole tree would not be constructed because it was unlimited. Only the successors were generated whenever needed. The search started from the root, which was the empty set of features in our project, and repeatedly chose the successor with the highest \( M_s \) at each node. The search terminated when 5 consecutive non-improving steps occurred. With the selected subset of features, we trained a linear SVM classifier (Chang and Lin, 2011).

D. Predicting new subjects

Given a new subject, we first normalized the contrast maps to the infant template space (Altaye et al., 2008), so that the given contrast maps were registered with the training contrast maps. A 972-D feature vector was then constructed with procedures described above, which was
subsequently filtered based on the feature selection results obtained on the training set. Finally, the formatted feature vector was fed to the trained classifier, yielding a decision score ($f_{MRI \_score}$) for the new subject based on the functional MRI data alone. The rule for classification was formulated as:

$$
\text{Class Label} = \begin{cases} 
H I, & \text{if } f_{MRI \_score} \geq \varepsilon_f \\
N H, & \text{otherwise}
\end{cases}
$$

(Equation 3.14)

E. Important features

The importance of a feature was measured as follows:

$$
I(f) = \left| \sum_{i=1}^{N} \sigma_i w_{if} \right|
$$

(3.15)

where $|$ was the absolute value function, $N$ was the total number of folds of cross-validation as described in the following part, $w_{if}$ was the SVM weight for feature $f$ during $i$-th fold of cross-validation, $\sigma_i = 1$ if the feature $f$ was selected in the $i$-th fold of cross-validation. Otherwise, $\sigma_i = 0$. For the ROIs that were merged into a joint feature through hierarchical clustering, the importance of such an ROI was equal to the importance of the feature, to which this ROI belonged.

3.2.4 Integrated model

To combine the sMRI and fMRI data, we designed a two-layer classification model (Figure 3-4). Given a training set, we trained two classifiers, namely sMRI classifier and fMRI classifier. Then we applied these two classifiers to the training set. As a result, we obtained two predicted scores for each training sample. Thus, the original feature space was transformed into a new two-dimensional feature space through these two classifiers. Finally, we trained a linear SVM classifier (with parameter $C=1$) in the new feature space to combine the two scores together.
When predicting new subjects, we first obtained the two predicted scores from the sMRI classifier and fMRI classifier, then fed these two predicted scores into the second layer classifier to yield the final decision score $y$. The decision rule was defined as follows:

$$y = f(C_{sum}, fMRI\_score) = w_1 \ast C_{sum} + w_2 \ast fMRI\_score + bias$$

(3.16)

$$\text{Class Label} = \begin{cases} HI, & \text{if } y \geq \epsilon_i \\ NH, & \text{otherwise} \end{cases}$$

(3.17)

where $w_1$, $w_2$ and $bias$ were the parameters in the SVM model, which were learnt from the training.

### 3.2.5 Validation of the classifier

Leave-one-out cross-validation (LOOCV) was employed to validate the three classifiers as follows. The total number of subjects was denoted as $N$. We performed $N$ experiments, each of which was called one fold of cross-validation. In the $n$-th ($n = 1, \ldots, N$) fold of cross-validation, the $n$-th subject was used for testing; while the others were used for training. Threshold $\epsilon_s$ was determined so that the false positive rate and false negative rate for the training brains were equal, while $\epsilon_f$ and $\epsilon_i$ were set to be 0. These thresholds were applied to the test images to assign them to be either NH or HI. The classification accuracy for all the $N$ subjects was reported as accuracy. Equal error rate (EER) accuracies were also determined based purely on the predicted scores of the testing brain images, e.g. the threshold $\epsilon_s/\epsilon_f/\epsilon_i$ were chosen so that the false positive rate was equal to false negative rate for the testing brains. In addition, area under curve (AUC) was also calculated to evaluate the performance of classifiers.
3.3 Results

3.3.1 Classifier performance

Performances of the three classifiers are shown in Table 3-1, and receiver operating curves (ROCs) are plotted in Figure 3-5. While the sMRI classifier and fMRI classifier performed well individually, their combination achieved a significant improvement in performance. The combined classifier yielded AUC and EER as high as 0.90 and 0.89, respectively. From the ROC, we can see that the sMRI classifier could not predict some of the positive subjects (HI) correctly even when the decision threshold was set to be very low, because the classifier did not reach 100% true positive rate even when the false positive rate approached 100%. However, the ROC for fMRI was in an opposite situation. The ROC did not reach 0% false positive rate even when the
true positive rate approached 0%, suggesting that fMRI classifier had difficulty in classifying some of the negative subjects (NH) correctly. As sMRI and fMRI classifiers were vulnerable to different types of errors, it was possible to combine them to overcome their individual limitations. To illustrate the reason why the combination can be successful, we plotted sMRI-fMRI scores in Figure 3-6. Simply speaking, fMRI classifier draws a horizontal line to separate the two groups of subjects based on the fMRI data, while sMRI classifier draws a vertical line to separate the two groups based on the sMRI data. Obviously, the two groups could not be perfectly separated by either a horizontal or a vertical line in Figure 3-6. However, by combining the fMRI and sMRI classifiers, the two groups of subjects were separable with a diagonal line as shown in the figures.

<table>
<thead>
<tr>
<th></th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
<th>AUC</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>sMRI</td>
<td>0.83</td>
<td>0.62</td>
<td>0.72</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>fMRI</td>
<td>0.76</td>
<td>0.72</td>
<td>0.74</td>
<td>0.83</td>
<td>0.76</td>
</tr>
<tr>
<td>sMRI+fMRI</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

### 3.3.2 Feature selection in sMRI analysis

In the analysis of sMRI data, image features were selected based on their likelihood scores. The total number of image features in a brain image ranged from 35,000 to 52,000. Most of these image features were noise features. The total number of selected features, i.e., healthy and patient features, ranged from 300 to 1,400 for different brains with a likelihood threshold of 0.9. Different choices of likelihood threshold for the sMRI feature selection resulted in different numbers of selected features and therefore different classification accuracies. Table 3-2 shows the relation between classification accuracy and the likelihood threshold. The classification accuracy did not change for likelihood threshold ranging from 0.7 to 1.1. The AUC changed
within a range of 0.09 with a peak where the likelihood threshold equaled 0.9. The EER accuracy varied within a range of 0.08. All three classification performance measures were stable with different likelihood thresholds.

Figure 3-5. ROCs of the three classifiers
3.3.3 Stability of feature selection in fMRI analysis

We have analyzed the stability of feature selection in the analysis of fMRI data. There were in total 972 features as the input for feature selection. Only 6.2% of the features (with a total number of 60) were selected at least once. For each fold of cross-validation, there were usually about 20 features selected for the training, generally 30% of which were consistently present in all folds of cross-validation. We calculated a stability index as follows (Kalousis et al., 2007):

\[
Sim(s_i, s_j) = \frac{|s_i \cap s_j|}{|s_i \cup s_j|}
\]  

(3.18)
\[
\text{index} = \frac{2}{c(c-1)} \sum_{i=1}^{c-1} \sum_{j=\{i+1\}}^{c} \text{Sim}(s_i, s_j)
\]

where \(c\) was the total number of rounds of feature selection, \(s_i\) and \(s_j\) were two sets of features selected during two runs, \(|s_i \cap s_j|\) was the cardinality of the intersection between \(s_i\) and \(s_j\), and \(|s_i \cup s_j|\) was the cardinality of the union of \(s_i\) and \(s_j\). Our feature selection yielded a stability index of 66.2%, which indicated that 66.2% of the selected features, on average, were common between any two runs of feature selection. Since the Euclidean distance was used in the hierarchical clustering, only very similar ROIs were merged. There was still considerable redundancy among features. For example, two ROIs, e.g. one from the contrast speech vs. silence and the other from the contrast noise vs. silence, were significantly correlated with class labels, and meanwhile they were also highly correlated with each other. Due to the large Euclidean distance between them, however, they were not merged during the hierarchical clustering. In feature selection, these two ROIs were treated as different features and selected interchangeably. This caused the calculated stability index lower than the actual value. In this regard, 66.2% represented very high stability.

### 3.3.4 Discriminative brain regions

For sMRI, we measured the importance of a SIFT feature with its likelihood score. In our project, however, the SIFT features usually had a scale of 10 mm or even larger, and correspondingly the side length of the appearance matrices was larger than 40 mm. Due to the large size of the SIFT features, it was more difficult and less useful to interpret the medical implications of such large brain regions.

With those considerations, we only focused on the highly predictive brain regions identified by the fMRI classifier. **Figure 3-7** shows the top 10 functional features extracted from
fMRI data that differentiate the HI and NH groups. Features are numbered from A to J in order. ROI A1 and A2 were merged during hierarchical clustering into a joint feature A. Similar procedures were performed for feature C, E, F, I and J. We can see that ROIs grouped together during hierarchical clustering are always from the same type of contrast maps (Table 3-3) and encompass adjoining or sometimes overlapping brain regions as designated by Brodmann’s Areas in the 4th column of Table 3-3.

Table 3-3. Characteristics of the top functional ROIs. STG is short for singular temporal gyrus, MTG for middle temporal gyrus

<table>
<thead>
<tr>
<th>Feature</th>
<th>number of voxels</th>
<th>contrasts</th>
<th>BA areas</th>
<th>Anatomical labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>16</td>
<td>speech vs. silence</td>
<td>Red Nuc.</td>
<td>Red Nuc.</td>
</tr>
<tr>
<td>A2</td>
<td>10</td>
<td>speech vs. silence</td>
<td>Thal., Sub Thal. Nuc</td>
<td>Thal., Sub Thal. Nuc.</td>
</tr>
<tr>
<td>B</td>
<td>19</td>
<td>speech vs. noise</td>
<td>22</td>
<td>STG</td>
</tr>
<tr>
<td>C1</td>
<td>19</td>
<td>noise vs. silence</td>
<td>33</td>
<td>Pregenual Cing. G.</td>
</tr>
<tr>
<td>C2</td>
<td>15</td>
<td>noise vs. silence</td>
<td>33,24</td>
<td>Pregenual Cing. G., Vent. Ant. Cing. G.</td>
</tr>
<tr>
<td>D</td>
<td>14</td>
<td>noise vs. silence</td>
<td>Auditory Nuclei</td>
<td>Pontine Auditory Nuclei</td>
</tr>
<tr>
<td>E1</td>
<td>12</td>
<td>speech vs. noise</td>
<td>32</td>
<td>Cing. G.</td>
</tr>
<tr>
<td>E2</td>
<td>11</td>
<td>speech vs. noise</td>
<td>32,10</td>
<td>Cing. G., Prefrontal Cortex</td>
</tr>
<tr>
<td>G</td>
<td>11</td>
<td>speech vs. silence</td>
<td>24,32</td>
<td>Vent. Ant. Cing. G., Cing. G.</td>
</tr>
<tr>
<td>I1</td>
<td>24</td>
<td>speech vs. silence</td>
<td>22,40,39</td>
<td>STG, Super Marg. G, Angu. G.</td>
</tr>
<tr>
<td>I2</td>
<td>30</td>
<td>speech vs. silence</td>
<td>40,39</td>
<td>Super Marg. G, Angu. G.</td>
</tr>
<tr>
<td>J1</td>
<td>253</td>
<td>noise vs. silence</td>
<td>Thal., Red Nuc.</td>
<td>Thal., Red Nuc.</td>
</tr>
<tr>
<td>J2</td>
<td>142</td>
<td>noise vs. silence</td>
<td>Thal., Mamillary Body</td>
<td>Thal., Mamillary Body</td>
</tr>
</tbody>
</table>
Figure 3-7. Important brain regions identified by fMRI classifier as differentiating the hearing impaired group (HI) from the normal hearing control group (NH).
3.4 Discussion

In this work, we have built a robust two-layer classifier that can accurately separate HI from NH infants. We realize that hearing in newborns can be accurately tested using the auditory brainstem response (ABR) evaluations or the otoacoustic emission (OAE) measures, it is thus not our intention to develop a tool for computer-aided diagnosis of hearing loss. Rather we provide a proof of principle that it is possible to accurately determine the functional, developmental status of the central auditory system in congenitally hearing impaired children based on MR images alone by utilizing machine learning techniques. Such success has been previously reported in other progressive diseases, such as Alzheimer’s disease (Cuingnet et al., 2011). However, for many progressive diseases, definite diagnosis is often difficult to establish, in which case the LOOCV approach may not be able to estimate the classifier performance accurately. Therefore, our dataset with solid labels corresponding to diagnostic categories of the participants has NH or HI enables us to make an objective evaluation of our algorithm, and demonstrate conclusively the feasibility of using machine learning in making automated diagnoses or prognoses based on imaging examinations. The approach described here may not be limited to a specific disease; essentially, any disease dataset with sMRI and fMRI brain images can be analyzed with our method providing sufficient training data is available.

A major innovation that makes highly accurate predictions possible in our approach is that we extracted high-level features instead of using each single voxel as a feature as in traditional approaches. We extracted SIFT features from sMRI and BoW features from fMRI. Both of these two types of features were on the region-level. For the BoW features, please note that we constructed our feature pool with all available samples, including the one used for testing during the cross-validation. We implemented a variant version of our algorithm, in which we extracted
ROIs based only on the training samples, and subsequently applied those ROIs to the testing sample directly. As expected, the variant algorithm performed slightly worse (AUC=0.81) than our original algorithm (AUC=0.83). Adding the ROIs from new samples requires us to retrain the classifier every time when new samples are available. As the feature pool becomes larger in the future, the retraining is not necessary.

Integration of different types of data, e.g. data from multiple modalities, has been demonstrated to be more powerful for classification (Fan et al., 2007, Fan et al., 2008, Tosun et al., 2010, Wang et al., 2012b). However, how to implement such integrations in the best way remains to be explored (Orru et al., 2012). Traditionally, features from different types of data are concatenated and a single classifier is trained (Fan et al., 2007, Fan et al., 2008, Tosun et al., 2010, Wang et al., 2012b). Specifically, the traditional integration method requires the training set to be organized into matrices, with each row representing a training sample and each column representing a feature. One matrix is constructed for one type of data, and subsequently all the matrices are concatenated into one big matrix, which serves as the input for classifier training. In our project, the fMRI data can be easily organized in this way. For sMRI, however, each training sample has a set of SIFT features, which can be treated as a set of words included in an article. Different articles have different sets of words. Thus, it is not easy to organize the sMRI data into a matrix as described above, and the traditional integration method is not applicable. Under such circumstances, we proposed a two-layer model to integrate the sMRI and fMRI data. Since the traditional approach was not applicable in our project, we did not compare their performances in the present study. Additionally, our two-layer model is also applicable when features from different modalities can be concatenated. In this case, one classifier is trained for one modality, and a second-layer classifier is subsequently used to integrate the multiple classifiers on the first-
layer. This approach is able to combine as many types of data as possible, without worrying about the high dimensionality or overfitting.

Although computer-aided diagnosis of hearing loss is not needed, our algorithm can potentially advance the study of congenital hearing loss mechanism by identifying discriminative brain regions as disease biomarkers for hearing impairment at various levels in the auditory system. Inspecting the most important features that differentiate children born with hearing impairment from children with normal hearing in this study, we see some features that are in line with hypotheses about under stimulation of auditory function in HI infants; while other observations already begin to add to our knowledge of how congenital deafness affects brain development and function. For example, features B, F, H, and I include known components of the auditory language network which our group and others have previously shown to be engaged by the narrative comprehension task (Schmithorst et al., 2006, Karunanayaka et al., 2007). These features include (B) the planum temporale and primary auditory cortex in the left hemisphere (including Wernicke’s area, the classical language recognition module), as well as the angular gyrus and supramarginal gyrus at the temporal parietal junction of the (F) left and (H, I) right hemispheres, known auditory and visual language association regions. Although all participants were bilaterally severely to profoundly hearing impaired, we observe left dominant auditory/language related activity present in components A, B, and F. In addition, components H and I contain right hemisphere auditory/language activity. Functional features such as these are not unexpected in terms of regions of differential cortical activation between HI and NH children listening to natural language as an auditory stimulus and it is reassuring to see these regions highlighted by our algorithm as potential biomarkers corresponding to hearing impairment. Similarly, there is evidence of differential activation in subcortical features corresponding to the
auditory brainstem pathways. Features A, D, and J include elements of the reticular auditory pathway of the brainstem have been identified by electrophysiological studies to have a key role in auditory perception of location of sounds as well as the ability to filter a source of sound in background noise. Roughly these features appear to encompass key elements of the auditory pathway at the level of the pons (D) including the cochlear nucleus, trapaoidal body, lateral lemniscus and superior olive on the right, (A) inferior colliculus, medial geniculate on the left and (J) thalamus bilaterally (Kretschmann and Weinrich, 1998). Although the resolution of the fMRI scans (4×4×5mm) is not sufficient to resolve these structures individually, differences in activation in these regions, as indicated by reference to the higher resolution anatomical images, suggest that brain stem auditory nuclei may be involved.

One feature that is conspicuously absent from those illustrated in Figure 3-7 is the primary auditory cortex (BA41). We expected that this region would be important in differentiating HI from NH participants and hoped that it could potentially become a biomarker for predicting outcome for hearing and language following cochlear implantation in HI infants as suggested by our earlier work (Patel et al., 2007). The sedation used in the present study is a likely confounding to primary auditory function and may be partly responsible for the absence of a functional MRI feature in primary auditory cortex that differentiates the groups (DiFrancesco et al., 2013). However, because Figure 3-7 highlights differences between the groups that optimally separate them, it is possible that brain regions beyond primary auditory cortex that are responsible for recognizing sounds as speech and for extracting and associating content are more differentially stimulated in a scenario where the hearing impaired brain receives a rare auditory input that is above the threshold it can detect. Vibrations, loud noise and other stimuli may occasionally stimulate the auditory cortex in a deaf infant so that it is capable of processing
sound and responds during our experiment in the same manner as the NH children who are receiving sound stimulation at the same relative SPL. However, unless the HI infant is participating in a successful hearing aid trial, it is much less likely that they are routinely subjected to an auditory stream of speech that is consistently above their hearing threshold and hence unintelligible. HI infants in this study were all severe to profoundly hearing-impaired and ultimately received a cochlear implant because they did not derive sufficient benefit from an external hearing aid. Though this explanation is speculative, it could explain why features B, C, E, F, G, H, and I seem to be more important in separating the HI and NH groups of infants based on brain activation during fMRI.

On the other hand, our analysis on the fMRI data in this study also identified a number of areas that are not necessarily expected to play a role in differentiating HI from NH children. In particular, several functional features also appear in various portions of the anterior cingulate cortex (ACC, BA 24,32,33): areas associated with attention management, conflict monitoring, and error detection (Weissman et al., 2005). These features may be related to responses in the HI group to the novel auditory stimulus. ACC features are present in all three contrasts (C2, E1, E2, and G), suggesting a difference in response to sound input in the HI group who do not typically receive an auditory input at a level above their auditory threshold. Important features are also present in secondary visual cortex (H) (BA18), associative visual cortex (BA19) and other subcortical regions; differentiating the two groups. These features provide clues about additional neuroimaging biomarkers that may be relevant to the future use of functional neuroimaging to guide predictions about speech and language outcomes in HI infants who receive a cochlear implant.
3.5 Conclusion

First, our study demonstrates that HI and NH infants can be differentiated by brain MR images, e.g. different fMRI contrasts in auditory language network and auditory brain stem nuclei. Based upon the discriminative features, a classification model can be built to predict whether an individual has normal hearing or impaired hearing. The discriminative features may also be used as objective biomarkers of hearing loss or used for further disease mechanism studies. Secondly, our two-layer model integrates sMRI and fMRI in an effective way. While our sMRI classifier and fMRI classifier work moderately well individually, the combination of the two classifiers gives birth to a much more powerful classifier, which corroborates the hypothesis that integration of multiple modalities improves classification accuracy. Besides, our integration approach is very flexible, and it can be easily extended to include many diverse types of data.
Chapter 4 A comprehensive comparison of BoW approach with two traditional feature extraction approaches

4.1 Overview

In Chapter 3, we did not compare our BoW algorithm with other feature extraction algorithms, although we theoretically discussed the superiorities of BoW features over voxel features. The comparison was made in this chapter. Two traditional feature extraction approaches were considered. One was to use each single voxel as a feature. The other was to segment the brain into ROIs, and calculated the mean image intensities within ROIs as features. We employed the most commonly used brain segmentation, AAL atlas, in this study. With features extracted from the above three approaches, we applied various feature selection algorithms, and compared their discriminative power for the classification of normal hearing infants vs. hearing impaired infants. As reviewed in Chapter 2, the feature selection algorithms included Principal Component Analysis (PCA), selecting the most active voxels/regions, selecting the most discriminative voxels/regions, and Recursive Feature Elimination (RFE).

4.2 Data

The data used in the current study came from the same dataset as the data used in Chapter 3, also describe in (Smith et al., 2011). So the data acquisition procedure was the same as that in Chapter 3. However, we reviewed the original dataset, and reprocessed the images. The preprocessing in this chapter was done using the standard procedures in SPM8. Images were realigned to remove residual motion effects (Thevenaz et al., 1998), normalized to the infant template in the MNI space (Friston et al., 1995, Altaye et al., 2008), and smoothed with an 8 mm Gaussian kernel. Finally, the general linear model (GLM) (Worsley et al., 2002) was used to
construct three separate contrast maps (speech vs. silence, speech vs. noise and noise vs. silence) for each individual, which were used for the studies in the current chapter and the next chapter as well.

Forty-four subjects were selected for the preprocessing. Among the 44 subjects, there were 4 individuals with ages above 24 months, which was much higher than that of the remaining subjects. For the study in this chapter, we excluded those 4 subjects to make the ages matched between normal hearing infants and hearing impaired infants. However, all the 44 subjects were used for the study in Chapter 5, the reason for which was explained in Chapter 5. Therefore, the current study included 40 participants in total. Nineteen participants were hearing impaired infants with SNHL (10 females, average age=14.6 months, range=8-24 months). The remaining 21 participants were normal hearing (NH) controls (16 females, average age=12.1 months, range=8-17 months).

4.3 Feature extraction

4.3.1 Voxel feature (VOX)

The feature extraction for VOX was simple. Each single voxel became a feature. The value of the feature was the image intensity of the voxel in the contrast map.

4.3.2 AAL feature

Since both AAL atlas and the contrast maps were normalized to the MNI space, voxels of AAL atlas were aligned with the voxels of the contrast maps, and consequently the region labels of the AAL atlas can be transferred to the contrast maps directly. The AAL atlas segments the brain into 116 regions based on brain anatomy. For each region, we calculated the mean contrast by
averaging the contrast value of the voxels within the corresponding region. In this way, each contrast map can be converted into a 116 dimensional feature vector.

### 4.3.3 BoW feature

Since we re-preprocessed the images, the BoW algorithm was revised a little bit to make it suitable for smoothed contrast maps. Given a contrast map, a series of characteristic contrast regions were derived by thresholding the contrast map and subsequently merging the selected voxels into spatially coherent regions. Based on the Student’s t-distribution, we had 7 thresholds, i.e., 1.671, 2.000, 2.390, 2.660, 2.915, 3.232 and 3.460, corresponding to the one-sided p-values of 0.05, 0.025, 0.01, 0.005, 0.0025, 0.001 and 0.0005, respectively. The 7 thresholds gave rise to 7 contrast ranges \([1.671, 2.000), [2.000, 2.390), \ldots, [3.460, +\infty)\). We also considered the corresponding negative contrast ranges, \((-\infty, -3.460], \ldots, (-2.390, -2.000], (-2.000, -1.671], \) because the contrast map speech vs. silence included both speech>silence and silence>speech. The positive contrast ranges came from the contrast speech>silence, while the negative contrast ranges came from the contrast silence>speech. Each contrast range was considered separately.

Take the contrast range \([1.671, 2.000)\) for example. We first selected the voxels with contrast value within this range. The selected voxels were then merged into regions by connecting the voxels adjacent to each other in the 3D space, in which each voxel had 26 neighbors if it was not on the border. Voxels that were connected made up a region, i.e., a connected component, which was added into the dictionary as a word. We extracted all the words occurring across all the subjects and all the contrast ranges to build a dictionary for each type of contrast map. The dictionary was then applied to each subject to convert his/her corresponding type of contrast map into a feature vector. The value of a feature was calculated as the mean contrast value of the voxels within the corresponding region.
While Chapter 3 combined the three contrast types together to make a long concatenated feature vector, we separated the three contrasts in this chapter as well as in the following chapter. We analyzed two sets of BoW features. BoW21 constructed the dictionaries exclusively from the 21 NH subjects, while the BoW40 used all the 40 subjects including the 19 SNHL subjects to construct the dictionaries. Hypothetically, the activated/deactivated brain regions from the NH subjects should be included in the dictionaries, because they represented the actual brain activation pattern in response to external auditory stimuli in sedated infants and were likely to distinguish between NH vs. HI infants. With those considerations, we did not consider the feature set which constructed the dictionary using the 19 SNHL subjects alone as we expected such a feature set would miss a number of features that would be critical for the classification of NH vs. HI.

4.4 Dimensionality reduction & Feature selection

The goal of feature selection was to find a subset of features to approximate the relationships between the feature vectors and the response variable, instead of using all the features, which may include irrelevant features. We applied four dimensionality reduction or feature selection approaches in this chapter. There was a common parameter for the above four approaches, which was the number of features to be selected. Grid search was used to optimize this parameter. Specifically, we consider 10 different values ($\{2^1, 2^2, \ldots, 2^{10}\}$) for this parameter, and reported the best performance across the 10 parameters.

4.4.1 Principal Component Analysis (PCA)

PCA is a widely used approach for dimensionality reduction. We start with a data set which is represented as a $m \times n$ matrix $X$, with each row representing a sample and each column representing a feature. PCA can be used to linearly transform the matrix $X$ into a new matrix $Y$, 

62
with $Y = PX$, where $Y$ is also a $m \times n$ matrix and $P$ is a $m \times m$ matrix. While columns in $X$ might be correlated with each other, columns in $Y$ are independent of each other. Usually, the columns in $Y$ are ordered by their variances. For example, the first column in $Y$ is called the first principal component, which is a linear combination of the columns in $X$. The data samples have the highest variance on the direction of the first principal component, and then the second principal component, and so on. We may take the first $N$ columns of $Y$ to represent the original matrix $X$, which might have explained at least 90% of the variance/information in $X$. In this way, the dimensionality of the feature space is reduced. For the implementation of PCA, we used the `princomp` function in MATLAB.

### 4.4.2 Most active voxels/regions

Given a vector $v = \{v_1, v_2, \ldots, v_n\}$, where $v_i$ is the brain activation of the $i$-th subject, the mean activation across this group of subjects was calculated as:

$$S = \frac{\text{mean}(v)}{\sqrt{\text{Var}(v)/\text{length}(v)}}$$  \hspace{1cm} (4.1)

where $\text{mean}(v)$, $\text{Var}(v)$ and $\text{length}(v)$ are the mean, variance and length of the vector $v$, respectively. For each voxel, we calculated an $S$ score for the patient group as $S_{\text{patient}}$ and an $S$ score for the healthy control group as $S_{\text{control}}$. The merit for this voxel was calculated as: $(S_{\text{patient}} + S_{\text{control}})/2$. Voxels with high merits were selected for model training.

### 4.4.3 Most discriminative voxels/regions

Two-sample t-test was used to select the voxels with high discriminative power. The $T$ statistics was calculated as:
\[
T = \frac{|\text{mean}(\mathbf{v}_{\text{patient}}) - \text{mean}(\mathbf{v}_{\text{control}})|}{\sqrt{\frac{\text{Var}(\mathbf{v}_{\text{patient}})}{\text{length}(\mathbf{v}_{\text{patient}})} + \frac{\text{Var}(\mathbf{v}_{\text{control}})}{\text{length}(\mathbf{v}_{\text{control}})}}}
\]

(4.2)

where \(\mathbf{v}_{\text{patient}}\) and \(\mathbf{v}_{\text{control}}\) were constructed in the same way as the vector \(\mathbf{v}\) in last section. Voxels with high \(T\) statistics were selected.

### 4.4.4 Recursive Feature Elimination (RFE)

RFE performs feature selection by removing the irrelevant features iteratively. The algorithm is usually used in combination with SVM. It starts by training an SVM model with all of the features. Based on the trained model, features are ranked based on their absolute weights. Features with the lowest absolute weights are discarded, and a new SVM model with the new feature set is trained. This process is repeated until the number of features reached the predefined threshold. There are two parameters for the RFE algorithm: the percentage of features to be removed in each iteration and the final number of features to be kept. The first parameter primarily affects the speed of training. If this parameter is small, the training will be slow, and it is less likely to remove relevant features when compared to large values. In our project, we set this parameter at 1%, and required the algorithm to remove one feature at a time if the total number of features was less than 100. The second parameter was optimized by grid search as described above.

### 4.5 Model learning and Model evaluation

We used the linear SVM model for all feature sets. The models with different feature sets were evaluated using the leave-one-out cross-validation approach as described in Chapter 3. Sensitivity, specificity, accuracy and AUC were calculated as performance statistics. When
comparing among different models, we focused on accuracy and AUC, which were two summary statistics for the classification performance.

4.6 Results

4.6.1 Baseline performance without feature selection

The baseline performance without feature selection was shown in Table 4-1. The “Combine” feature sets were generated by concatenating the feature vectors from the three contrast types. Due to the high dimensionality of VOX, we did not consider the “Combine” feature set for VOX. From Table 4-1, we can see that the contrast speech vs. silence outperformed the other two contrasts, which was consistent across the four feature types. For the contrast speech vs. silence, VOX, BoW40 and BoW21 achieved comparable AUCs, but outperformed AAL. The contrast speech vs. noise demonstrated very poor performance, which might suggest that this contrast had very limited predictive power for the classification NH vs. HI. For BoW features, the Combine feature set achieved comparable AUC as the contrast speech vs. silence, but with improved accuracy, which might indicate that the contrast noise vs. silence included some predictive information that was complementary with the contrast speech vs. silence. Finally, the performance for BoW21 was comparable to BoW40, but with less number of features. Since both BoW21 and BoW40 were extracted by our BoW approach, we focused on BoW21 for the remaining analysis in this chapter.
Table 4-1. Baseline performance without feature selection

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. Silence</td>
<td>26767</td>
<td>63.2%</td>
<td>66.7%</td>
<td>65.0%</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>57.9%</td>
<td>42.9%</td>
<td>50.0%</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>21.1%</td>
<td>33.3%</td>
<td>27.5%</td>
<td>0.25</td>
</tr>
<tr>
<td>BoW40</td>
<td>Speech vs. Silence</td>
<td>1071</td>
<td>68.4%</td>
<td>61.9%</td>
<td>65.0%</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>1164</td>
<td>57.9%</td>
<td>47.6%</td>
<td>52.5%</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>739</td>
<td>26.3%</td>
<td>42.9%</td>
<td>35.0%</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>2974</td>
<td>73.7%</td>
<td>66.7%</td>
<td>70.0%</td>
<td>0.74</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. Silence</td>
<td>658</td>
<td>79.0%</td>
<td>57.1%</td>
<td>67.5%</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>63.2%</td>
<td>42.9%</td>
<td>52.5%</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>26.3%</td>
<td>38.1%</td>
<td>32.5%</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>79.0%</td>
<td>66.7%</td>
<td>72.5%</td>
<td>0.73</td>
</tr>
<tr>
<td>AAL</td>
<td>Speech vs. Silence</td>
<td>116</td>
<td>57.9%</td>
<td>61.9%</td>
<td>60.0%</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>116</td>
<td>47.4%</td>
<td>28.6%</td>
<td>37.5%</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>116</td>
<td>31.6%</td>
<td>28.6%</td>
<td>30.0%</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>348</td>
<td>47.4%</td>
<td>52.4%</td>
<td>50.0%</td>
<td>0.55</td>
</tr>
</tbody>
</table>

4.6.2  **Principal Component Analysis (PCA)**

Using PCA for feature selection, the performance for BoW21 had been significantly improved for the contrast speech vs. silence, and Combine. The AUCs were comparable for these two feature sets, but speech vs. silence showed higher accuracy than Combine. This indicated that most of the predictive power came from the contrast speech vs. silence. The other two contrasts might have some predictive power, but did not contribute a lot to the classification, when the three contrasts were integrated. For the other two feature types (VOX and AAL), the contrast speech vs. silence was clearly much better than the other two contrasts as well. If we focused on the contrast speech vs. silence, BoW21 was clearly better than VOX and AAL, although the performance of AAL had been improved greatly with feature selection.
Table 4-2. Classification performance using PCA for feature selection

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. Silence</td>
<td>26767</td>
<td>63.2%</td>
<td>76.2%</td>
<td>70.0%</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>79.0%</td>
<td>57.1%</td>
<td>67.5%</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>26.3%</td>
<td>38.1%</td>
<td>32.5%</td>
<td>0.31</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. Silence</td>
<td>658</td>
<td>73.7%</td>
<td>90.5%</td>
<td>82.5%</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>21.1%</td>
<td>47.6%</td>
<td>35.0%</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>57.9%</td>
<td>57.1%</td>
<td>57.5%</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>79.0%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.84</td>
</tr>
<tr>
<td>AAL</td>
<td>Speech vs. Silence</td>
<td>116</td>
<td>73.4%</td>
<td>71.4%</td>
<td>72.5%</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>116</td>
<td>42.1%</td>
<td>71.4%</td>
<td>57.5%</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>116</td>
<td>0.0%</td>
<td>100%</td>
<td>52.5%</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>348</td>
<td>57.9%</td>
<td>66.7%</td>
<td>62.5%</td>
<td>0.60</td>
</tr>
</tbody>
</table>

4.6.3 Most active voxels/ Most discriminative voxels/RFE

The results for these three feature selection algorithms are shown in Table 4-3, 4-4, and 4-5. We can see that BoW21 was always better than VOX and AAL.

Table 4-3. Classification performance by selecting the most active voxels

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. Silence</td>
<td>26767</td>
<td>5.3%</td>
<td>95.2%</td>
<td>52.5%</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>31.6%</td>
<td>90.5%</td>
<td>62.5%</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>10.5%</td>
<td>100%</td>
<td>57.5%</td>
<td>0.38</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. Silence</td>
<td>658</td>
<td>84.2%</td>
<td>76.2%</td>
<td>80.0%</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>84.2%</td>
<td>38.1%</td>
<td>60.0%</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>47.4%</td>
<td>57.1%</td>
<td>52.5%</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>84.2%</td>
<td>76.2%</td>
<td>80.0%</td>
<td>0.74</td>
</tr>
<tr>
<td>AAL</td>
<td>Speech vs. Silence</td>
<td>116</td>
<td>79.0%</td>
<td>57.1%</td>
<td>67.5%</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>116</td>
<td>68.4%</td>
<td>38.1%</td>
<td>52.5%</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>116</td>
<td>0.0%</td>
<td>90.5%</td>
<td>47.5%</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>348</td>
<td>57.9%</td>
<td>81.0%</td>
<td>70.0%</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Table 4-4. Classification performance by selecting the most discriminative voxels

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. Silence</td>
<td>26767</td>
<td>52.6%</td>
<td>52.4%</td>
<td>52.5%</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>52.6%</td>
<td>57.1%</td>
<td>55.0%</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>0.0%</td>
<td>100%</td>
<td>52.5%</td>
<td>0.50</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. Silence</td>
<td>658</td>
<td>84.2%</td>
<td>76.2%</td>
<td>80.0%</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>73.7%</td>
<td>66.7%</td>
<td>70.0%</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>36.8%</td>
<td>42.9%</td>
<td>40.0%</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>63.2%</td>
<td>66.7%</td>
<td>65.0%</td>
<td>0.67</td>
</tr>
<tr>
<td>AAL</td>
<td>Speech vs. Silence</td>
<td>116</td>
<td>63.2%</td>
<td>71.4%</td>
<td>67.5%</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>116</td>
<td>57.9%</td>
<td>66.7%</td>
<td>62.5%</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>116</td>
<td>21.1%</td>
<td>61.9%</td>
<td>42.5%</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>348</td>
<td>52.6%</td>
<td>61.9%</td>
<td>57.5%</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 4-5. Classification performance using RFE for feature selection

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. Silence</td>
<td>26767</td>
<td>52.6%</td>
<td>66.7%</td>
<td>60.0%</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>68.4%</td>
<td>52.4%</td>
<td>60.0%</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>47.4%</td>
<td>38.1%</td>
<td>42.5%</td>
<td>0.31</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. Silence</td>
<td>658</td>
<td>79.0%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>79.0%</td>
<td>61.9%</td>
<td>70.0%</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>42.1%</td>
<td>42.9%</td>
<td>42.5%</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>73.7%</td>
<td>66.7%</td>
<td>70.0%</td>
<td>0.78</td>
</tr>
<tr>
<td>AAL</td>
<td>Speech vs. Silence</td>
<td>116</td>
<td>73.7%</td>
<td>66.7%</td>
<td>70.0%</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>116</td>
<td>42.1%</td>
<td>61.9%</td>
<td>52.5%</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>116</td>
<td>5.3%</td>
<td>81.0%</td>
<td>45.0%</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>348</td>
<td>52.6%</td>
<td>61.9%</td>
<td>57.5%</td>
<td>0.57</td>
</tr>
</tbody>
</table>

4.7 Discussion

Compared with VOX features, our BoW features, which were calculated on the region-level, were much less sensitive to registration errors. An accurate alignment of images from different subjects is very difficult and currently not possible. In addition, BoW features considerably reduced the dimensionality of feature space, which not only made our classification problem easier to handle, but also helped to reduce the problem of over-fitting. At last, the classification
model based on BoW features was more interpretable, because it involved fewer features consisting of continuous regions instead of scattered voxels. These features can then be related more easily to disease etiology, diagnosis and prognosis.

AAL features were also extracted on the region-level and avoided the above disadvantages of VOX features. However, the AAL atlas was defined based on brain anatomy, without considering the activation pattern of the brain. An AAL region might mix-up the voxels with positive contrast and the voxels with negative contrast, which could negate the signal when we calculated the mean contrast for the region. In comparison, the BoW approach separated the positive contrast from negative contrast, and derived the activated/deactivated regions from the contrast maps themselves. Obviously, BoW regions were able to capture the activation patterns much better than AAL regions. That’s probably the reason why BoW features outperformed AAL features.

For the three contrasts, the contrast speech vs. silence was clearly more predictive than the other two contrasts. The contrast noise vs. silence also demonstrated some predictive power, while the contrast speech vs. noise showed very limited discriminative power. Without feature selection, the combination of three contrasts seemed to provide a better classifier than the one based on contrast speech vs. silence alone. Applying feature selection approaches, however, this improvement seemed to disappear.
Chapter 5 A semi-supervised SVM model for predicting the language outcomes following cochlear implantation based on pre-implant brain fMRI imaging

The work in this chapter is an extension of the work in previous two chapters. While the previous two chapters focused on the classification of Normal Hearing (NH) vs. Hearing Impaired (HI) infants, we developed a machine learning model in the current chapter to predict whether or not a cochlear implant (CI) candidate will develop effective language skills within two years after the CI surgery by using the pre-implant brain fMRI data from the candidate. The language performance was measured two years after the CI surgery by the Clinical Evaluation of Language Fundamentals-Preschool, Second Edition (CELF-P2). Based on the CELF-P2 scores, the CI recipients were designated as either effective or ineffective CI users. We trained both supervised models and semi-supervised models to classify CI users as effective or ineffective.

For feature extraction from the fMRI data, we constructed contrast maps using the general linear model, and then utilized the Bag-of-Words (BoW) approach that was used in the previous two chapters to convert the contrast maps into feature vectors. Compared with the conventional feature extraction approach, which used each single voxel as a feature, our BoW approach gave rise to much better performance for the classification of effective versus ineffective CI users. We also compared the predictive power for different auditory stimuli. The contrast of speech versus silence was shown to surpass the contrast of noise versus silence as well as that of speech versus noise. Furthermore, the semi-supervised model with the feature set extracted by the BoW approach from the contrast of speech versus silence achieved a leave-one-out cross-validation AUC as high as 0.97. Recursive feature elimination unexpectedly revealed that two features were
sufficient to provide highly accurate classification of effective versus ineffective CI users based on our current dataset. These two features are potential biomarkers for the prediction of CI outcomes. One of the two brain regions is located in the left hemisphere speech recognition and language association areas in the superior and middle temporal gyri, and aligns with our original hypothesis that brain activity in this area might be predictive of outcome from cochlear implantation in infants. The other brain region, located in right cerebellar structures, is somewhat surprising and may provide new insight about neural circuitry supporting language and audition in the developing brain. Additionally, our study also demonstrated the superiority of the semi-supervised model over the supervised model. It is always worthwhile to try a semi-supervised model when unlabeled data are available.

5.1 Introduction

As introduced in Chapter 3, cochlear implant (CI) is effective for restoring hearing, even in severely to profoundly deaf patients with hearing thresholds of 75dB HL and above (Geers et al., 2009, Hayes et al., 2009, Moog and Geers, 2010, Geers and Hayes, 2011). In infants and toddlers with pre-lingual or congenital sensori-neural hearing loss (SNHL), more than two decades of accumulated data show that many of these children can develop and continue to maintain good speech and language abilities with the use of a CI, even in the long term (Beadle et al., 2005, Geers et al., 2011, Geers and Sedey, 2011, Ruffin et al., 2013). However, variability in speech and language outcomes among this age group of CI patients remains high and individual outcomes may be difficult to predict (Zaidman-Zait and Most, 2005, Lazard et al., 2010, Niparko et al., 2010, Tobey et al., 2013). The reasons underlying the varied benefits across different individuals are not always clear. Furthermore, current behavioral methods used to predict language outcomes for a CI candidate prior to surgery may be inaccurate, particularly in infants.
Improved prognostic information would be helpful to clinicians and parents in setting expectations during the CI decision process, particularly given the high medical cost and anesthetic risks of this surgery. The motivation for the current study using fMRI and machine learning classification of pre-implant brain activation to auditory stimulation in CI candidates is to develop a neurobiological biomarker for speech and language outcomes.

Numerous studies investigating factors that influence language outcomes following cochlear implantation have been reported in the literature. Nikolopoulos et al. (Nikolopoulos et al., 1999) first studied the influence of age at implantation with 126 pre-lingually deafened children younger than 7 years of age at the time of implantation. Regression analysis and Spearman rank correlation coefficients revealed that language outcome was negatively correlated with age at implantation. Since then, several studies investigating the influence of age at implantation on speech and language outcomes have been published (Baumgartner et al., 2002, Manrique et al., 2004, Svirsky et al., 2004, Connor et al., 2006, Svirsky et al., 2007), for different age of participants, etiology of deafness and method for measuring language skills. In addition, mutations in gap junction protein beta2 (GJB2) were found to be a common cause of SNHL. Influence of GJB2 mutations on cochlear implantation outcomes was analyzed in (Bauer et al., 2003, Cullen et al., 2004, Sinnathuray et al., 2004a, Sinnathuray et al., 2004b), by comparing the language performances between groups with and without GJB2 mutations. Other influencing factors include inner ear malformation (Eisenman et al., 2001, Kim et al., 2006), meningitis (El-Kashlan et al., 2003), communication mode (oral vs. total) (Osberger et al., 1998, Osberger and Fisher, 2000, Kirk et al., 2002), pre-implant speech recognition skills (Zwolan et al., 1997, Osberger and Fisher, 2000) and pre-implant residual hearing (Gordon et al., 2001, Niparko et al., 2010), parent-child interactions (Niparko et al., 2010), socioeconomic status (Niparko et al.,
Approximately 50% of the variability in post-implant speech perception outcomes was explained by factors like duration of hearing loss before implantation, length of implant use, mode of communication and implant characteristics (Sarant et al., 2001). Although a variety of influencing factors have been investigated, a predictive model combining these variables has not been developed. Further, despite extensive pre-implant social, behavioral and clinical work-ups by pediatric cochlear implant teams, there continues to be variability in outcomes that does not appear to be accounted for by any of these parameters.

Within the past decade, functional Magnetic Resonance Imaging (fMRI) has been discussed as a way to assess auditory function in the brains of children as well as adults (Scheffler et al., 1998, Anderson et al., 2001, Lazeyras et al., 2002, Patel et al., 2007, Propst et al., 2010). With improvements in acquisition, preprocessing and analysis, it has been suggested that pre-implant fMRI could be translated into an objective predictor for CI outcomes (Patel et al., 2007). Indeed, the hypothesis motivating the design for our original fMRI study in infants with congenital SNHL was that the pre-implant cortical activation patterns revealed by fMRI during infancy would correlate with auditory performance two years after the CI surgery. Meanwhile, machine learning methods have begun to demonstrate success for analyzing neuroimaging data and show promise for translation of neuroimaging findings in populations to making predictions for individual patients (De Martino et al., 2008, Pereira et al., 2009, Cuingnet et al., 2011). In this chapter, we attempted to develop a machine learning model based on pre-implant fMRI data to predict the language outcomes two years after cochlear implantation in congenitally deaf infants with SNHL. The Support Vector Machine (SVM) model we developed uses pre-implant fMRI data from an individual CI candidate to predict whether or not the candidate will develop
effective language skills within two years after the cochlear implantation. This type of prognostic model could be extremely useful and is currently not available to clinicians by any other means.

5.2 Materials and Methods

5.2.1 Participants
As described in last chapter, the work in current chapter included 44 participants in total (Smith et al., 2011). Twenty-three participants had SNHL (12 females, average age=20.0 months, range=8-67 months). All of the SNHL children received the CI surgery. Their MRI data were acquired before the surgery. The remaining 21 participants were normal hearing (NH) controls (16 females, average age=12.1 months, range=8-17 months). The inclusion criteria for selecting participants were the same as that in Chapter 3, e.g., gestational age of at least 36 weeks and normal otoacoustic emissions for control group, normal brain anatomy for both patient group and control group, except that we extended the age range to maximize the sample size in the current chapter. As we noticed, the average age of the NH children was not perfectly matched to the average age of the SNHL children in the present study. However, this age difference will not invalidate our analysis as discussed in section 5.4 Discussion.

5.2.2 Cochlear implantation outcomes
Two years after the CI surgery, we administered a battery of tests to assess hearing, speech, language and cognitive function in the CI recipients. The tests were used to evaluate the CI recipients’ auditory, speech, and language outcomes following CI at a point in development when standardized behavioral measures of these skills could be used. For the current study, we used the data from the Clinical Evaluation of Language Fundamentals-Preschool, Second Edition (CELF-P2) (Wiig et al., 2004) as the primary language outcome measure of interest. The CELF-
P2 detects language delay or language impairment in children between the ages of three and seven years. The subtests of CELF-P2 focus on different language domains such as word structure, sentence structure, expressive vocabulary, concepts and following directions. These subtests help in the assessment of both receptive and expressive components of language. CELF-P2 is standardized on more than 1,500 children including children with hearing impairment (HI). Additionally, age equivalent norms are available for direct comparison between the target and control populations. CELF-P2 is routinely used in the clinic as a diagnostic as well as therapeutic tool to evaluate and monitor the progress in children’s language abilities. Thus, this valid and reliable test (Friberg, 2010) was used in the present study at the two-year follow-up stage to evaluate CI recipients’ language skills post-implantation. Sixteen out of the 23 CI recipients had 2-year follow-up scores for the CELF-P2. Follow-up scores for the remaining 7 children were not available for reasons such as family moving away from the area or a toddler unwilling to comply with the testing during a clinical follow up visit. There were 5 scores/indices for the CELF-P2, namely the core language score, receptive language index, expressive language index, language content index, and language structure index. The follow-up scores for the 16 participants are listed in Table 5-1.
Each of the five scores/indices of the CELF-P2 used in the present study provides a standard score ranging from 45-155 with a mean of 100 and a standard deviation of 15 (Table 5-1). The score 45 corresponds to a percentile rank of <0.1, while a score of 155 corresponds to a percentile rank of >99.9. In children with hearing loss and particularly with severe to profound congenital SNHL, the scores are substantially lower than the maximum value. The five scores/indices are highly correlated with each other. Based on the 16 samples in Table 5-1, we calculated the pair-wise correlations between the five scores/indices. The pair-wise Pearson’s correlation coefficient ranged from 0.90 to 0.95, and the Spearman’s correlation coefficient ranged from 0.87 to 0.92. Given the continuous outcome scores, a regression model might be more desirable than a classification model. However, regression function estimation is more challenging and requires more samples (Devroye et al., 1996, Wang et al., 2010). Considering our limited sample size as well as the conspicuous gap in the outcome scores (second to the last

<table>
<thead>
<tr>
<th>participant index</th>
<th>core language</th>
<th>receptive language</th>
<th>expressive language</th>
<th>language content</th>
<th>language structure</th>
<th>sum</th>
<th>effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>225</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>225</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>225</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>45</td>
<td>56</td>
<td>45</td>
<td>59</td>
<td>250</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>53</td>
<td>50</td>
<td>57</td>
<td>55</td>
<td>265</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>59</td>
<td>53</td>
<td>53</td>
<td>50</td>
<td>268</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>53</td>
<td>59</td>
<td>57</td>
<td>59</td>
<td>57</td>
<td>285</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>63</td>
<td>65</td>
<td>67</td>
<td>71</td>
<td>331</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>69</td>
<td>63</td>
<td>75</td>
<td>69</td>
<td>81</td>
<td>357</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>77</td>
<td>67</td>
<td>69</td>
<td>71</td>
<td>79</td>
<td>363</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>65</td>
<td>63</td>
<td>77</td>
<td>75</td>
<td>88</td>
<td>368</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>69</td>
<td>71</td>
<td>75</td>
<td>79</td>
<td>83</td>
<td>377</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>73</td>
<td>69</td>
<td>81</td>
<td>75</td>
<td>86</td>
<td>384</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>73</td>
<td>73</td>
<td>83</td>
<td>77</td>
<td>91</td>
<td>397</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>79</td>
<td>69</td>
<td>81</td>
<td>79</td>
<td>90</td>
<td>398</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>81</td>
<td>75</td>
<td>87</td>
<td>79</td>
<td>90</td>
<td>412</td>
<td>Yes</td>
</tr>
</tbody>
</table>
column in Table 5-1), we decided to train a classification model at present. To assign the class labels based on the CELF-P2 scores, we performed a k-means clustering with k=2. The two clusters were labeled as effective and ineffective CI users respectively. The effective group included 9 subjects (4 females, average age=21.1 months) with high follow-up scores, while the other cluster with 7 subjects (3 females, average age=19.7 months) was ineffective. As shown in Table 5-1, participants 1-7 were ineffective subjects with class label -1, and the remaining were effective subjects with class label +1. The class labels for the 7 children (5 females, average age=18.7 months) without follow-up scores were unknown. We then trained classification models to separate the effective from the ineffective CI-users.

It is well-known that pre-implant residual hearing is a good indicator for the subsequent success of cochlear implantation. Specifically, children with more residual hearing are likely to be effective CI-users, while those with less residual hearing tend to be ineffective CI-users. Our present work would be less meaningful, if the effective and ineffective CI-users in our project could be distinguished merely based on the pre-implant hearing thresholds. To exclude this possibility, we plotted the pre-implant hearing thresholds and post-implant language test scores in Figure 5-1. Obviously, the effective and ineffective CI-users were not separable based on the pre-implant hearing thresholds alone. Further, our previous regression analysis using age at implantation and pre-implant hearing threshold as independent variables also failed to predict the post-implant language test scores with a satisfactory accuracy.
Figure 5-1. The pre-implant hearing threshold and post-implant language test score for the participants. The post-implant language test score is the sum of the five CELF-P2 test scores (Table 5-1). Red circle points represent ineffective CI-users, Green square points represent effective CI-users. Blue diamond points represent unlabeled samples. Since the unlabeled samples do not have the post-implant language test scores, we set their values to be 300 in order to show them in this figure. Furthermore, two out of the seven unlabeled samples have the same pre-implant hearing threshold of 100. They are overlapped in the figure. Thus, it appears that there are only 6 diamond points in the figure.

5.2.3 MRI data acquisition & preprocessing

This section is exactly the same as that in Chapter 4.

5.2.4 Feature extraction

The feature extraction approach is exactly the same as that in Chapter 4. Also similar to Chapter 4, we analyzed two sets of BoW features in the current chapter. BoW21 constructed the dictionaries exclusively from the 21 NH subjects, while BoW44 used all the 44 subjects including the 23 SNHL subjects to construct the dictionaries.
For comparison purposes, we also trained models with voxel features. In this approach, each voxel became a single feature, the value of which was the image intensity of this voxel in the contrast maps (Ryali et al., 2010, Nouretdinov et al., 2011, Rizk-Jackson et al., 2011, Brodersen et al., 2012, Oliveira et al., 2013, Hart et al., 2014). For convenience, we denoted this approach as VOX.

5.2.5 Supervised SVM model

The basics for the supervised linear SVM model was described in section 2.5 Model learning. The parameter $C$ was set to be 1 according to our previous experiences.

5.2.6 Semi-supervised SVM model

The inputs for the semi-supervised SVM model training are a set of labeled samples $D_1 = \{(X_1, y_1), \ldots, (X_M, y_M)\}$ and a set of unlabeled samples $D_2 = \{X_1, \ldots, X_K\}$, where $M$ is the number of labeled samples and $K$ is the number of unlabeled samples. In our project, we had 7 unlabeled samples, whose follow-up scores were missing. The semi-supervised model has the same format as that of the supervised model, namely $\hat{y} = wX + b$, but with a different objective function as defined in Equation (5.1).

$$
\begin{align*}
\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{M} \xi_i + U \sum_{j=1}^{K} \xi_j^* \\
\text{s.t. } \forall_{i=1}^{M}: y_i(wX_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \\
\forall_{j=1}^{K}: y_j(wX_j + b) \geq 1 - \xi_j^*, \xi_j^* \geq 0 \\
\frac{1}{K} \sum_{j=1}^{K} \max(0, y_j) = r
\end{align*}
$$

(5.1)

where the first term measures the margin, the second and third term measure the training errors on labeled samples and unlabeled samples respectively, $C$ is the regularization parameter for the
labeled samples, and $U$ is the regularization parameter for unlabeled samples, $r$ is the ratio of positive samples within the unlabeled samples, which is a user specified parameter. The parameter $C$ was set to be 1 as the supervised model. The parameter $U$ is usually set to be equal to $C$, which is also followed by the SVM\textsuperscript{light} package.

The training process of a semi-supervised model includes three steps: (1) Train a supervised model based on the labeled samples, followed by application of this initial model to the unlabeled samples. The $K \times r$ unlabeled samples with the highest predicted scores are assigned to +1, and the remaining are assigned to -1. (2) Assign a temporary parameter $U_{tmp}$ and re-train a new model with all of the samples, including the unlabeled samples. Switch the labels of a pair of unlabeled samples according to a certain rule to achieve the maximal drop of the objective function in Equation (5.1). Repeat this process until no pair of the unlabeled samples meets the switching rule. (3) Increase the value of $U_{tmp}$ and then repeat step (2). When $U_{tmp} \geq U$, terminate the training process and output the model. In this project, both the supervised and semi-supervised models were trained using the SVM\textsuperscript{light} package (Joachims, 1999). Linear kernel was used for the SVM models.

5.2.7 Feature selection

We employed the Recursive Feature Elimination (RFE) approach for feature selection. The basic idea of RFE was described in section 4.4.4 Recursive Feature Elimination (RFE).

5.2.8 Predicting new subjects

Given a new sample, whose class label was unknown while feature vector $X$ was available, we calculated the $\hat{y}$ by inserting $X$ into the learned model $\hat{y} = wX + b$. If $\hat{y} \geq \delta$, the new sample was predicted to have class label +1. Otherwise, its class label was predicted to be -1. The
threshold $\delta$ was set to be 0 by default. The semi-supervised model and the supervised model had different $w$ and $b$, and consequently their predictions were different.

5.2.9 Model evaluation

We employed the Leave-One-Out Cross-Validation (LOOCV) approach to evaluate the supervised models as well as the semi-supervised models when the parameter $r$ was already known before the model training. For the supervised model, we performed a LOOCV on the 16 labeled samples. Specifically, we performed 16 rounds of training and testing, each round of which was called one fold of cross-validation. In the $k$-th fold, the $k$-th labeled subject was used for testing, and the remaining 15 subjects were used for training. In this way, each labeled subject was used for testing once. The model was evaluated based on the predictions on the 16 testing samples accumulated across the 16 folds of cross-validation. For the semi-supervised model with parameter $r$ set to be a predefined value, the evaluation approach was generally the same as that applied to the supervised model. We also performed 16 rounds of training and testing, with each of the labeled samples left-out for testing once. The only difference was that, for each fold of cross-validation, the semi-supervised model had 22 training samples, including 15 labeled samples and 7 unlabeled samples, while the supervised model had only 15 training samples. Please note the unlabeled samples were used only for training but not for testing, and both the supervised model and semi-supervised model were tested on the same 16 labeled samples.

For the semi-supervised model whose parameter $r$ was unknown before the model training and needed to be optimized during the model training, we employed a nested LOOCV approach for model evaluation. A flowchart for the nested LOOCV is shown in Figure 5-2. As we see from Figure 5-2, the nested LOOCV consisted of an outer LOOCV and an inner LOOCV. The
outer LOOCV was used for model evaluation, and the inner LOOCV was used for parameter selection. In the $k$-th iteration of the outer LOOCV, the $k$-th labeled subject was used for testing, and the remaining 22 subjects, including 15 labeled samples and 7 unlabeled samples, were used for training and parameter selection. Since parameter $r$ was the ratio of positive samples within the unlabeled samples, we considered $r$ starting from 0 to 1, stepping by 0.1, which gave us 11 different values for $r$. Under each $r$ value, we performed an inner LOOCV on the 22 subjects, which was the same as the LOOCV procedure described in the previous paragraph except that we had only 15 instead of 16 rounds of training and testing here, because the 22 subjects only included 15 labeled samples. Then we selected the $r$ value that led to the highest F-measure, which was calculated based on the predictions on the 15 labeled samples. The F-measure was defined in Equation (5.2).

$$
sensitivity = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}
$$

$$
specificity = \frac{\text{true negative}}{\text{true negative} + \text{false positive}}
$$

$$
F\text{-measure} = \frac{2 * sensitivity * specificity}{sensitivity + specificity}
$$

(5.2)

The F-measure defined as the harmonic mean of sensitivity and specificity promotes a balance between sensitivity and specificity of the classifier, and is often used as a single measure of classification performance. Using the selected $r$ value, we trained a model with the 22 subjects, and applied the trained model to the $k$-th labeled subject for testing. Thus, we completed one iteration of the outer LOOCV. There were 16 iterations for the outer LOOCV, with each of the 16 labeled samples left-out for testing once. The nested LOOCV was used only when the parameter $r$ was tuned automatically. If $r$ was predefined to be a fixed value such as 0.6, the inner LOOCV was no longer needed because there was no need to optimize the parameter $r$ and
the LOOCV as described in the previous paragraph was used for the model evaluation in such a case. More concretely, the nested LOOCV was only used in the section Semi-supervised model with automatic parameter selection. Based on the results in this section, parameter $r$ was set to be predefined values in the remaining analysis of this chapter, and therefore the LOOCV was used for model evaluation after this section, except for the blind test on NH subjects.

Based on the predictions on the testing samples, we calculated the sensitivity, specificity, accuracy and area under the receiver operating characteristic curve (AUC) to estimate the generalization performance of the models. Furthermore, we calculated the Pearson’s correlation coefficient (Pcorr) as well as the Spearman’s rank correlation coefficient (Scorr) between the predicted scores and the sum of the CELF-P2 test scores, which were listed in the second to the last column in Table 5-1. Pcorr was used to quantify the linear relationship, while Scorr was used to measure the monotonic relationship. A model with high correlations was desirable.
Figure 5-2. A flowchart for the nested LOOCV that we used to evaluate the semi-supervised model when the parameter $r$ was unknown before the model training. The outer LOOCV included 16 rounds of training and testing. For each round, one labeled sample was left-out for testing, and the remaining 15 labeled samples and the 7 unlabeled samples were used for training. We submitted these 22 samples to the inner LOOCV box. All the activities in the inner LOOCV box were confined to these 22 samples. The goal of the inner LOOCV box was to optimize the parameter $r$. Different $r$ values, ranging from 0 to 1 and stepping by 0.1, were tried in the inner LOOCV box. Specifically, we performed a LOOCV on the 22 samples for each value of $r$. An inner LOOCV included 15 rounds of training and testing. For each round, one labeled sample was left-out for testing, and the remaining 14 labeled samples and the 7 unlabeled samples were used for training. The LOOCV performance was calculated based on the predictions on the 15 testing samples accumulated across the 15 rounds of testing. The $r$ value that achieved the best LOOCV performance was selected. With the optimized parameter $r$, we jumped out of the inner LOOCV box and returned to the outer LOOCV box. Then we trained a model with the optimized parameter $r$ using the 7 unlabeled samples and the 15 labeled samples in the blue ellipse, and applied this model to the left-out labeled sample for testing. Thus, we completed one round of training and testing for the outer LOOCV. This process was repeated 16 times, with each labeled sample left-out for testing once. The predictions on those 16 testing samples were used for the evaluation of the model.
5.3 Results

5.3.1 Supervised model

The LOOCV performance for the supervised SVM models is shown in Table 5-2. Among the performance measures, we focused on Pcorr and Scorr. We can see that the contrast speech vs. silence outperformed the other two contrasts, which was consistent across the three types of features. The VOX approach obtained almost zero correlations for the contrasts noise vs. silence and speech vs. noise, which suggested that these two contrasts had very limited predictive power. Combining these two contrasts with the contrast speech vs. silence did not improve the classification performance, which implied that including these two contrasts in the training added noisy features and disturbed the classification. In comparison with the VOX approach, our BoW approaches achieved better performance for all the three contrasts. Additionally, the BoW21 returned comparable performance as BoW44, but with fewer features.

Table 5-2. LOOCV performance for the supervised SVM models. The feature vectors from the three contrasts were concatenated to form the feature vector for the “Combine”. Due to the large number of features, we did not train the model with the three contrasts combined for the VOX.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Pcorr</th>
<th>Scorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. silence</td>
<td>26767</td>
<td>33.3%</td>
<td>42.9%</td>
<td>37.5%</td>
<td>0.51</td>
<td>0.29</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>44.4%</td>
<td>42.9%</td>
<td>43.8%</td>
<td>0.54</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>66.7%</td>
<td>57.1%</td>
<td>62.5%</td>
<td>0.59</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>BoW44</td>
<td>Speech vs. silence</td>
<td>1156</td>
<td>33.3%</td>
<td>71.4%</td>
<td>50.0%</td>
<td>0.56</td>
<td>0.35</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>1216</td>
<td>55.6%</td>
<td>71.4%</td>
<td>62.5%</td>
<td>0.63</td>
<td>0.21</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>803</td>
<td>66.7%</td>
<td>71.4%</td>
<td>68.8%</td>
<td>0.67</td>
<td>0.17</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>3175</td>
<td>77.8%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.65</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. silence</td>
<td>658</td>
<td>44.4%</td>
<td>85.7%</td>
<td>62.5%</td>
<td>0.73</td>
<td>0.38</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>55.6%</td>
<td>57.1%</td>
<td>56.3%</td>
<td>0.60</td>
<td>0.21</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>66.7%</td>
<td>71.4%</td>
<td>68.8%</td>
<td>0.68</td>
<td>0.19</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>66.7%</td>
<td>57.1%</td>
<td>62.5%</td>
<td>0.59</td>
<td>0.16</td>
<td>0.01</td>
</tr>
</tbody>
</table>

5.3.2 Semi-supervised model with automatic parameter selection

For the experiment in the current section, the parameter $r$ was optimized automatically. The model was evaluated using the nested LOOCV approach (Figure 5-2). Results are shown in...
Table 5-3. In comparison with the supervised models whose performance is shown in Table 5-2, the semi-supervised models achieved comparable performance for VOX, and better performance for both BoW21 and BoW44 across all of the three contrasts. We also noticed that the correlations (Pcorr and Scorr) for the contrast speech vs. silence were considerably higher than the other two contrasts (Table 5-3). Furthermore, the correlations for BoW21 were higher than or at least as good as the BoW44 across different contrasts. The BoW21 with the contrast speech vs. silence achieved AUC of 0.97, Pcorr of 0.60 and Scorr of 0.76, which significantly surpassed other combinations. Analyzing its automatically selected parameters across different folds of cross-validation, we noticed that 15 out of 16 folds had selected 0.6 for the ratio \( r \), which corresponded to 4 effective and 3 ineffective children within the unlabeled samples. This ratio was very reasonable, because it was consistent with the ratio on the labeled samples (9 effective vs. 7 ineffective). The labeled samples and unlabeled samples came from the same distribution, and therefore their ratios of positive samples were expected to be close to each other. The only fold that selected a different ratio had selected the ratio 0.7 corresponding to 5 effective and 2 ineffective, which was very close to the ratio selected during other folds. Hence, we were confident that the parameter selection process was very stable across different folds of cross-validation. For the BoW44 with contrast speech vs. silence, the parameter selection was not as stable as BoW21, but it also selected the ratio 0.6 in 12 out of 16 folds. The VOX approach for the contrast speech vs. silence selected the ratio 0.6 in only 4 folds. Five other ratios were selected at least once, but none of them were selected for more than 4 folds. Therefore, the parameter selection was considerably unstable for the VOX approach, which might be a reason why the semi-supervised model with VOX features did not achieve an improvement in classification performance when compared with the corresponding supervised model.
Table 5-3. Nested LOOCV performance for the semi-supervised models. The parameter ratio $r$ was selected automatically.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Contrast</th>
<th># of Features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Pcorr</th>
<th>Scorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>Speech vs. silence</td>
<td>26767</td>
<td>44.4%</td>
<td>42.9%</td>
<td>43.8%</td>
<td>0.56</td>
<td>0.23</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>26767</td>
<td>44.4%</td>
<td>85.7%</td>
<td>62.5%</td>
<td>0.57</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>26767</td>
<td>66.7%</td>
<td>57.1%</td>
<td>62.5%</td>
<td>0.63</td>
<td>0.17</td>
<td>0.06</td>
</tr>
<tr>
<td>BoW44</td>
<td>Speech vs. silence</td>
<td>1156</td>
<td>66.7%</td>
<td>71.4%</td>
<td>68.8%</td>
<td>0.73</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>1216</td>
<td>55.6%</td>
<td>85.7%</td>
<td>68.8%</td>
<td>0.73</td>
<td>0.37</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>803</td>
<td>77.8%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.71</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>3175</td>
<td>77.8%</td>
<td>57.1%</td>
<td>68.8%</td>
<td>0.65</td>
<td>0.19</td>
<td>0.16</td>
</tr>
<tr>
<td>BoW21</td>
<td>Speech vs. silence</td>
<td>658</td>
<td>77.8%</td>
<td>85.7%</td>
<td>81.3%</td>
<td><strong>0.97</strong></td>
<td><strong>0.60</strong></td>
<td><strong>0.76</strong></td>
</tr>
<tr>
<td></td>
<td>Noise vs. Silence</td>
<td>806</td>
<td>55.6%</td>
<td>71.4%</td>
<td>62.5%</td>
<td>0.70</td>
<td>0.37</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Speech vs. Noise</td>
<td>434</td>
<td>77.8%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.75</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>1898</td>
<td>77.8%</td>
<td>57.1%</td>
<td>68.8%</td>
<td>0.78</td>
<td>0.32</td>
<td>0.22</td>
</tr>
</tbody>
</table>

5.3.3 Semi-supervised model with fixed parameter $r$

Since the contrast speech vs. silence outperformed the other two contrasts and all the three feature sets selected the ratio $r$ of 0.6 most frequently for this contrast, we performed a LOOCV on the contrast speech vs. silence with the parameter $r$ fixed to be 0.6. Results for this experiment are shown in Table 5-4. The performance of BoW21 did not change when compared to its performance in Table 5-3. This was expected because only one fold had changed the ratio from 0.7 to 0.6. However, locking $r=0.6$ for the BoW44 model changed the ratio in 4 folds, and achieved better performance. VOX feature set achieved significantly higher AUC and correlations, when comparing its performance in Table 5-4 with that in Table 5-3. To make our comparisons more convincing, we used the R package pROC (Robin et al., 2011) to compare two AUCs. A two-sided p-value was calculated to show if two AUCs were significantly different. We set the significance level at 0.05. In comparison with the supervised models, the AUCs for the semi-supervised models were significantly higher for BoW44 (p-value=0.02) and BoW21 (p-value=0.04), but not for VOX (p-value=0.09). Comparing among the semi-supervised models with different feature sets, the AUC for BoW21 was shown to be significantly higher than the
AUC of VOX (p-value=0.03), but not significantly higher than the AUC of BoW44 (p-value=0.08). Meanwhile, we did not detect significant difference between BoW44 and VOX for the semi-supervised models. Since the combination of BoW21 and contrast speech vs. silence was consistently better than all other combinations, we focused on this feature set in the remaining analysis of this chapter, with the ratio of positive samples within the unlabeled samples fixed at 0.6 unless stated explicitly as a different value.

Table 5-4. LOOCV performance for the semi-supervised models with $r = 0.6$ for the contrast speech vs. silence.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Pcorr</th>
<th>Scorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOX</td>
<td>66.7%</td>
<td>42.9%</td>
<td>56.3%</td>
<td>0.70</td>
<td>0.49</td>
<td>0.42</td>
</tr>
<tr>
<td>BoW44</td>
<td>77.8%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.83</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>BoW21</td>
<td>77.8%</td>
<td>85.7%</td>
<td>81.3%</td>
<td>0.97</td>
<td>0.60</td>
<td>0.76</td>
</tr>
</tbody>
</table>

5.3.4 Why did semi-supervised model perform better?

Adding the seven unlabeled samples to the training set increased the sample size by approximately 45%. We speculated that these unlabeled samples contributed to the characterization of the distribution of the two groups of samples, e.g., effective and ineffective CI-users, and consequently helped to identify the accurate hyperplane separating the two groups of samples. To support our speculation, we analyzed the models across different folds of cross-validation. As described above, the models were defined by the weight vector $w$ and the bias $b$. Considering that $w$ specified the orientation of the separating hyperplane, we measured the difference/distance between two $w$ vectors as $1 - \cos \theta$, where $\theta$ was the included angle between the two $w$ vectors. Each fold of cross-validation yielded a model, and each model had a weight vector $w$. We had 16 models, each of which was trained with one of the 16 labeled samples left out for testing, and then we obtained 120 unique distances by calculating the pairwise distance among the 16 weight vectors. Thus, we had 120 distances for the supervised
approach and the semi-supervised approach, respectively. We noticed that the 120 distances were not independent of each other. For example, let three models be \( m_1, m_2, \) and \( m_3 \). Let the distance between \( m_1 \) and \( m_2 \) be \( d_{12} \), the distance between \( m_2 \) and \( m_3 \) be \( d_{23} \), and the distance between \( m_1 \) and \( m_3 \) be \( d_{13} \). According to the triangle inequality theorem, we would know that 

\[
|d_{12} - d_{23}| < d_{13} < d_{12} + d_{23}.
\]

In other words, \( d_{13} \) is not independent of \( d_{12} \) and \( d_{23} \).

Therefore, the paired two sample t-test or Wilcoxon signed-rank test, which required the samples to be independent to each other, was not appropriate for testing if the means of the 120 distances were significantly different between supervised and semi-supervised model. Therefore, we compared the two sets of distances directly by subtracting the semi-supervised distance from the supervised distance, as shown in Figure 5-3. Among the 120 distances, the semi-supervised approach was only slightly higher than the supervised approach for 10 distances, suggesting that the weight vectors from the semi-supervised models were much more similar to each other than that of the supervised models across different folds of cross-validation. Furthermore, we also calculated the standard deviation (std) for the bias \( b \) across different folds of cross-validation. The std for the supervised model was 0.08, which was also higher than the 0.04 for the semi-supervised model. These results confirmed that the semi-supervised models were more stable and consistent across different folds of cross-validation.
Figure 5-3. Stability of the weight vectors across different folds of cross-validation. One-hundred-and-twenty unique pair-wise distances were calculated among the weight vectors, for the semi-supervised model and the supervised model respectively. Each bar represents one of the 120 unique distances. y-axis is the distance from the supervised model minus the distance from the semi-supervised model.

Figure 5-4 shows the distance matrices calculated from the weight vectors as described above. From this figure, we can see that the hyperplane learned by the supervised learning for folds 1, 8, and 13 had considerable deviation from other folds, while the semi-supervised model only had a small deviation for the fold 13. When the hyperplane deviated by a large amount, the predicted value for the testing sample was very inaccurate, which had a great effect on the AUC score, given the limited sample size. AUC, the area under the receiver operating characteristic (ROC) curve, evaluates the ranking of the predicted values for the testing samples. If the predicted value for any negative sample is lower than the predicted value for any positive sample, then it is a perfect ranking and the AUC will reach 1. Figure 5-5 shows the ROC curves for the supervised model and semi-supervised model. In the ROC curve, each vertical segment represents a positive sample and each horizontal segment represents a negative sample. As shown in Figure 5-5, we marked the 1st, 8th, and 13th testing sample on the ROC curve of the
supervised model. We can see that the 1\textsuperscript{st} and 8\textsuperscript{th} testing samples were positive samples, but the predicted values for these two samples were lower than almost all negative samples. As a result, the ROC curve cannot climb higher than 0.8 until the false positive rate reached as high as 0.8. The 13\textsuperscript{th} sample was a negative sample, but its predicted value was higher than most positive samples. Accordingly, the ROC curve was forced to move horizontally even though the true positive rate was relatively low. Due to the influence of these three samples, the area under the ROC curve cannot be high for the supervised model. Conversely, the ranking of the predicted values from the semi-supervised model was very close to the perfect ranking, except that the predicted value for the 13\textsuperscript{th} sample was higher than two positive samples but lower than all other positive samples. This explains the very high AUC of the semi-supervised model.

![Pair-wise distance among the weight vectors across different folds of cross-validation.](image)

**Figure 5-4.** Pair-wise distance among the weight vectors across different folds of cross-validation. The left panel is for the supervised model, and the right panel for the semi-supervised model. Each panel is a symmetrical distance matrix. The color represents the distance, with hot colors representing large distance values and cool colors representing small distance values. The two panels use the same color scale as shown by the color bar.
Figure 5. ROC curves for the supervised model and semi-supervised model. The models were trained on the feature set extracted by the BoW21 from the contrast speech vs. silence. The parameter $\tau$ for the semi-supervised model was fixed to be 0.6. The ROC curves and AUCs were derived based on the predicted values on the testing samples with the LOOCV approach.

We show a two-dimensional toy example in Figure 5 to make the above explanation more straightforward. In this example, we also had 9 positive, 7 negative and 7 unlabeled samples. The ratio of positive samples on the unlabeled samples was set to be 0.6. For this example, the supervised model achieved LOOCV AUC of 0.82, while the semi-supervised model achieved an AUC of 0.92. From the hyperplanes learned across different folds of cross-validation, we noticed that samples A and B were two special samples. They had determined that the area covered by the yellow ellipse belonged to the positive group. If the sample A or B was left out for testing, the supervised model had only one training sample within the yellow ellipse, leaving this area almost uncovered by the training samples. Based on the distribution of training samples, the model learning algorithm was likely to derive a hyperplane classifying the yellow ellipse as negative. Thus, the learned hyperplane deviated considerably from the actual hyperplane represented by the red line in the figure. Thanks to the unlabeled samples, especially
the two within the yellow ellipse, the distribution of samples stayed unchanged after leaving sample A or B out for testing. Accordingly, the semi-supervised model learning algorithm returned a hyperplane very close the actual hyperplane. In addition, we noticed that 5 out of 7 unlabeled samples were classified to the positive group in the semi-supervised model, although the user specified ratio was 4 positive vs. 3 negative. This observation implied that the semi-supervised learning does not force the model to follow the pre-specified ratio. On the contrary, it considers the sample distribution in general and finds out the hyperplane minimizing the objective function, even at the cost of some training errors from the unlabeled samples.

Based on the evidence above, we concluded that the semi-supervised model is more stable and accurate when compared to the supervised model. Due to the high feature-to-sample ratio, the distribution of samples in the feature space was not apparent. In such case, the supervised model is sensitive to the elimination of certain samples. For the semi-supervised model, however, including the unlabeled samples makes the distribution of samples better-defined. Thus, the semi-supervised model remains stable even after removing some samples from the training set.
Figure 5-6. A two-dimensional example for the comparison between supervised model and semi-supervised model. The red line represents the hyperplane learnt with all available data, 16 samples for the supervised model and 23 samples for the semi-supervised model. The two black lines in each panel represent the hyperplanes learnt with sample A or B left out for testing, respectively. Since the two hyperplanes are very close to each other, they appear to be completely overlapped.
5.3.5 Blind test on normal hearing infants

Although the semi-supervised model was indeed more stable across different folds of cross-validation, one could still argue that the supervised model trained with all of the labeled samples would be as accurate as the semi-supervised model. To rule out this possibility, we performed a blind test on the NH samples. Theoretically, all of the NH participants had normal hearing and should be categorized to the effective group, although they did not have CELF-P2 scores to provide a label. As we described in section 5.2.4 and section 4.3.3, we had constructed a dictionary for the contrast speech vs. silence. This dictionary was applied to each NH infant to convert his/her contrast map into a 658-dimensional feature vector. We then trained a supervised model with the 16 labeled samples, as well as a semi-supervised model with all of the 23 samples, including 7 unlabeled samples, and applied these two classification models to classify the NH infants. The supervised model successfully classified 66.7% of the NH infants as effective, while the semi-supervised model correctly classified 81.0%, which was close to the LOOCV classification accuracy on the labeled samples. This result further confirmed that the semi-supervised model was more accurate than the supervised model.

5.3.6 Semi-supervised model by adding the unlabeled samples one by one

We investigated how the classification performance changed as we added the unlabeled samples one by one. As we changed the set of unlabeled samples to be added to the training, the parameter $r$ changed accordingly. If parameter $r$ was unknown, we needed to perform nested LOOCV to evaluate the model. For simplicity and consistency, we first trained a semi-supervised model using all of the 23 samples to assign labels for the unlabeled samples. Subsequently, the predicted labels for the unlabeled samples were used to calculate the parameter $r$, which was used as the input for model training for the experiment in this section. For example,
we first added one unlabeled sample, which gave us a sample size of 17 including 16 labeled samples and 1 unlabeled sample. If this unlabeled sample was an effective CI-user according to the predictions above, we set parameter $r$ to be 1.0. After that, we added the 2\textsuperscript{nd} unlabeled sample. If the 2\textsuperscript{nd} unlabeled sample was an ineffective CI-user, we set the parameter $r$ to be 0.5, given the 1\textsuperscript{st} unlabeled sample was an effective CI-user. Then, we added the 3\textsuperscript{rd} unlabeled sample, 4\textsuperscript{th} unlabeled sample and so on. Models were tested using the LOOCV approach, since $r$ was already determined before the model training. The order in which the samples were added was generated randomly. Considering the possible bias effect of the order, we tried two sets of orders, and added the samples in both forward order and reverse order. The results are shown in Table 5-5. The sample indexes for the two orders were not the same, e.g., the 7\textsuperscript{th} sample from Order 1 and the 7\textsuperscript{th} sample from Order 2 were not the same sample. From Table 5-5, we found that the classification performance was the same as or very similar to that of the supervised model, when we added only one or two unlabeled samples. Adding the 4 positive unlabeled samples alone also did not improve the classification performance. Meanwhile, the improvement was negligible when only the 3 negative unlabeled samples were added to the training set. The improvement became significant only when we added both positive and negative unlabeled samples. Furthermore, the performance improved gradually as we added more samples. There seemed to be a jump in performance when we added 5 or 6 unlabeled samples. We speculated that adding 5 or 6 unlabeled samples reached a point where the sample distributions were well-defined and the distinction between the two groups of samples became obvious, which might have explained the performance jump.
**Table 5-5. LOOCV performance of the semi-supervised model when adding the unlabeled samples one by one.**

<table>
<thead>
<tr>
<th>Random Order</th>
<th>Forward order</th>
<th>Reverse Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample index</td>
<td>Label</td>
<td>Ratio</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>0.6</td>
</tr>
</tbody>
</table>

**5.3.7 Semi-supervised model with artificial unlabeled samples**

We also investigated whether the improvement in classification performance was dependent on these particular 7 unlabeled samples. Could we still obtain a performance improvement if we were given another set of unlabeled samples? We assumed that the labeled samples were good representations of the actual sample populations, e.g., CI effective users and CI ineffective users. Under this hypothesis, the mean and standard deviation for each sample population can be estimated using the labeled samples. Thus, the simulated samples were generated by adding white noise to the population mean as formulated in Equation (5.3).

\[
x = m + \text{alpha} \times s \times \text{rand}
\]  

(5.3)

where \(x\) is the feature vector for the simulated sample, \(m\) and \(s\) are the mean and standard deviation of the labeled samples respectively, \(x, m, s\) are 658-dimensional vectors since there are 658 features for the contrast speech vs. silence. \(\text{rand}\) is also a 658-dimensional vector containing...
pseudorandom values drawn from the standard normal distribution, which was generated by the `randn` function in MATLAB. \( alpha \) is a parameter used to control the noise strength. We estimated the mean \( m \) and standard deviation \( s \) for the CI-effective group and CI-ineffective group separately, and generated 5 simulated samples for each group. Thus, we had 26 samples, including 16 labeled samples and 10 unlabeled samples. A semi-supervised model with parameter \( r \) equal to 0.5 could be trained. Still, the LOOCV approach was used to evaluate the model. For each \( alpha \), we performed this experiment 10 times and calculated the mean and standard deviation for the performance measures across the 10 runs. The results are shown in Table 5-6.

We can see that adding the simulated unlabeled samples to the training set helped to improve the classification performance when compared to the supervised model (Table 5-2). This result suggested that the performance improvement for the semi-supervised model did not depend on the particular 7 unlabeled samples in our original data set. The semi-supervised model is expected to outperform the supervised model even with another set of unlabeled samples. Although the semi-supervised model with the simulated unlabeled samples performed better than the supervised model, it did not perform as good as the semi-supervised model with the 7 actual unlabeled samples (Table 5-4). This was very reasonable, because the simulated samples were derived from the distribution of the labeled samples, and they did not contribute as much information as the 7 actual samples about the actual sample distribution.
Table 5-6. LOOCV performance of the semi-supervised model with simulated unlabeled samples.

<table>
<thead>
<tr>
<th>α</th>
<th>AUC</th>
<th>Pcorr</th>
<th>Scorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.83±0.02</td>
<td>0.49±0.01</td>
<td>0.53±0.02</td>
</tr>
<tr>
<td>0.2</td>
<td>0.82±0.02</td>
<td>0.48±0.02</td>
<td>0.52±0.03</td>
</tr>
<tr>
<td>0.3</td>
<td>0.78±0.02</td>
<td>0.45±0.02</td>
<td>0.48±0.03</td>
</tr>
<tr>
<td>0.4</td>
<td>0.76±0.01</td>
<td>0.44±0.01</td>
<td>0.45±0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>0.76±0.01</td>
<td>0.43±0.03</td>
<td>0.45±0.02</td>
</tr>
</tbody>
</table>

5.3.8 Model with feature selection

In RFE, the percentage of features to be removed in each iteration was set to be 1%, and we required the algorithm to remove one feature at a time if the total number of features was less than 100. For the number of features to be kept in the model, there was not an effective way to determine the threshold before the experiment. Therefore, we tried different thresholds, and performed a LOOCV at each threshold. The classifier performance is shown in Figure 5-7. We plotted the AUC, Pcorr and Scorr under different thresholds. We also plotted the 95% confidence intervals for the AUCs, which were calculated with the pROC package (Robin et al., 2011). When the number of selected features was set to be 1, the classification problem was not solvable with the default slack variables defined in the SVM\(^{\text{light}}\) package. Thus, the SVM\(^{\text{light}}\) package attempted to relax the slack variables gradually. However, the training did not converge after 96h of computation. From Figure 5-7, we can see that the confidence intervals of the AUCs overlapped, with the number of selected features ranging from 2 to 650, indicating that the performances, despite random fluctuations, did not change significantly over this interval. In other words, two features would be enough to separate the effective group from the ineffective group. Hence, we set the number of selected features at 2, and performed a LOOCV with RFE for feature selection. We achieved classification accuracy of 93.8%, AUC of 0.92, Pcorr of 0.78 and Scorr of 0.72. This excellent performance confirms that two features are sufficient to
separate the effective group from the ineffective group using the semi-supervised SVM model based on features extracted from the fMRI contrast of speech vs. silence.

We also analyzed the performance of the supervised model with RFE for feature selection with the selected number of features set at 2. With the LOOCV approach, we achieved accuracy of 68.8%, AUC of 0.71, Pcorr of 0.56 and Scorr of 0.45, which was much lower than that of the semi-supervised model.

![Figure 5-7. LOOCV performance for the semi-supervised model using RFE for feature selection with the number of selected features changing from 2 to 650.](image)

5.3.9 Validate the feature selection algorithm

We hypothesized that the two features selected by the RFE algorithm have strong discriminative power in distinguishing effective from ineffective CI users. These selected features may serve as biomarkers for predicting the CI outcomes. Ideally we would like to predict language performance outcome based on the fMRI contrast of speech vs. silence. Our features came from the brain regions that responded to the speech stimulus. There might be a chance that all or most of the features were good predictors. Any random selection of two features may yield a good model. If this is true, the features selected by the RFE algorithm will be unreliable and cannot be used as biomarkers. To exclude this possibility, we randomly selected two features from the 658 features to train a semi-supervised model, and used the LOOCV approach to evaluate the model.
We repeated this experiment 100 times and observed a considerable number of runs in which the classification problem was not solvable unless the slack variables were relaxed, indicating that the effective group and ineffective group were not separable based on the two randomly selected features. Furthermore, none of the 100 runs achieved an AUC higher than 0.92, which was the performance when we employed the RFE algorithm for automatic feature selection. Only 3 out of the 100 runs achieved an AUC above 0.8. This result indicated that not all features or combinations of features were equally good predictors for the classification of effective vs. ineffective CI users. We concluded that the RFE algorithm correctly selected the features with the strongest discriminative power.

5.3.10 Stability of feature selection

Because the features were selected automatically for each fold of cross-validation, they might vary across different folds. We had 16 folds of cross-validation in total, and selected 2 features at each fold, which led to 32 features in total. We calculated the occurrence frequency for each feature that had occurred for at least once. Besides, there could be multiple features that came from the same brain region and were highly correlated with each other. Such redundancy was attributed to the inherent properties of our feature extraction approach. For example, some characteristic contrast regions might occur in more than one subject as described in section 4.3.3. In such cases, there were multiple copies of this region in the vocabulary. The features corresponding to such regions were highly correlated. Such features were considered to be redundant and selected interchangeably during the feature selection process. The 64 selected features across different folds of cross-validation, i.e., 32 from the semi-supervised model and 32 from the supervised model, involved 13 different features. Among the 13 features, a pair of features had a correlation as high as 0.99, and we had verified that this pair of features came
from approximately the same brain region. Therefore, we treated this pair of features as one single feature when we calculated the occurrence frequency. Except for this pair of features, the correlations between other pairs of selected features were mostly below 0.8 with a few pairs above 0.8 but below 0.9. Thus, all of the other selected features were treated as different features. As shown in Figure 5-8, the feature selection was clearly more stable for the semi-supervised model when compared to the supervised model. Due to the limited number of samples in the supervised learning, the selected features in each fold of cross-validation might fit the training samples too well to be generalizable to the testing samples. This was a possible explanation for the unstable feature selection and unsatisfactory LOOCV performance for the supervised model.

Comparing Figure 5-8B with Figure 5-8A, the most frequently selected feature in Figure 5-8B and that from Figure 5-8A were the same feature. However, the second most frequently selected features did not match, and the correlation between these two features was only 0.46. As we mentioned above, features were eliminated one by one at the late stage of RFE. We trained a semi-supervised model with all the 23 samples, using RFE for feature selection. Let the last 10 features be labeled as A to J. If we reduced the number of selected features from 10 to 2, the first feature to be eliminated was the feature J, followed by feature I and H, and so on. The last two features kept in the model were feature A and B. We found that the two most frequently selected features from Figure 5-8B corresponded to the features A and B, and the second most frequently selected feature in Figure 5-8A corresponded to the feature D. Based on these observations, we concluded that the selected features from the semi-supervised model were consistent with the selected features from the supervised model. The semi-supervised model did not select certain features to fit the unlabeled samples. Therefore, we excluded the possibility that some unlabeled samples might be outliers, which forced the feature selection to be fixed on certain features and
consequently improved the stability of feature selection. Instead, the improved stability should be attributed to the improved statistical power of the training set due to the inclusion of unlabeled samples.

![Figure 5-8](image)

**Figure 5-8.** Histogram of selected features across different folds of cross-validation. The horizontal axis is the feature index, and the vertical axis is the occurrence frequency of a feature. (A) Supervised model. (B) Semi-supervised model.

### 5.3.11 Discriminative brain regions

We trained a semi-supervised model with all of the 23 subjects and used the RFE for feature selection with the number of selected features set to be 2. Then, we back-projected the two
discriminative brain regions onto the infant template (Altaye et al., 2008) as shown in Figure 5-9. As we mentioned above, these two features are actually the features 1 and 2 in Figure 5-8B. The first predictive feature corresponds to a brain region located in the left superior and middle temporal gyri and aligns with our original hypothesis that brain activity in this area might be predictive of outcomes following cochlear implantation in infants. However, using univariate correlation or regression analysis with age at implantation and pre-implant hearing threshold as covariates, we have not been able to find persuasive predictive value by looking at a region of interest in this part of the brain alone. This implies that one single brain region is not enough for the classification. Only the combination of multiple brain regions makes a good prediction for the language function for infants receiving a CI. To verify this observation, we showed the distribution of samples in Figure 5-10 using the brain activities of the two discriminative brain regions. As we can see, the CI-effective and CI-ineffective users were not separable using either a horizontal line or a vertical line, which confirmed that only one brain region was not enough for the classification. Combining these two brain regions, the two groups of individuals were separable by a diagonal line. The other important region is located in the right cerebellum, whose predictive power was underestimated according to our original hypothesis. However, based on a substantial number of scientific publications supporting the role of cerebellum in supporting language functions (see section 5.4 Discussion), this discovery is not so surprising as it is unconventional. This finding demonstrates the advantage of machine learning techniques, which can automatically detect the predictive features and draw our attention to features that are important but beyond our prior knowledge.
Figure 5-9. Two discriminative brain regions from the contrast speech vs. silence. Images are displayed in neurological orientation using the xjView toolbox (xjView). The coordinate of the center of the first region is (-54, -70, 13). This region is located in an area corresponding to left superior and middle temporal gyri. The second region is located in the right cerebellum. Its central coordinate is (10, -86, -27).
Figure 5-10. Distribution of samples within the two-dimensional space. Horizontal axis represents the brain activity in left superior and middle temporal gyri (first brain region in Figure 5-9), and the vertical axis represents the brain activity in right cerebellum (second brain region in Figure 5-9).

5.4 Discussion

In this chapter, we presented a semi-supervised SVM model for predicting whether or not a prelingually deaf in infant or toddler receiving a CI will develop effective language skills within two years after the surgery. Such prognostic information could be extremely useful and is currently not available to clinicians by any other means. The average cost of cochlear implantation in the United States is $60,000 including device, surgery and post-implantation therapy fees (Battey, 2007). A reliable predictive tool can guide pre-operative counseling, help to
calibrate expectations, influence post-CI speech and language therapy and prevent subsequent disappointment. Non-invasive neurobiological information about developing auditory and language networks in the brain available via fMRI and accurate interpretation of such data using the approach we have developed may also guide timely intervention with CI, and consequently maximize the benefits of CI. The findings of the present study demonstrate the remarkable power of a semi-supervised machine learning approach to the analysis of group fMRI data. Where simple linear regression models between fMRI statistical parameter maps and hearing outcome measures have failed to provide a method by which fMRI data from individual subjects can be used to make a prognosis about possible speech and language outcomes, two features extracted by a machine learning algorithm appear to be able to provide us with a method of doing so. Even with a limited sample size, we have demonstrated a classification accuracy of 93.8% based on two features from brain activation maps of infants listening to natural language during an fMRI scan. Clearly the methodology needs further exploration and verification with a larger sample size. As the sample size increases, the classification problem becomes more complicated, and it may require additional features to make an accurate classification. But at this stage machine learning classification of fMRI data appears to offer promise in producing an objective prognosis for speech and language outcomes in individual prelingual CI recipients. For the two brain regions highlighted by our current algorithm, we discuss below their functions, and the probable biological mechanisms underpinning their involvement in the classification of effective and ineffective CI users.

**Left Superior and Middle Temporal Gyri** Wernicke first described the role of the superior temporal gyrus (STG) in speech perception (Wernicke, 1874). Auditorily presented speech stimuli are known to activate the STG (Mummery et al., 1999, Patel et al., 2007, Vannest et al.,
while lexical/semantic processing requires inputs from the middle temporal gyrus (MTG) (Vandenberghe et al., 1996, Holland et al., 2007). The left STG has been shown to analyze phoneme and word forms (DeWitt and Rauschecker, 2013). Jamison et al. demonstrated, using an fMRI paradigm, that the right STG is specialized for detecting spectral changes even for non-speech stimuli (Jamison et al., 2006) whereas the left STG is more highly attuned to temporal variations typical of auditory speech inputs. Recently, Duffy et al. reviewed the utility of the Frequency Modulated Auditory Evoked Response (FMAER) technique in the diagnosis of childhood language impairment (Duffy et al., 2013). The FMAER is based on the detection of quick frequency modulation (FM) changes in speech and its source can be traced to posterior STG bilaterally. The authors found that the FMAER was absent in children with language impairments, especially those with speech comprehension deficits. Children with hearing impairment are known to have language delays and/or deficits (Ching et al., 2010, Yoshinaga-Itano, 2014). Thus, it is reasonable that our SVM models, both the semi-supervised model and the supervised model, identified the STG as one of the areas differentiating between effective and non-effective CI users based on the variable language proficiency of children with congenital hearing impairment. Peterson et al., using positron emission tomography (PET), found that the left STG was activated in a speech comprehension task in post-lingually deaf adults but not in pre-lingually deaf adults (Petersen et al., 2013). They attributed this finding to the exposure to language in the post-lingual HI group. In the current study, it is possible that the left STG was identified in the classification of effective vs. non-effective CI users as this brain area in the former group may be better tuned to analyzing incoming speech stimuli. At the same time, it is important to note the differences between the two studies before such direct comparisons can be made; e.g., in terms of populations studied (adult CI recipients vs.
infant/toddler CI candidates), stimulus presentation (monaural vs. binaural) and subject state during acquisition (awake vs. sedated).

In an early landmark PET study aimed at parcellating the brain areas responsible for different aspects of sentence comprehension, Mazoyer et al. found that the left STG and MTG were activated in response to stories in the native tongue (Mazoyer et al., 1993). By comparing different contrasts, they concluded that “the activations in the left middle temporal gyrus… reflect processing beyond the single-word level” (p. 469), i.e., syntactic processing in addition to phonological and lexical analysis. Using PET, Kang et al. studied the fluorodeoxyglucose (FDG) uptake in the brains of 87 children with congenital hearing impairment ranging from 1-15 years of age (Kang et al., 2004). Giraud and Lee re-analyzed their data and found an age-dependent increase in metabolism in both the superior and middle temporal gyri on the left side (Giraud and Lee, 2007). Following a subset of the same participants as they became CI recipients, the authors found the left prefrontal and parietal areas correlated positively with speech perception scores irrespective of the age at implantation or duration of deafness. The authors suggested that these age-dependent and independent hypermetabolic changes may indicate a 2-step cortical reorganization process in children with congenital HI. In the present study, a similar cortical reorganization may have played a role in the effective group resulting in greater language gains post-implantation. In another study of neural activation using PET, Giraud et al. found the STG and MTG to be highly active in CI users in response to unrelated sentences (Giraud et al., 2000). In addition, the anterior portion of the left MTG was also activated in response to a story-listening task. This observation indicates that the left STG and MTG continue to play a role in speech processing/language acquisition after cochlear implantation. Our results suggest that pre-CI activation of left STG in response to a natural speech stimulus is one of two
key features in the fMRI results that are predictive of later language outcomes for prelingually deaf children receiving a CI.

**Right Cerebellum** For over 200 years, the cerebellum was primarily considered to be the center of motor control (Rolando, 1809, Holmes, 1939). However, this view has been challenged in the past two decades (Habas, 2001). In one of the early neuroimaging studies supporting cerebellum as a sensory center, Gao et al. demonstrated dentate nuclei to be active in response to cutaneous stimulation (Gao et al., 1996). Later, overt and covert speech production abilities were associated with cerebellar involvement in both normal (Riecker et al., 2000, Seger et al., 2000, Frings et al., 2006) and disordered (Eckert et al., 2003) populations. However, studies have also indicated cerebellar recruitment in speech reception tasks. Paphathanassiou et al., in a PET study, found right cerebellar activation in normal hearing adults in response to a story listening task – similar to the one used in the present study (Paphathanassiou et al., 2000). Redcay et al., using event-related fMRI, observed right cerebellar activation in normal hearing toddlers in response to natural speech as compared to silence, a similar contrast to the speech vs. silence contrast used in the present study (Redcay et al., 2008). Ackermann et al. have also shown cerebellar involvement in the discrimination of vowel duration and voice onset time – important aspects of speech perception (Ackermann et al., 1997). Strelnikov et al. studied regional cerebral blood flow (rCBF) changes in response to prosodic cues in normal hearing adults (Strelnikov et al., 2006). They found right cerebellar activation when participants listened to sentences with intonation patterns. They concluded that the right cerebellum (in addition to the right dorsolateral prefrontal cortex) plays an important role in extracting syntactic and prosodic information (such as pauses and associated pitch changes) from natural sentences.
Fabbro et al. observed morpho-syntactic and speech comprehension deficits in patients with focal lesions involving the right cerebellum and vermis (Fabbro et al., 2000). In the current study, the right vermis was identified as one of the regions in the cerebellum to be a biomarker of the effective vs. ineffective classification of CI users. The vermis has long been considered as the limbic cerebellum (Anand et al., 1959) in that it has cerebro-cerebellar projections to the cerebral limbic system and it mediates some emotional responses (Timmann and Daum, 2007). It is possible that the vermis is tuned to tapping the emotional content in speech stimuli, even under sedation. Alternatively, the right cerebellum may “reflect some basic low level aspect of neural processing that may be relevant to speech but cannot be a consequence of accessing the speech system itself” (p. 1761) (Johnsrude et al., 1997).

In a review of cerebellar functions, Marien et al. observed that “the cerebellum modulates cognitive functioning of at least those parts of the brain to which it is reciprocally connected” and is involved in “various non-motor language processes such as lexical retrieval, syntax, and language dynamics” (p. 580) (Marien et al., 2001). This view was recently corroborated by Murdoch based on additional evidence using neuroimaging studies (Murdoch, 2010). In the review, Murdoch emphasized the supportive function of the cerebellum in language tasks as opposed to direct involvement – which, in part, may explain the lack of explicit language dysfunction as a result of direct injury to the cerebellum.

**Cerebro-Cerbellar Interaction** In the present study, two regions – right cerebellar vermis and left cerebrum (temporo-parietal) – were identified to successfully discriminate between effective and ineffective CI users. This observation may be explained by the right cerebellar-left cerebral pairing observed in neuroimaging and electrophysiological studies of language lateralization (Desmond et al., 1998, Gronholm et al., 2005). Papanthasissiou et al. found a
coupling of the traditional left cerebral language areas with right cerebellar regions for both speech comprehension and production (Papathanassiou et al., 2000). Strelnikov et al. found deactivations in the superior temporal gyrus and the right cerebellum in response to degraded speech in normal hearing listeners (Strelnikov et al., 2011). Wong et al. also identified similar areas but with activation instead of deactivation, likely the effect of differences in intelligibility of speech stimuli used in the two studies (Wong et al., 2008). Although noise was not explicitly added to speech stimuli in the current study, the effects of acoustic MRI scanner noise interleaved with the story segments certainly present a noisy background to the subject who also has a poorly performing auditory system due to congenital deafness.

In the present study, we compared the predictive power of three different feature sets, namely VOX, BoW21 and BoW44. The BoW features exhibited much better performance than the VOX features, which was expected as we analyzed in the previous two chapters. The classification performance for BoW21 was better than or at least as good as BoW44. BoW21 feature set only included the regions active in the NH infants, while BoW44 included the active regions from both NH controls and SNHL patients. It is likely that regions detected in the 21 NH participants included all the relevant brain regions for classification of effective vs. ineffective CI users. Regions from the SNHL patients did not add any information for the classification process and simply introduced noise into the classification due to aberrant and inconsistent activation patterns. Also notice that the age and gender were not perfectly matched between the SNHL group and the NH group. However, we did not think that this age/gender difference would weaken our analysis or jeopardize our conclusion as we were classifying effective vs. ineffective CI-users rather than NH vs. SNHL. The average age of NH children was 8 months younger than the average age of SNHL children. Although such an age difference during infancy is likely to
significantly affect auditory activation patterns, the feature set (BoW21) extracted based on the activation pattern of the NH children was shown to perform well in the classification of effective vs. ineffective CI-users, even in an older group of infants and toddlers. Admittedly, it would be ideal to use a better age/gender-matched group of NH children, which may help to further improve the classification accuracy. Since our approach works on our current dataset, which represents the worst situation, it is not unreasonable to expect it will work on an ideal dataset.

Correlations (Pcorr and Scorr) for the contrast speech vs. silence were considerably higher than the other two contrasts. The human auditory system is highly attuned to human speech sounds and recent theories suggest that the right and left hemisphere structures are specifically tuned to temporal and spectral features of the speech waveform respectively (Hickok and Poeppel, 2007, Ghitza et al., 2012, Poeppel et al., 2012, Wang et al., 2012c). Human speech therefore activates a more extensive auditory network than any other form of auditory stimulation. When speech is contrasted with silence we can expect the maximum difference in brain activity to occur within this auditory language network. Therefore, it is not surprising that we find the greatest correlations between the features from the fMRI contrast for speech vs. silence with the CELF-P2 scores as shown in Table 5-2 and Table 5-3. Somewhat more surprising is that the auditory system of an infant with severe to profound SNHL still responds maximally to this type of stimulus during fMRI with auditory stimulation prior to cochlear implantation, while he/she is still deaf. The fact that this contrast produces the best classification performance for our chosen model suggests that infants who have the greatest response to human speech (more specifically the voice of the mother of infants and toddlers themselves) are most likely to develop good language capabilities with cochlear implantation.
In this work, we also compared the supervised model with semi-supervised model. A straightforward explanation was included to illustrate the superiority of the semi-supervised model, where the unlabeled samples helped to characterize the distribution of samples in the feature space, and therefore the classifier was able to find a more optimal hyperplane to separate the different groups of samples. Furthermore, adding the unlabeled data to the analysis increased the statistical power of training data due to a larger sample size, which helped to highlight the most discriminative features during feature selection. Although the semi-supervised model showed a remarkable power in the current project, we are not saying a semi-supervised model is always better than a supervised model. As shown in the section 5.3.6 **Semi-supervised model by adding the unlabeled samples one by one**, adding either unlabeled positive samples alone or unlabeled negative samples alone produced a classification performance the same as or very close to that of the supervised model. The superiority of a semi-supervised model over supervised model depends on how much the unlabeled samples contribute to the characterization of the distribution of different groups of samples. Based on our experience, a semi-supervised model will perform at least as well as a supervised model, and therefore it is always worthwhile to try a semi-supervised model when unlabeled data are available. Please note this is only an empirical observation, a mathematical proof is needed to support this conclusion. Besides, we also compared the semi-supervised SVM model with two other related models, namely transductive SVM (TSVM) model and standard logistic regression model. The TSVM model was trained using SVM\textsuperscript{light} package with the parameter $r$ set to be 0.6, which was the ratio of positive samples among the labeled samples (9 positive samples vs. 7 negative samples). The logistic regression model was trained using the “glmfit” function in MATLAB. The two models were evaluated using the LOOCV approach as well. Results were summarized in **Table 5-7**. We
can see that the performance of the TSVM model was worse than that of the semi-supervised SVM model (Table 5-3) for most of the feature sets. This was expected for reasons as explained below. In our project, we had 16 labeled samples and 7 unlabeled samples. Let the number of positive samples within the 7 unlabeled samples be N. During the LOOCV process, the TSVM model added the left-out sample to the training set as an unlabeled sample. If the left-out sample was positive, the parameter $r$ (which was the ratio of positive samples within the unlabeled samples) would be $(N+1)/8$. If the left-out sample was negative, the parameter $r$ would be $N/8$. Thus, the parameter $r$ kept changing, and there was not a good way to optimize the parameter $r$.

For the semi-supervised model, however, the parameter $r$ was a constant $(N/7)$, and it can be optimized using a nested LOOCV as illustrated in Figure 5-2. Due to the non-optimal parameter $r$, the performance of TSVM model was adversely affected. Furthermore, the standard logistic regression model demonstrated almost random classification for most of the feature sets. This was because the number of samples was much smaller than the number of features for our dataset, in which case the logistic regression model without regularization was not estimable. As a result, the logistic regression model was unreliable and performed poorly for the LOOCV.
Table 5-7. LOOCV performance for the transductive SVM (TSVM) model and logistic regression (LR) model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Type</th>
<th>Contrast</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Pcorr</th>
<th>Scorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSVM</td>
<td>BoW44</td>
<td>Speech vs. silence</td>
<td>88.9%</td>
<td>28.6%</td>
<td>62.5%</td>
<td>0.80</td>
<td>0.36</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noise vs. Silence</td>
<td>77.8%</td>
<td>57.1%</td>
<td>68.8%</td>
<td>0.66</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech vs. Noise</td>
<td>33.3%</td>
<td>71.4%</td>
<td>50.0%</td>
<td>0.55</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combine</td>
<td>77.8%</td>
<td>71.4%</td>
<td>75.0%</td>
<td>0.76</td>
<td>0.40</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>BoW21</td>
<td>Speech vs. silence</td>
<td>77.8%</td>
<td>42.9%</td>
<td>62.5%</td>
<td>0.67</td>
<td>0.37</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noise vs. Silence</td>
<td>77.8%</td>
<td>57.1%</td>
<td>68.8%</td>
<td>0.74</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech vs. Noise</td>
<td>44.4%</td>
<td>71.4%</td>
<td>56.3%</td>
<td>0.52</td>
<td>0.00</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combine</td>
<td>88.9%</td>
<td>57.1%</td>
<td>75.0%</td>
<td>0.66</td>
<td>0.37</td>
<td>-0.02</td>
</tr>
<tr>
<td>LR</td>
<td>BoW44</td>
<td>Speech vs. silence</td>
<td>44.4%</td>
<td>71.4%</td>
<td>56.3%</td>
<td>0.51</td>
<td>0.08</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noise vs. Silence</td>
<td>77.8%</td>
<td>42.9%</td>
<td>62.5%</td>
<td>0.62</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech vs. Noise</td>
<td>77.8%</td>
<td>42.9%</td>
<td>62.5%</td>
<td>0.62</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combine</td>
<td>33.3%</td>
<td>71.4%</td>
<td>50.0%</td>
<td>0.56</td>
<td>-0.07</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>BoW21</td>
<td>Speech vs. silence</td>
<td>22.2%</td>
<td>71.4%</td>
<td>43.8%</td>
<td>0.43</td>
<td>-0.09</td>
<td>-0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noise vs. Silence</td>
<td>33.3%</td>
<td>71.4%</td>
<td>50.0%</td>
<td>0.51</td>
<td>-0.02</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech vs. Noise</td>
<td>55.6%</td>
<td>42.9%</td>
<td>50.0%</td>
<td>0.54</td>
<td>-0.17</td>
<td>-0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combine</td>
<td>44.4%</td>
<td>57.1%</td>
<td>50.0%</td>
<td>0.60</td>
<td>-0.07</td>
<td>0.05</td>
</tr>
</tbody>
</table>

In order to develop a prognostic tool that will eventually be useful clinically, several improvements to the model would be needed as future efforts. First, a larger sample size is needed to train an initial regression model instead of a classification model. We will continue to use the semi-supervised learning for the regression model, considering the difficulties in recruiting participants and there will still be unlabeled samples whose follow-up scores are not obtainable among the newly recruited infants. An additional benefit of a larger training dataset would be the possibility of constructing a more complete assessment of patient outcome based on additional measures of speech, language and cognitive ability for each participant. In the current analysis we have classified patients as effective and ineffective CI users based exclusively on the CELF-P2 scores. While this test provides a comprehensive assessment of language fundamentals in pre-school-aged children, it does not include measures of speech ability or general cognitive function. Clearly the effectiveness of CI usage, even during the early developmental stages, should be based on a broader range of cognitive abilities than language alone, even though a high degree of correlation might be expected among such assessments. Availability of comprehensive
neurocognitive data for a larger patient population would allow for a more accurate evaluation of the effectiveness of CI usage, which in turn should improve training of the model and result in a more accurate predictive model.

Another potential research direction might involve improving the feature extraction algorithm. For the current BoW algorithm, we defined the thresholds based on the p-values for the T-statistics. Optimization of these thresholds might maximize the brain regions while preserving the homogeneity of the contrast intensities within a single region. Finally, we focused on fMRI image features alone in the present study. In future work, the age at implantation and the pre-implant hearing threshold should be included as two additional features in a clinically relevant predictive model, since these variables are known to account for much of the variance in CI outcomes.

5.5 Conclusion

In this study, we have confirmed that our BoW approach is more accurate than the conventional approach for feature extraction to enhance performance of a machine learning approach to making predictions about future clinical outcomes based on fMRI data alone. Not surprisingly, fMRI measures of brain activity stimulated by human speech provided contrasts that were most predictive of language outcomes after cochlear implantation. Semi-supervised learning made the maximal use of the available data, and provided a stable and accurate classification model for predicting the CI outcomes. Capitalizing on the excellent performance of the semi-supervised model, we have validated the hypothesis that pre-implant cortical activation patterns revealed by fMRI during infancy correlate with language performance two years after cochlear implantation. By using the recursive feature elimination algorithm for feature selection, we discovered that two features from the fMRI contrast map for speech vs. silence were sufficient for classifying
effective from ineffective CI users based on our current dataset. We highlighted these two features as discriminative brain regions. One of these two regions is located in an area corresponding to left superior and middle temporal gyri. The left STG is implicated in spectral, phonemic and lexical processing of human speech. The left MTG is involved in syntactic processing. These observations have been made not only in congenitally hearing impaired children but also in young CI recipients. These findings suggest that the left STG and MTG play an important role in speech processing/language acquisition even in congenitally deaf infants and toddlers. In the present study, cortical development in these areas may have played a role in the effective group, resulting in greater language gains post-implantation. The second region is located in the right cerebellum. The involvement of the right cerebellum in the speech vs. silence contrast suggests that it may play a role in extracting syntactic and prosodic information from natural speech and points to the supportive function of the cerebellum in linguistic tasks. Based on this preliminary result we are optimistic that a reliable machine learning model based on a larger training set can eventually be applied in the clinical setting to provide specific prognostic information to patients considering cochlear implantation.
Chapter 6 Functional Volume Outperforms Anatomical Volume for Automatic Diagnosis of Attention Deficit Hyperactivity Disorder

In the previous three chapters, we studied task fMRI data. In this chapter, we focused on resting-state fMRI (rs-fMRI). While the contrast maps generated from general linear model were usually used as brain activity maps for task fMRI, (fractional) amplitude of low frequency fluctuations (ALFF/fALFF) maps could be considered as brain activity maps for rs-fMRI. In this chapter, we proposed a new approach for feature extraction from fALFF maps and applied this approach to the classification of Attention Deficit Hyperactivity Disorder (ADHD) patients vs. typically developing controls (TDC), using the ADHD-200 dataset.

Since anatomical brain volume shrinkage was one of the most replicated findings for ADHD, we calculated regional brain volumes as predictors for the classification of ADHD patients vs. TDC. Conventionally, volume of a brain region was considered to be an anatomical feature and quantified using structural magnetic resonance images. In the present study, however, we have initially proposed to measure the volumes of brain regions using fALFF maps. Brain volumes measured from fALFF maps were denoted as functional volumes, which quantified the volumes of brain tissue that were actually active during fMRI imaging. We compared the predictive power of functional volumes with that of the regional brain volumes measured from anatomical images, which was denoted as anatomical volumes. The former demonstrated higher discriminative power than the latter for the classification of ADHD patients vs. TDC. Combining with our two-step feature selection algorithm, our SVM classification model using functional volumes achieved a balanced accuracy of 67.7%, which was 16.1% higher than that of a relevant
model published previously (Sato et al., 2012b). Furthermore, our classifier highlighted 10 brain regions that were most discriminative in distinguishing between ADHD patients and TDC. These 10 regions were mainly located in cerebellum posterior lobe, occipital lobe, temporal lobe, frontal lobe and parietal lobe. Potentially, our present study using functional images would provide some new perspectives about the brain regions affected by ADHD.

6.1 Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is a psychiatric disorder characterized by clinical symptoms of inattention, impulsivity, and hyperactivity. This condition affects 5-8% school age children, and usually persists into adolescence and adulthood. Clinical diagnosis of ADHD is based on behavioral information gathered from parents and school. Depending on the number and type of symptoms, a child can be diagnosed with one of three ADHD presentations: primarily inattentive (ADHD-I), primarily hyperactive (ADHD-H) or combined subtype (ADHD-C) (Association, 2013). Despite its high prevalence, the precise etiology and pathogenesis of ADHD remains unclear.

In recent years, magnetic resonance imaging (MRI) studies of patients with ADHD have demonstrated possible physiological underpinnings of the disorder. Modern machine learning techniques with a large-scale dataset may help to identify reliable neuroimaging biomarkers, which may offer some clues of the physiological basis of the disorder. Towards this aim, the ADHD-200 Consortium organized the ADHD-200 global competition, making hundreds of anatomical and functional images publicly available (Consortium, 2012). The ADHD-200 competition released the demographic and clinical data, anatomical and resting-state functional MR images for 973 participants accumulated from 8 independent sites: Bradley Hospital/Brown University, Kennedy Krieger Institute, NeuroIMAGE Sample, New York University Child Study
Spontaneous low frequency fluctuations in blood oxygen level dependent (BOLD) activity are a fundamental feature of the brain at rest. The relative magnitude of these fluctuations is usually measured by the amplitude of low frequency fluctuations (ALFF) (Zang et al., 2007) or fractional amplitude of low frequency fluctuations (fALFF) (Zou et al., 2008). ALFF is a voxel-by-voxel calculation of the power spectrum of the BOLD fMRI time series. fALFF is the ratio of power spectrum of low-frequency (0.009-0.08 Hz) to that of the entire frequency range. ALFF/fALFF was widely used to study the abnormal spontaneous brain activities in various diseases, such as schizophrenia (Hoptman et al., 2010), amnestic mild cognitive impairment (Han et al., 2011, Han et al., 2012), Parkinson’s disease (Skidmore et al., 2013), major depressive disorder (Jiao et al., 2011, Wang et al., 2012a), etc. Studies in ADHD also reported decreased/increased ALFF/fALFF in various brain regions (Zang et al., 2007, Yang et al., 2011, An et al., 2013). fALFF has been evaluated in the ADHD-200 competition dataset. In particular, Cheng et al. used the fALFF coefficient at each voxel as an indicator of ADHD status and then applied two-sample t-test to select significant voxels for subsequent model training (Cheng et al., 2012). Combining fALFF with regional homogeneity (ReHo) and information from brain networks, they achieved a cross-validated accuracy of 76.15% for the classification of ADHD patients vs. healthy controls on the dataset collected from Peking University. Sato et al. (Sato et al., 2012b) applied the brain parcellation defined by CC400 atlas, which divided the brain into 351 regions. The mean fALFF within each brain region was calculated and used as a predictor in their classification model. Based on the whole ADHD-200 dataset, Sato et al. suggested that the combination of ALFF and ReHo contained information to distinguish ADHD patients from
healthy controls, but with limited discriminative power. CC400 atlas was generated via a two-level spatially constrained spectral clustering algorithm (Craddock et al., 2012) using the ADHD-200 dataset, and was made publicly available by NeuroBureau at the competition website. CC400 atlas was widely used in studies on the ADHD-200 dataset (Colby et al., 2012, Dai et al., 2012a, Sato et al., 2012a, Sato et al., 2012b, Sato et al., 2013).

Through a close examination, we observed that there were many voxels whose fALFF coefficients were zeros. According to the definition of fALFF (Zou et al., 2008), a voxel exhibited zero fALFF coefficient only when there was no fMRI signal at this voxel. In other words, the voxels with zero fALFF coefficients were inactive during fMRI imaging. In this case, any voxel-level measures, e.g. ReHo and fALFF, would have a magnitude of zero. Thus, a zero fALFF coefficient may probably be attributed to the absence of brain tissue or inactive necrotic tissue at this voxel. On the other hand, CC400 atlas was supposed to cover only voxels within the brain. All the voxels covered by CC400 atlas should not have zero fALFF coefficients, which contradicted our actual observation. Furthermore, we observed individual differences for the number of voxels with zero fALFF coefficients, although the images from different subjects had been normalized to the standard template and were assumed to be perfectly overlapped. The individual differences were illustrated in Figure 6-1. Please note we looked at voxels with non-zero instead of zero fALFF coefficients in Figure 6-1 for visualization considerations. A fALFF coefficient was considered to be non-zero as long as it was not exactly equal to zero. Therefore, a voxel was either a zero voxel or non-zero voxel, and individual difference analysis on the non-zero voxels was equivalent to the individual difference analysis on zero voxels. Our above observation of the voxels with zero fALFF coefficients within the brain as well as their individual differences seemed to be ignored by previous studies, which calculated an overall
mean fALFF within a brain region without giving particular attention to the voxels with zero fALFF coefficients (Sato et al., 2012b). In our study, we proposed to use the count of non-zero voxels as a measure of brain volume that was actually active during fMRI imaging. We named this brain volume as **functional volume**, which was the counterpart of **anatomical volume**.

Conventionally, volume of a brain region was considered to be an anatomical feature and usually calculated based on anatomical magnetic resonance images. To the best of our knowledge, functional volume was initially proposed by us in the current study. We first compared the functional brain size between ADHD patients and healthy controls, since studies based on anatomical images reported that the brains of children and adolescents with ADHD were 3-4% smaller than those of children who did not have this disorder (Castellanos et al., 2002). In addition to the total brain size analysis, we also investigated how ADHD affected the functional volume of different brain regions. Finally, we applied the CC400 atlas to calculate the functional volumes for brain regions, and used them as predictors for the classification of ADHD patients vs. TDC, given that decreased anatomical volume was one of the most replicated evidences for ADHD (Castellanos et al., 1996, Berquin et al., 1998, Castellanos et al., 2001, Castellanos, 2002, Mostofsky et al., 2002, Carmona et al., 2005, Mackie et al., 2007, Wang et al., 2007, Carmona et al., 2009, Batty et al., 2010, Montes et al., 2011, Qiu et al., 2011, Lopez-Larson et al., 2012). We compared the predictive power of our functional volumes with other relevant feature sets, including regional mean fALFF (Sato et al., 2012b) and anatomical volumes. We also compared our model using functional volumes with the model based on demographic data, since demographic data were shown to outperform resting state fMRI measures in the ADHD-200 global competition (Brown et al., 2012). Besides, we trained models by integrating the functional/anatomical volumes with the demographic characteristics, given the
possibility that brain volume might be correlated with the personal characteristics and integrating these two types of information might improve the classification performance. Our goal for this study was to verify that if fMRI images could provide some additional information that was not included in the anatomical images about the brain volume abnormality in ADHD patients, and consequently lead to a better classification model for the automatic diagnosis of ADHD.
Figure 6-1. A. We calculated the total number of voxels with non-zero fALFF coefficient for each subject, and summarized the distribution of subjects using a histogram. Horizontal axis is the total number of voxels with non-zero fALFF coefficient within a brain. Vertical axis is the number of subjects. B. We calculated the number of subjects with non-zero fALFF coefficient voxel-by-voxel. The color of a voxel represents the number of subjects with non-zero fALFF coefficient at this voxel. Dark red indicates very few subjects have non-zero fALFF coefficient at a voxel, while white indicates that almost all subjects have non-zero fALFF coefficient at a voxel. Regions with colors from light red to yellow have a high individual difference.
6.2 Materials and Methods

6.2.1 Participants

In order to avoid the systematic differences caused by scanner hardware and scanning protocols across different sites (Stonnington et al., 2008, Moorhead et al., 2009, Huppertz et al., 2010, Abdulkadir et al., 2011, Kostro et al., 2014), we used the data from New York University Child Study Center (NYU) only. This center was selected due to its largest sample size among all sites. Although the Peking dataset had a sample size comparable to NYU, the Peking dataset included 3 batches with different scanning parameters, e.g. different voxel sizes, which may introduce undesirable heterogeneity to the data, and was not used for the present study. The NYU dataset included 263 subjects in total. After excluding the subjects whose image quality was questionable for either fMRI images or anatomical images, there left 215 subjects, which were used for the present study. The quality control assessments (usable vs. questionable) based upon visual inspection were provided by ADHD-200 Consortium. One-hundred and seventeen out of the 215 subjects were ADHD patients (86 male, average age=11.3 years) and the remaining 98 subjects were TDC (50 male, average age=12.4 years).

6.2.2 Preprocessing of Images

The anatomical images were preprocessed using SPM8 with standard procedures. Images were segmented to generate grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) density maps, which were subsequently normalized to the MNI template with modulation.

The preprocessing of the raw fMRI data was carried out by NeuroBureau using the Athena pipeline. The Athena pipeline consisted of the following steps:

(1) Exclude the first 4 echo-planar (EPI) volumes.

(2) Slice timing correction.
(3) Deoblique dataset.

(4) Motion correction by registering the EPI volumes to the first volume.

(5) Spatial normalization.

(6) Extract the white matter (WM) and cerebrospinal-fluid (CSF) time-courses.

(7) Regress out WM, CSF, motion time courses from EPI data.

(8) Temporal band-pass filter \((0.009<f<0.08 \text{ Hz})\).

(9) Spatially smooth the filtered data using a Gaussian filter (full width at half maximum = 6mm).

After preprocessing, NeuroBureau also generated the fALFF maps and made them publicly available at the competition website. For the present study, we used the fALFF maps and the normalized tissue density maps as the inputs for the machine learning analysis.

### 6.2.3 Features

As in (Sato et al., 2012b), we employed the CC400 atlas to segment the brain into regions. CC400 was a data-driven atlas that clustered the voxels into functionally coherent and spatially continuous regions based on the fMRI time courses of the voxels. CC400 atlas was generated via a two-level normalized cut (Ncut) spectral clustering algorithm. On the first level, subject specific parcellation was generated within each subject. The similarity between two voxels was calculated as the Pearson’s correlation coefficient between the fMRI time courses of the two voxels. On the second level, it was the group level clustering. Based on the clustering results from the first level, an \(N\times N\) adjacency matrix was constructed for each subject, with \(N\) being the number of voxels within the brain. The adjacency matrices were averaged across subjects to form a group coincidence matrix that was fed to the Ncut algorithm to generate group-level parcellation. Since all the images were normalized to the standard MNI space, voxels from
CC400 atlas and voxels from fALFF maps or tissue density maps were assumed to be aligned. The parcellation labels from CC400 atlas can be directly transferred to the fALFF maps or the tissue density maps. Each region defined in the atlas corresponded to a feature, e.g. a fALFF map can be transformed into a 351-D feature vector using the CC400 atlas. The approaches for calculating the feature values were described below.

Functional Volume (FV) For each brain region defined in the CC400 atlas, we calculated the percentage of voxels with non-zero fALFF coefficients to quantify the functional volume for this region. A voxel was considered to be a non-zero voxel as long as its fALFF coefficient was greater than zero. The threshold zero instead of other small numbers, e.g. 10^{-3}, was determined based on the distribution of fALFF coefficients of the voxels. Figure 6-2 shows the empirical cumulative distribution of the fALFF coefficients for 9 subjects, which were randomly selected from the 215 participants. From Figure 6-2, we can see that fALFF coefficients were either exactly equal to zero or generally above 0.4. Therefore, any thresholds in the range [0, 0.4] would have the same results. Although the functional regional volumes were calculated based on fALFF maps, they actually characterized the volumes of the brain regions that were active during fMRI imaging, rather than fALFF itself. We would obtain the same set of numbers from ReHo maps, because a voxel with zero fALFF coefficient will also have zero ReHo coefficient. Zero voxels were aligned between ReHo maps and fALFF maps.

Regional Mean fALFF We applied two different approaches to calculate the regional mean fALFF. First, we calculated regional mean fALFF using the traditional method by including voxels with zero fALFF coefficients (Sato et al., 2012b), and denoted it as fALFF1. We noticed that fALFF1 was highly correlated with functional regional volume. For example, a brain region including many voxels with zero fALFF coefficients will have a low regional mean
fALFF as well as a small functional volume. Therefore, fALFF1 contained much information from functional regional volume, which made it difficult to determine if it was the fALFF information or the functional volume information that was actually relevant to the classification of ADHD patients vs. TDC. To exclude functional volume information from fALFF1, we calculated fALFF2 by excluding the voxels with zero fALFF coefficients when calculating the regional mean fALFF. In specific, we calculated the regional mean over the voxels with non-zero fALFF coefficients and denoted it as fALFF2.

![Cumulative Probability](image)

**Figure 6-2.** Empirical cumulative distribution function for the fALFF coefficients.

*Anatomical Volume* As described in section *Preprocessing of images*, we obtained three tissue density maps, namely GM, WM and CSF, for each subject from the segmentation of the
anatomical image. We applied the CC400 atlas to the tissue density maps, and calculated the regional mean tissue density as GM, WM and CSF volume, respectively.

**Demographic Variables** The NYU dataset provided information for 8 demographic variables: gender, age, handedness, verbal IQ, performance IQ and Full4 IQ, Full2 IQ and medication status. The IQ scores of the participants were evaluated using the Wechsler Abbreviated Scale of Intelligence (WASI). Full2 IQ and Full4 IQ were different estimations of the full-scale IQ score. Only the former 6 demographic variables were used for the present study due to the substantial missing data for the latter two variables. Among the six used variables, only gender was categorical and the remaining five variables were continuous. The gender and age information were available for all the 215 participants. There were a couple of missing values for the other 4 demographic variables. Before the imputation of the missing values, the 6 demographic variables were scaled using Equation (6.1).

\[
x' = \frac{x - \min(x)}{\max(x) - \min(x)}
\]  

(6.1)

The above scaling enabled the different demographic variables to contribute equally during the imputation of missing values. We used the nearest-neighbor method for imputation, with k=10. We also tried k=20, the imputed data only changed a little bit when compared to that generated with k=10. Given a subject with missing data, we calculated the Euclidean distance between this subject and all other subjects based on the demographic variables for which this subject had non-missing values. Then, we found the 10 subjects with the smallest Euclidean distance and calculated a weighted mean for each demographic variable whose value was missing for the current subject based on the values of these 10 neighbors. The weights were calculated as the reciprocals of the Euclidean distances. The weighted mean was subsequently used as the imputed

6.2.4 Feature selection

In this study, we also designed a two-step feature selection algorithm by integrating prior knowledge with the recursive feature elimination (RFE) algorithm. As we mentioned before, the functional volume varied across different subjects even though the images had been normalized to the standard template. We speculated that the above variance was resulted from two possible reasons: individual difference and ADHD status. We would like to exclude the brain regions with only small fluctuations which were likely to be caused by individual difference, because such random fluctuations will not provide any predictive information for the classification of ADHD vs. TDC, or even confound the classification. In order to select the brain regions that were likely to be affected by ADHD, we calculated the variance for each feature/region across different subjects. The features were subsequently sorted according to their variance in descending order. For feature selection, the first step was to select the top N (denoted as $topN$) features with the highest variance. The rationale to do so was that high variance was likely to be caused by ADHD instead of random fluctuations. Using the selected features from step one, step two was to perform a RFE to pick out the features that were correlated with the class labels. The RFE algorithm was described in section 4.4.4.

The parameters for feature selection were set as follow. Parameter $p$ in the RFE algorithm, which was the percentage of features to be removed in each iteration, was set to be 1% as described in section 4.4.4. We requested the algorithm to remove one feature at a time when the number of features in the model was below 100. For the parameter $topN$ in above variance selection as well as $threN$ in RFE algorithm (which was the number of features to be kept in the
final model), there was not a good way to determine the optimal values. We considered 3 values for \( topN \), namely 100, 200 and 351, and 10 different values for \( threN \), starting from 10 to 100 in steps of 10. We showed the classification performance using different combinations of \( topN \) and \( threN \) in the results section.

The above feature selection approaches were only applied to image features. To integrate the demographic variables with the image features in the classification model, we treated the demographic variables differently from the image features. The demographic variables were always retained in the model, while the image features were ranked and removed based on their variance or weights.

### 6.2.5 Classification model learning

We used the linear SVM model. The basic idea of SVM model was described in section 2.5 Model learning. After model training, the predicted model is denoted as \( \hat{y} = wX - b \), which can be used to classify a new sample as either an ADHD patient or TDC. Given the data (images and demographic information) of the new sample, we first calculate the feature vector \( X \) in the same way as that applied to training samples described in section 6.2.3 Features, then format \( X \) to \( X^e \) according to the feature selection results, and finally insert \( X^e \) into the model \( \hat{y} = wX - b \) to obtain a predicted score \( \hat{y} \) for the new sample. If \( \hat{y} \geq 0 \), the new sample is classified as an ADHD patient. Otherwise, it is a TDC.

### 6.2.6 Model evaluation

We employed the 10-fold cross-validation approach for model evaluation. In this approach, the original samples were randomly partitioned into 10 equally sized subsets. The randomness in splitting the samples prevented a fair comparison among different classifiers, e.g. classifiers with different feature sets, if different partitions were used for different classifiers. Therefore, we
generated the cross-validation partition only once, and applied this partition to all classifiers. This guaranteed that all the classifiers had the same partition. Nevertheless, a particular partition may favor one classifier over another. To avoid this problem, we ran 20 rounds of cross-validation. We generated 20 partitions beforehand and applied them to all of the classifiers. The mean performance as well as standard deviation across the 20 rounds of cross-validation were calculated for each classifier, and compared among different classifiers.

Based on the predictions for the testing samples, we calculated sensitivity, specificity, accuracy, and area under receiver operating characteristic curve (AUC) to evaluate the performance of the classifiers. Since the ADHD-200 dataset was an imbalanced dataset, balanced accuracy, which was expressed as \( \frac{(\text{sensitivity} + \text{specificity})}{2} \), was used to accommodate this imbalance in previous studies (Sato et al., 2012a, Lim et al., 2013). Thus, we also calculated the balanced accuracy, and focused on this measure when we compared among different classifiers.

6.2.7 Identification of Important Features

Feature importance was measured as the accumulated weights across different folds of cross-validation. Since we had 20 rounds of 10-fold cross-validation, there were in total 200 runs of feature selection and model learning. The importance of a feature was calculated with Equation (6.2).

\[
\text{importance}_i = \sum_{k=1}^{200} \sigma_i^{(k)} |w_i^{(k)}|
\]

where \( |w_i^{(k)}| \) was the absolute weight from the SVM model for the \( i^{th} \) feature during the \( k^{th} \) run, \( \sigma_i^{(k)} = 1 \) indicated that the \( i^{th} \) feature was selected by the feature selection algorithm during the \( k^{th} \) run, while \( \sigma_i^{(k)} = 0 \) indicated that this feature was not selected.
6.2.8 Linear Regression Analysis

A linear regression model has the general form as shown in Equation (6.3).

\[ Y = X\beta + \varepsilon \]  

(6.3)

where \( Y \) is an n-dimensional vector of a continuous response variable, \( X \) is the design matrix whose dimension is \( n \times p \), \( \beta \) is the p-dimensional vector of coefficients to be estimated, and \( \varepsilon \) is an n-dimensional vector of random errors. \( n \) is the number of samples and \( p \) is the number of predictors. Generally, we have Gauss–Markov assumptions on the random error vector \( \varepsilon \). It states that all random errors \( \varepsilon_i \) (\( i = 1, \ldots, n \)) are mutually independent and have constant variance \( \sigma^2 \) and zero mean. Under this assumption, the best linear unbiased estimator of \( \beta \) is \( \hat{\beta} = (X'X)^{-1}X'Y \), which can be obtained using ordinary least square. The variance of \( \hat{\beta} \) is \( \sigma^2(X'X)^{-1} \). In order to make inference or perform statistical testing, we often impose normality assumption on the random errors, i.e. \( \varepsilon \sim N(0, \sigma^2I) \), where \( I \) is the identity matrix. Under normality assumption, we can perform statistical tests on any coefficient in the model. Generally we use t-test to test whether a single coefficient equals to zero or not. It has the form

\[ T = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 c_{ii}}} \]

where \( T \) is the t-statistic, \( c_{ii} \) is the i-th diagonal element of matrix \( (X'X)^{-1} \),

\[ \hat{\sigma}^2 = \frac{1}{n-p} \sum_{k=1}^{n} (Y - \hat{Y})^2. \]

Using the above linear regression model, we compared the functional brain size between ADHD patients and TDC. Similar to the calculation of functional regional volume, functional brain size for a subject was calculated as the percentage of voxels with non-zero fALFF coefficients within the whole brain. The CC400 atlas was used as a mask for the brain. A voxel covered by the CC400 atlas was considered to be a voxel within the brain. Otherwise, it was a voxel outside the brain and was not included for the calculation of brain size. Using the brain
size calculated from fALFF maps, we analyzed the effect of ADHD disorder on the brain size. We performed a linear regression analysis as described above with brain size as the response variable and the ADHD index as one covariate. ADHD index was an overall measure of symptom severity. The higher the ADHD index, the more severe the symptom is. We used ADHD index instead of ADHD diagnosis labels because binarization would lose some information. Besides, personal characteristics, such as age and gender, may also affect the brain size. Therefore, we included the 6 demographic variables in the linear model as well.

In order to analyze how the ADHD disease affected different brain regions, we performed a linear regression analysis for each single brain region. The analysis was the same as the brain size analysis except that we used the functional regional volume instead of the functional whole brain size as the response variable. The predictors still included the ADHD index and the 6 demographic variables as above.

6.3 Results

6.3.1 Linear Regression Analysis Results

For the functional brain size analysis, the linear regression model was summarized in Table 6-1. As expected, the coefficient for ADHD index was negative, which indicated that the more severe the ADHD symptoms they have, the smaller their brains are. However, the coefficient for the ADHD index was not significantly different from zero according to the t-test results. Handedness and IQ scores were showed to be non-significant either. On the other hand, age and gender exhibited significant influence on the functional brain size, although the images had been normalized to the standard template. Due to their significant effects on brain size, it was reasonable to integrate the demographic variables or at least age and gender with the functional
volume features. A model integrating these two types of information was likely to have an improved classification performance.

**Table 6-1. Functional brain size analysis: a summary for the linear regression model.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient Estimate</th>
<th>Std. Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.622</td>
<td>0.152</td>
<td>6.12e-5</td>
</tr>
<tr>
<td>ADHD index</td>
<td>-0.045</td>
<td>0.041</td>
<td>0.273</td>
</tr>
<tr>
<td>Gender</td>
<td>0.060</td>
<td>0.021</td>
<td>0.005</td>
</tr>
<tr>
<td>Age</td>
<td>-0.122</td>
<td>0.039</td>
<td>0.002</td>
</tr>
<tr>
<td>Handedness</td>
<td>0.104</td>
<td>0.061</td>
<td>0.092</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>-0.350</td>
<td>0.559</td>
<td>0.531</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>-0.278</td>
<td>0.505</td>
<td>0.583</td>
</tr>
<tr>
<td>Full IQ</td>
<td>0.576</td>
<td>0.881</td>
<td>0.514</td>
</tr>
</tbody>
</table>

Among the 351 brain regions defined in the CC400 atlas, there were 78 regions whose volumes were exactly the same across all the subjects. We assumed that there was no disease effect on those regions, and excluded them from our analysis. Therefore, we performed a linear regression analysis for each of the 273 remaining regions and employed the False Discover Rate (FDR) procedure (Benjamini and Hochberg, 1995) to adjust the p-value for multiple testing. Twenty-six out of the 273 analyzed brain regions showed significant disease effect, e.g. p-value for the coefficient of ADHD index was below 0.05. None of the 26 regions survived through p-value correction using the FDR procedure. In our project, however, traditional p-value correction methods might be too stringent for two reasons. First, some brain regions had only subtle variance in functional volume across different individuals, which was likely to be random noise caused by individual difference instead of actual change in functional volume caused by disease. We were not likely to detect significance for those brain regions. Including those brain regions in the analysis, however, increased the number of multiple testing, which would decrease the discovery rate for other regions. Secondly, there were correlations between different brain regions. Therefore, FDR procedure (Benjamini–Hochberg) which relies heavily on independence
assumption may give a too stringent control (Benjamini and Hochberg, 1995). There are some other multiple testing procedures for dependence hypothesis (Benjamini and Yekutieli, 2001, Sun and Cai, 2009), but it is hard to verify whether they truly work for our problem since the dependence structure of different brain regions was unknown. So we decided not to implement those methods. If we cut the significance threshold at the uncorrected p-value of 0.01, six regions exhibited significant effects, and all the 6 regions had negative coefficients, which suggested that ADHD index was negatively correlated with functional volume. We projected the 6 regions to the brain space, and showed it in Figure 6-3. As we can see, these 6 regions were adjacent to each other and located in the frontal lobe exclusively.
Figure 6-3. Linear regression analysis: brain regions exhibited significant disease effect (uncorrected p-value < 0.01). The image was displayed in the neurological orientation using the xjview toolbox.

6.3.2 Classification performance without feature selection

The classification performance without feature selection was summarized in Table 6-2. The model using demographic characteristics alone achieved a balanced accuracy of 58.5%, which surpassed all image feature sets except for functional volume (FV) that had a balanced accuracy of 59.6%. Integrating the demographic variables with the image features helped to improve the classification performance when compared to the image features alone, except for fALFF1.
Nevertheless, none of the integrated feature sets, such as fALFF1+Demo and fALFF2+Demo (Table 6-2), outperformed the demographic characteristics alone, except for FV+Demo. Although FV was calculated from fALFF maps, it outperformed fALFF1 and fALFF2, especially after integrating demographic information into the model. Compared with fALFF2, fALFF1 achieved better performance, which was expectable since fALFF1 included much FV information as we explained in the section 6.1 Introduction. fALFF2 without integrating demographic information exhibited almost random classification with AUC and balanced accuracy around 0.5, which was much lower than that of fALFF1 and FV. This might indicate that FV encoded most of the predictive information, while fALFF itself had very limited predictive power. Furthermore, FV also demonstrated better performance than anatomical features (GM/WM/CSF). In summary, functional volume exhibited the highest discriminative power, outperforming the demographic variables, fALFF features and anatomical features.

Table 6-2. Classification performance without feature selection. Demo is short for demographic variables; FV is short for functional volume; fALFF1/fALFF2 is short for regional mean fALFF1/fALFF2; GM/WM/CSF is short for GM/WM/CSF volume; X+Demo is short for the combination of X feature set and demographic variables.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>sens. (%)</th>
<th>spec. (%)</th>
<th>accu. (%)</th>
<th>AUC</th>
<th>(sens+spec)/2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demo</td>
<td>70.9±2.4</td>
<td>46.2±1.8</td>
<td>59.6±1.6</td>
<td>0.65±0.02</td>
<td>58.5±1.6</td>
</tr>
<tr>
<td>FV</td>
<td>67.7±3.0</td>
<td>51.6±2.7</td>
<td>60.3±2.1</td>
<td>0.62±0.02</td>
<td>59.6±2.1</td>
</tr>
<tr>
<td>fALFF1</td>
<td>63.5±3.1</td>
<td>51.9±2.4</td>
<td>58.2±2.3</td>
<td>0.60±0.02</td>
<td>57.7±2.3</td>
</tr>
<tr>
<td>fALFF2</td>
<td>64.2±2.5</td>
<td>39.7±3.2</td>
<td>53.0±1.7</td>
<td>0.52±0.02</td>
<td>52.0±1.8</td>
</tr>
<tr>
<td>GM</td>
<td>63.3±1.8</td>
<td>46.7±3.8</td>
<td>55.7±1.8</td>
<td>0.56±0.02</td>
<td>55.0±2.0</td>
</tr>
<tr>
<td>WM</td>
<td>56.9±2.5</td>
<td>47.7±2.6</td>
<td>52.7±2.1</td>
<td>0.51±0.02</td>
<td>52.3±2.1</td>
</tr>
<tr>
<td>CSF</td>
<td>56.4±2.8</td>
<td>39.6±3.4</td>
<td>48.7±2.7</td>
<td>0.49±0.03</td>
<td>48.0±2.8</td>
</tr>
<tr>
<td>FV+Demo</td>
<td>68.2±3.3</td>
<td>54.5±2.4</td>
<td>62.0±2.3</td>
<td>0.64±0.02</td>
<td>61.4±2.2</td>
</tr>
<tr>
<td>fALFF1+Demo</td>
<td>62.5±2.5</td>
<td>52.4±2.7</td>
<td>57.9±2.0</td>
<td>0.62±0.02</td>
<td>57.4±2.0</td>
</tr>
<tr>
<td>fALFF2+Demo</td>
<td>62.3±2.6</td>
<td>44.1±3.6</td>
<td>54.0±2.4</td>
<td>0.58±0.02</td>
<td>53.2±2.5</td>
</tr>
<tr>
<td>GM+Demo</td>
<td>62.8±2.4</td>
<td>51.7±2.1</td>
<td>57.7±1.6</td>
<td>0.61±0.01</td>
<td>57.3±1.6</td>
</tr>
<tr>
<td>WM+Demo</td>
<td>58.3±2.2</td>
<td>50.2±2.5</td>
<td>54.6±1.9</td>
<td>0.55±0.02</td>
<td>54.2±1.9</td>
</tr>
<tr>
<td>CSF+Demo</td>
<td>62.1±3.0</td>
<td>49.2±2.8</td>
<td>56.2±2.2</td>
<td>0.56±0.02</td>
<td>55.6±2.2</td>
</tr>
</tbody>
</table>
6.3.3 Classification performance with feature selection

Since integrating demographic variables tended to improve the classification performance, we only considered the models integrating demographic variables for this section. Performance under different parameters was shown in Figure 6-4. FV and fALFF1 exhibited similar pattern. The performance of topN=200 and topN=351 was close to each other, but obviously worse than that of topN=100, no matter what the value of threN was. topN=351 represented the situation where we did not preselect the brain regions with high variance and submitted all the features to the RFE process, which was actually the standard RFE. Our results suggested that a pre-selection based on variance before RFE helped to improve the classification performance. In another word, including the features with small variance would disturb the feature selection algorithm. When topN=100, the performance tended to reach a relatively stable state when threN=50 for FV and threN=30 for fALFF1. We reported the classification performance in Table 6-3 for FV with topN=100, threN=50 as well as the performance for fALFF1 with topN=100, threN=30. For all the other feature sets including fALFF2, GM, WM and CSF, we reported their best performance across all the combinations of topN and threN in Table 6-3. Clearly, this might overestimate the actual classification performance for those four feature sets. Nevertheless, the performance for those four features sets was still much worse than the performance of FV and fALFF1.
Figure 6-4. Classification performance with feature selection using different parameters.

Table 6-3. Classification performance with feature selection

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>sens. (%)</th>
<th>spec. (%)</th>
<th>accu. (%)</th>
<th>AUC</th>
<th>(sens+spec)/2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FV+Demo</td>
<td>78.1±2.9</td>
<td>57.3±2.3</td>
<td>68.6±1.7</td>
<td>0.71±0.01</td>
<td>67.7±1.7</td>
</tr>
<tr>
<td>fALFF1+Demo</td>
<td>75.2±2.2</td>
<td>58.8±2.5</td>
<td>67.7±1.5</td>
<td>0.72±0.02</td>
<td>67.0±1.5</td>
</tr>
<tr>
<td>fALFF2+Demo</td>
<td>66.1±3.3</td>
<td>48.4±3.4</td>
<td>58.0±2.8</td>
<td>0.64±0.02</td>
<td>57.2±2.8</td>
</tr>
<tr>
<td>GM+Demo</td>
<td>67.0±3.2</td>
<td>50.9±2.4</td>
<td>59.7±2.3</td>
<td>0.64±0.02</td>
<td>58.9±2.2</td>
</tr>
<tr>
<td>WM+Demo</td>
<td>69.5±3.1</td>
<td>54.9±4.4</td>
<td>62.9±3.0</td>
<td>0.66±0.02</td>
<td>62.2±3.1</td>
</tr>
<tr>
<td>CSF+Demo</td>
<td>70.1±1.8</td>
<td>55.7±1.9</td>
<td>63.5±1.4</td>
<td>0.67±0.02</td>
<td>62.9±1.4</td>
</tr>
</tbody>
</table>

6.3.4 Important features

The feature importance was calculated based on the model integrating FV and demographic variables. Feature selection was performed with topN = 100 and threN = 50. The feature importance was calculated using Equation (6.2). Features were ranked according to their importance in descending order. We projected the top 10 features to the brain space as shown in Figure 6-5. The anatomical information for those 10 brain regions was summarized in Table 6-4.
Feature F was one of the 6 significant regions (uncorrected p-value < 0.01) detected by the linear regression analysis. Besides, some important regions, e.g. region A and G, appeared to be adjacent to each other. Therefore, we marked the top 10 regions in a single brain, and showed it in Figure 6-6. The predictive regions were mainly located in cerebellum posterior lobe, occipital lobe, temporal lobe, frontal lobe and parietal lobe. Those findings were consistent with the findings from anatomical images as we discussed in section 6.4 Discussion.

Figure 6-5. The top 10 discriminative brain regions that distinguished between ADHD patients and TDC. Images were displayed in the neurological orientation using xjview toolbox.
Table 6-4. Anatomical information for the top 10 discriminative brain regions shown in Figure 6-5.

<table>
<thead>
<tr>
<th>Region Index</th>
<th>Central Coordinates</th>
<th>Anatomical Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(-4,-100,-22)</td>
<td>Left Occipital Lobe</td>
</tr>
<tr>
<td>B</td>
<td>(24,-84,-54)</td>
<td>Right Cerebellum Posterior Lobe</td>
</tr>
<tr>
<td>C</td>
<td>(-40,-68,-62)</td>
<td>Left Cerebellum Posterior Lobe</td>
</tr>
<tr>
<td>D</td>
<td>(24,-72,54)</td>
<td>Right Parietal Lobe</td>
</tr>
<tr>
<td>E</td>
<td>(40,-52,-62)</td>
<td>Right Cerebellum Posterior Lobe</td>
</tr>
<tr>
<td>F</td>
<td>(-32,28,42)</td>
<td>Left Frontal Lobe</td>
</tr>
<tr>
<td>G</td>
<td>(-20,-100,-22)</td>
<td>Left Occipital Lobe</td>
</tr>
<tr>
<td>H</td>
<td>(60,-68,-22)</td>
<td>Right Temporal Lobe</td>
</tr>
<tr>
<td>I</td>
<td>(32,-68,-66)</td>
<td>Right Cerebellum Posterior Lobe</td>
</tr>
<tr>
<td>J</td>
<td>(-48,-76,-46)</td>
<td>Left Cerebellum Posterior Lobe</td>
</tr>
</tbody>
</table>
Figure 6-6. The top 10 regions in Figure 6-5 were combined and displayed in the brain space simultaneously.

6.4 Discussion

In the present study, we have initially proposed the concept of functional volume. We compared the discriminative power of functional volume with that of anatomical volume for the classification of ADHD patients vs. TDC, and the former demonstrated better performance than the latter. One possible explanation for this might be that some brain areas may have necrotic or functionally deficient brain tissues. Such necrotic tissues can be captured by anatomical images, and prevent an accurate measure of brain volume using anatomical images. By contrast,
functional volume would automatically exclude such necrotic tissues, because necrotic tissues do not have fMRI signal. Thus, functional volume may serve as a better measure for brain volume than anatomical volume. Future work is needed to verify this theory.

Another major contribution of this study was that we had improved the accuracy of ADHD diagnosis. Although functional volume had comparable performance as fALFF1, our study shed light on why the regional mean fALFF distinguished between ADHD patients and TDC. According to our analysis, it was the functional volume other than fALFF itself that encoded the discriminative information. Upon this observation, we made two major improvements based on the model in (Sato et al., 2012b). The first improvement was to integrate the demographic information with the regional mean fALFF. Since the regional mean fALFF mainly captured brain volume information and brain volume was likely to be related to demographic characteristics, integration of these two types of information became straightforward. Otherwise, combination of fALFF maps with demographic characteristics might be a little bit strange. Inspired by the observation in this study, we designed a two-step feature selection algorithm by integrating the variance threshold approach with the standard RFE algorithm. This two-step algorithm worked well for both functional volume and fALFF1, and performed much better than the standard RFE algorithm. The above two improvements helped to improve the balanced accuracy from 57.7% to 67.0% for the fALFF1 feature set. Finally, (Sato et al., 2012b) reported that combination of ReHo and fALFF contained relevant information to discriminate ADHD patients from TDC. They compared 10 different classifiers and suggested that all classifiers provided almost the same performance for the classification of ADHD patients vs. TDC. Therefore, we calculated the regional mean ReHo and fALFF using the CC400 atlas as in (Sato et al., 2012b), and then used them as features to train a linear SVM model. The model was
evaluated in the same way as all the other models in this study. This SVM model based on ReHo and fALFF achieved a balanced accuracy of 51.6%, which was 16.1% lower than our model using functional volume with feature selection.

Furthermore, we highlighted 10 brain regions that distinguished ADHD patients from TDC. There were plenty of studies that compared regional brain volume between ADHD patients and TDC. However, previous studies were usually based on anatomical images with a limited sample size, and analyzed the data using univariate approaches. The regions affected by ADHD remained controversial. Hopefully, our present multivariate analysis based on functional images with a very large sample size would provide some new perspectives about the brain regions affected by ADHD. Our classification algorithm detected functional volume differences in cerebellum posterior lobe, occipital lobe, temporal lobe, frontal lobe and parietal lobe. While previous ADHD morphometry research has implicated nearly the entire cerebral cortex as well as many brain substructures (e.g., caudate) as being smaller or less developed in children with ADHD than controls, the areas implicated in this study include several brain regions that have been consistently implicated across studies or have been shown to have a large between-group effect size. For example, Valera et al. (Valera et al., 2007) found that the brain region with the largest reduction in patients with ADHD compared to controls was the cerebellum, both in specific regions (i.e., inferior vermis) as well as more globally (i.e., both right and left cerebellum). Another common finding in ADHD structural imaging studies is that patients with ADHD have globally reduced grey matter in the cortices (Shaw et al., 2006, Valera et al., 2007). However, these reductions in cortical grey matter are often found to be greatest in frontal regions (Valera et al., 2007). Though we did find reduced left frontal lobe volume among patients with ADHD compared to typically developing controls, the cortical areas that demonstrated reduced
volume included occipital (left), parietal (right), and temporal (right) regions as well. Our findings of more diffuse reductions across the cortex may be the result of the sensitivity of the fALFF indicator to not only determine if matter exists at each voxel as is done when typically measuring brain volumes but instead lets us determine whether matter in each voxel is functional (at-rest). Perhaps the maturational delays observed in ADHD brain development reflect not only development of brain tissue but also development of the networks that integrate newly-developed brain cells into active networks. It should be noted that our study did not detect smaller volumes in many sub-cortical structures, such as the caudate, putamen, and globus pallidus, which have repeatedly been found to be reduced in patients with ADHD (Ellison-Wright et al., 2008, Nakao et al., 2011, Frodl and Skokauskas, 2012). This is expected because a grey matter mask was applied to select the voxels to be included in the CC400 atlas. As a result, the striatal regions are excluded by the CC400 atlas.

Although CC400 atlas was clearly a functional atlas generated from fMRI data, it was also used for the anatomical images in this project. For a fair comparison, we had tried the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) for the GM, WM and CSF maps. The AAL atlas is defined based on brain anatomy. It divides the brain into 116 regions, including 90 cerebrum regions and 26 cerebellum regions. We calculated the regional mean tissue density as the features. The approach to calculate the feature values was the same as the approach described in paragraph Anatomical Regional Volume, except replacing the CC400 atlas with the AAL atlas. The model training and evaluation was the same as the other models in this study. The classification performance without feature selection was shown in Table 6-5. For feature selection, we applied the standard RFE algorithm, since there were only 116 features. The percentage of features to be removed at each iteration was set to be 1% and the number of
features to be kept in the final model was set to be 10 to 100, stepping by 10. The best classification performance across different parameters was shown in Table 6-6. As we can see, performance for the anatomical features using AAL atlas was still worse than the performance of functional volume features. Further, CC400 atlas might not be the optimal way to segment the brain for calculating functional volumes either. As we can see in Figure 6-5, the top 10 discriminative regions were adjacent to each other, suggesting that the regions could be further merged into larger regions. A brain segmentation characterizing the functional shrinkage pattern in brains of ADHD patients may further improve the classification performance for the functional volume features. Investigating the optimal way of segmenting the brain to maximize the classification performance might be a future direction of work.

### Table 6-5. Classification performance of anatomical volume using AAL atlas without feature selection

<table>
<thead>
<tr>
<th>features</th>
<th>sens. (%)</th>
<th>spec. (%)</th>
<th>accu. (%)</th>
<th>AUC</th>
<th>(sens+spec)/2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>70.3±2.7</td>
<td>42.4±2.5</td>
<td>57.6±1.7</td>
<td>0.57±0.01</td>
<td>56.4±1.7</td>
</tr>
<tr>
<td>WM</td>
<td>66.4±2.3</td>
<td>44.2±2.9</td>
<td>56.3±2.0</td>
<td>0.56±0.02</td>
<td>55.3±2.1</td>
</tr>
<tr>
<td>CSF</td>
<td>67.3±2.2</td>
<td>34.5±2.4</td>
<td>52.3±1.6</td>
<td>0.54±0.02</td>
<td>50.9±1.6</td>
</tr>
<tr>
<td>GM+Demo</td>
<td>67.1±2.0</td>
<td>49.6±2.1</td>
<td>59.1±1.5</td>
<td>0.65±0.01</td>
<td>58.3±1.5</td>
</tr>
<tr>
<td>WM+Demo</td>
<td>65.5±1.7</td>
<td>52.9±2.7</td>
<td>59.8±1.9</td>
<td>0.63±0.01</td>
<td>59.2±1.9</td>
</tr>
<tr>
<td>CSF+Demo</td>
<td>67.1±2.6</td>
<td>51.9±2.9</td>
<td>60.2±2.5</td>
<td>0.64±0.02</td>
<td>59.5±2.5</td>
</tr>
</tbody>
</table>

### Table 6-6. Classification performance of anatomical volume using AAL atlas with feature selection

<table>
<thead>
<tr>
<th>features</th>
<th>sens. (%)</th>
<th>spec. (%)</th>
<th>accu. (%)</th>
<th>AUC</th>
<th>(sens+spec)/2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM+Demo</td>
<td>67.8±1.9</td>
<td>49.9±2.5</td>
<td>59.7±1.7</td>
<td>0.65±0.01</td>
<td>58.9±1.7</td>
</tr>
<tr>
<td>WM+Demo</td>
<td>66.0±2.6</td>
<td>53.0±2.6</td>
<td>60.0±2.1</td>
<td>0.64±0.01</td>
<td>59.5±2.1</td>
</tr>
<tr>
<td>CSF+Demo</td>
<td>67.2±2.1</td>
<td>52.4±2.7</td>
<td>60.4±2.1</td>
<td>0.63±0.02</td>
<td>59.8±2.1</td>
</tr>
</tbody>
</table>
6.5 Conclusion

This study started out to build an accurate classifier for the automatic diagnosis of ADHD. Through a close examination of existing classification models, we first observed that there was a set of voxels whose fALFF coefficients were zero, although those voxels were within the boundary of the brain. According to the definition of fALFF, those voxels with zero fALFF coefficients were functionally deficient, and therefore we proposed to quantify the functional volume of the whole brain or a brain region using fALFF maps. The functional brain size measured from fALFF maps exhibited significant correlations with demographic characteristics such as age and gender, although the images had been normalized to the standard template. This result not only pointed to the weakness of current spatial normalization algorithm, but also demonstrated the essentiality to control those personal variables while comparing the functional brain size between different groups of individuals, e.g. comparing the functional brain size of the patient group with that of the healthy controls. Otherwise, the analysis result might be misleading. Furthermore, we calculated the functional volumes for the brain regions defined in the CC400 atlas, and used them as features for the classification of ADHD patients vs. TDC. Functional volumes demonstrated much higher discriminative power than anatomical volumes which were calculated in the same way using the tissue density maps. With our two-step feature selection algorithm, the model based on functional volumes also exhibited much better classification performance in comparison with related models in literature. Finally, our results also demonstrated that fALFF itself had very limited predictive power for the classification of ADHD patients vs. TDC. The regional mean fALFF calculated using the traditional method was highly correlated with functional volume. Cautions should be taken when we attempt to compare the fALFF between different groups of individuals, especially when there is brain shrinkage for one
group of individuals. In fact, this observation is not limited to fALFF alone. Any voxel-wise features, e.g. ReHo, would have the same problem. It might be not fair enough to claim that some brain regions have decreased fALFF or ReHo, if the decreases in fALFF/ReHo are in fact caused by the reduced brain volume.
Chapter 7 Conclusion & Future Directions

Machine learning techniques have demonstrated remarkable power for the analysis of fMRI data. They have been widely used for the automatic diagnosis of brain disorders, and showed great success in a variety of diseases (Shinkareva et al., 2006, Demirci et al., 2008, Zhu et al., 2008, Shen et al., 2010, Fan et al., 2011, Mourao-Miranda et al., 2011, Rizk-Jackson et al., 2011, Dai et al., 2012b, Weygandt et al., 2012, Anderson et al., 2014, Cao et al., 2014, Gamboa et al., 2014, Watanabe et al., 2014, Challis et al., 2015, Du et al., 2015, Kim et al., 2015, Rosa et al., 2015). For the three applications in this dissertation, machine learning models also achieved relatively high classification accuracy, although they were not ready to be used in clinical settings yet.

Take the prediction of CI outcomes for example. Our previous simple linear regression analysis, with age at implantation, pre-implant hearing threshold and brain activity at a region of interest as the independent variables, failed to provide a method by which fMRI data from individual subjects can be used to make a prognosis about possible speech and language outcomes. However, two features extracted by our machine learning algorithm appear to be able to provide us with a method of doing so. Based on this preliminary result we are optimistic that a reliable machine learning model based on a larger training set can eventually be applied in the clinical setting to provide specific prognostic information to patients considering cochlear implantation.

Machine learning models also demonstrated significant superiority for deriving disease biomarkers. For example, our machine learning model highlighted two brain regions as biomarkers for the prediction of CI surgery outcomes. One of these two brain regions is aligned with our original hypothesis, while previous univariate correlation analysis or multivariate regression analysis failed to provide persuasive predictive value by looking at a region of interest in this part of the brain alone. The other brain region, whose predictive power was
underestimated according to our original hypothesis, appeared to be surprising at the first thought. Through a deeper investigation, we found substantial number of scientific publications supporting this discovery and achieved a better understanding of the brain function. Our work demonstrated the advantage of machine learning techniques, which can automatically detect the predictive features and draw our attention to features that are important but beyond our prior knowledge.

Although machine learning techniques showed great success for the analysis of fMRI data, it should be admitted that machine learning analysis of fMRI data is still in its early stage. The success in this field would require the close collaboration of neuroimaging experts with machine learning experts. This thesis may serve as a good start for both neuroimaging experts and machine learning experts who attempt to do research in this field. Additionally, public databases with preprocessed MRI images should be built to encourage more machine learning experts to work on the neuroimaging data. The Autism Brain Imaging Data Exchange (ABIDE) and ADHD-200 are two initiatives in this direction.

As we explained in Chapter 2, feature extraction was the most challenging part for the machine learning analysis of fMRI data, because it was very difficult to define discriminative features for the fMRI data due to its complex format. Most of previous studies used each single voxel as a feature, which was obviously not the optimal way to define features for reasons as we explained in Chapter 4. Deep learning, which was known as the next-generation machine learning, has gained popularity in recent years for the analysis of MR images (Hjelm et al., 2014, Suk et al., 2014, Kim et al., 2015, Zhang et al., 2015). The input for deep learning could be a matrix with each row representing a participant and each column representing a voxel. The deep learning architecture would be able to generate high-level features automatically by combining
the voxels in some manner. Compared with the traditional machine learning architecture, the superiority of deep learning for the MR images is quite obvious. In traditional machine learning, we construct hand-crafted features according to our prior knowledge, and therefore the features that we can obtain are usually limited by our prior knowledge. On the contrary, deep learning algorithms start out with the data, and generate data-driven features, which are supposed to characterize the structure of data better than the hand-crafted features. For example, we may define regions of interests according to our prior knowledge, but have limited knowledge about the role of previously undiscovered regions as key elements in brain networks related to a certain cognitive task. However, the deep learning algorithms might be able to detect such network-level patterns and derive numerical measures to characterize those patterns automatically. Due to the above superiorities, moving from the traditional machine learning to deep learning might be a future trend in this field.

Another future direction of research might be the integration of multiple modalities in a single machine learning model. Data obtained from different imaging modalities are supposed to contain information that is complementary to each other. This theory has been proven in previous studies (Szaflarski et al., 2010, Wang et al., 2012d), as well as our work in Chapter 3. Nevertheless, how to integrate the information from different modalities is still a research issue waiting for further investigations.
References


Geers AE, Hayes H (2011) Reading, writing, and phonological processing skills of adolescents with 10 or more years of cochlear implant experience. Ear and hearing 32:49S-59S.

Geers AE, Sedey AL (2011) Language and verbal reasoning skills in adolescents with 10 or more years of cochlear implant experience. Ear and hearing 32:39S-48S.


Rolando L (1809) Saggio sopra la vera struttura del cervello dell'uomo e degli animali, e sopra le funzioni del sistema nervoso. Sassari.


Schmithorst VJ, Holland SK, Dardzinsk BJ (2010) CCHIPS: Cincinnati Children's Hospital Image Processing Software


Wernicke C (1874) Der aphasische Symptomencomplex.


xjView http://www.alivelearn.net/xjview.


Publications during Doctoral Study

*Peer-reviewed Publications*


Deshpande AK, **Tan L**, Lu LJ, Altaye M and Holland SK. fMRI as a Pre-Implant Objective Tool to Predict Post-Implant Oral Language Outcomes in Children with Cochlear Implants. Ear and Hearing. (Submitted)


*Book Chapter*