"It is certain that there may be extraordinary mental activity with an extremely small absolute mass of nervous matter: thus the wonderfully diversified instincts, mental powers, and affections of ants are notorious, yet their cerebral ganglia are not so large as the quarter of a small pin's head. Under this point of view, the brain of an ant is one of the most marvelous atoms of matter in the world, perhaps more so than the brain of a man."

- Charles Darwin

"No bird soars too high, if he soars with his own wings."

-William Blake

"We have existence and it's all we share"

"You're a slave to money, then you die"

"There's no time, no space, no law. We're out here on our own"

"History will have a place for us, it may take three albums, but we'll be there"

-Richard Ashcroft

"Some people want it to happen, some wish it would happen, others make it happen."

"I can accept failure, everyone fails at something. But I can't accept not trying.

-Michael Jordan

University of Alberta

Dimensionality Reduction for fMRI Diagnostic Systems

by

Gagan Sidhu

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Computing Science

Fall 2012
Edmonton, Alberta

*This work is dedicated to my Mother and Father, Her, and all the ****** dat been down from day one.*

Abstract

Functional Magnetic Resonance Imaging (fMRI) measures the dynamic activity of each voxel of a brain. This dissertation addresses the challenge of learning a *diagnostic classifier* that uses a subject's fMRI data to distinguish subjects with neuropsychiatric disorders from healthy controls. fMRI intrinsically possess *spatial* and *temporal* dimensions, given by a waveform over hundreds of time points at each of 10^5 spatial locations. Given training data of only dozens to hundreds of subjects, standard learning algorithms will over-fit – i.e., do well on the training data, but poorly on novel instances. We address this by reducing the dimensionality, using several variants of Principal Component Analysis (PCA). We evaluate the performance of the PCA Variants on two datasets: Attention-Deficit Hyperactivity Disorder (ADHD) [a large public dataset of 668 subjects, used for the ADHD200 competition] and First Episode Psychosis [involving 34 subjects]. Our empirical studies show that using *non-linear* PCA to reduce fMRI dimensionality over *both* the spatial and temporal dimensions is statistically better, with respect to the classification task, than using a linear mapping to reduce over only the spatial or only the temporal dimension.

Acknowledgements

Where to start? I am a product of the people who given me the honor of being in their company. Throughout my life, there have been innumerable people that gave me so much more than I ever gave them; I was blessed with family and friends who exercised an extraordinary, and often undeserving, amount of patience for my behavior. This list covers everyone who has "put up with me" over an extended period of time, and of course there is no order:

Lauren Styles a great woman who made me want to be a great man.

Tina and Gurmeet Sidhu, and Gurdev Gill (nanima) Their endless support and blind faith allowed me to figure out my life. Even during the rough times in my early undergraduate years, my parents never stopped supporting me.

Justin Sidhu My brother always kept me out of trouble, even when my mouth would ask for it. If it wasn't for his brotherly love, I would be in a different position than I am today.

Kuljit Thiara One of my dearest and most noble friends that always gave me the truth and put me before himself.

Indy Sagoo Another one of my dearest childhood friends who looked past my lapses in judgment and malicious behavior as a youth. It was his friendship that led me to believe everyone could be as good of a friend, which was not the case.

Manjinder Sagoo My familiarity, knowledge and expertise with computers is because of this man's mentorship in my youth. His influence instilled my insatiable curiosity for the computer.

Michael Robert Webb On one of the last days of in the Winter semester of my fourth undergraduate year, Mike asked me a question for which I had no answer:

> *I do well in school to pay for my education, and my girlfriend of 8 years works to support both of us. Why are you proud of a 3.0 GPA when you're not putting forth a full effort, especially when your school is paid for?*

It was this quote that motivated me to achieve my potential.

Dr. Russell Greiner One full year after Mike said the above, I had turned around my GPA significantly as I wanted to do a Masters. However,

four of my five undergraduate years were very poor; I had no right to even *think* of graduate school. It was because of Dr. Russell Greiner that I was admitted into the program, and I am hopeful this dissertation reflects my immeasurable gratitude for his support.

Dr. Brenda Leskiw It was because of Dr. Leskiw that I was able to graduate with a BSc. in Computing Science, even though three of my five undergraduate years were very poor. She gave me a "second chance", and I am hopeful that senior undergraduates realize that there is hope if you are willing to work hard.

I would like to also thank my teachers at Terrace Heights Elementary School:

Mrs. Catherine Suen was my grade 1 teacher. She made sure I knew how to properly put on *and* tie my shoes because I never cared to learn how to do either; boy, did she ever make me care. I learned how to tie and put on my shoes over one 15 minute recess.

Mrs. Frances Stead was my grade 3 teacher that exercised an extreme amount of patience with me. She read us The Sandwich by Ian Wallace [33], which was a story that gave me courage to resist peer pressure. She also read us poems from the great Shel Silverstein, a man whose work also had a large impact on me.

Mr. Jeff Huculak was my grade 6 teacher, who I did not like initially because he wasn't the "old" grade 6 teacher named Mr. Norris. Mr. Huculak gave me "the talk" during recess; he told me things change in life and it's what we make of these changes that determine our happiness. He couldn't have been more right: he turned out to be one of my favorite teachers. P.S. I still remember that you and I cried at the grade 6 "grad" Mr. Huculak ;) You had a legitimate excuse though, as we were the first class you taught.

Mr. Stewart was my principal from kindergarten to grade 3. His discipline was amongst the most effective at deterring my often-outlandish behavior. I will never forget the Friday assemblies we would have before Christmas break, where we would sing carols all day in the gymnasium and he would sing/play piano, or the time when he managed to get our school 5 computers equipped with the internet, which changed my life forever. May you rest in peace.

A big thanks to my teachers at Ottewell Junior High School: Mrs. Park, Ms. Russell (Mrs. Faye), Mr. Woodward, Mr. Fennell, Mrs. Armstrong, Mrs. Rebman, Mr. Loxam, and of course Mr. Feary. All of you were responsible for giving a surreal, and memorable, junior high school experience.
My teachers at Harry Ainlay Composite High School, all of whom were my "handlers" in some capacity:

Mr. Keir Jenkins, aka "The Chem God" was my Science 10IP teacher. He emphasized the importance of "doing your homework and not just doing well on the exams". It took 8 years for this lesson to sink in.

Mrs. Linda-Rae Carson was my History 10IB teacher. She spotted my ADHD while I was in her amazing History class; it was because of her observation that I was formally diagnosed 3 years later. She always gave me a "little extra" leeway because of my eagerness and constructive contributions to the class discussions. History 10IB was among the funnest classes I have ever been apart of.

Mr. Steve Hardy was my Math 10I and 30I teacher. He warned that my shortcuts would create "holes in [my] math", and boy was he right: by the time my undergraduate degree was completed, I had *finally* patched most of these holes.

Mrs. Edie Ferris was my Chem 15IB/20IB teacher that put up with too many of my shenanigans. I will never forget the time she said "No" when I asked her to buy a pop from the vending machine, and then I asked her if I could "go get a drink" and came back with a pop. This act was a culmination of "button-pushing" that earned me a suspension for "willful defiance"– a priceless/humorous memory that I will never forget. I always admired, and respected, her knowledge and passion for Chemistry–she made it fun.

Mr. David Visser was my Physics 30 teacher. He always encouraged me to achieve my potential, and had a passion for physics that I carry with me to this day.

Mr. Leno Delcioppo was my Chemistry 30 teacher. Man, Chemistry was never supposed to be fun. Not only did Mr Delcioppo know his stuff, he also shared plenty of humorous life stories that made me wiser. I was spoiled having Mr. Jenkins, Mrs. Ferris and Mr. Delcioppo for high school Chemistry.

Mrs. Christine Peters was my principal who always gave me *just enough* rope (Thanks Mrs. Peters!).

To all of the guys on EFNet, many of whom I've known longer than a decade: I ain't got nothin' but love. - Broly.

Contents

Chapter 1

Introduction

Over time, Moore's law has modeled the increasing availability of computational resources, motivating multiple industries/disciplines to use these resources to improve their products and services. Algorithms in Machine Learning, which is a discipline that focuses on using past data to optimize some performance criterion [2], have improved multiple industries' products and services, and the increased availability of computational resources has further broadened their application to cutting-edge medical research that involves *high dimensional data* such as microarrays [35], Single Nucleotide Polymorphs (SNP) [6], and functional Magnetic Resonance Imaging (fMRI) [9].

In this dissertation, we use Machine Learning algorithms to analyze functional Magnetic Resonance Images, which are images that provide an indirect measure of brain activity. Unlike Magnetic Resonance Imaging (MRI), which is used primarily to measure structural images of the organs, fMRI measures hemodynamic changes in the brain caused by changes in neural activity [21].

The most common form of fMRI measures the blood oxygenation level dependent (BOLD) signal, which is an indirect measure of neuronal activity based on changes in blood oxygenation, blood volume, and blood flow rate. Full details of how changes in neuronal activity cause changes in the BOLD signal are not completely understood [23]. Evaluating the changes in BOLD contrast could assist in providing a neurological basis for diagnosing neuropsychiatric disorders. Evaluating these changes across the entire brain using techniques from Machine Learning may be one way to extract this neurological basis.

Psychiatrists diagnose neuropsychiatric disorders, defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM) [1], using clinical assessments that include: evaluating the background demographics, collecting first and third party observations, and a structured psychiatric interview with the subject [27]. However, clinical assessments are highly dependent on the training of the interviewer [34], and it is feasible that this dependence can be reduced if clinicians have information about the subject's neurobiology to supplement their diagnosis. Furthermore, the diagnosis of specific neuropsychiatric disorders is imperfect. For example, bipolar disorder has a low diagnostic reliability, and is consequently more likely to be missed than correctly

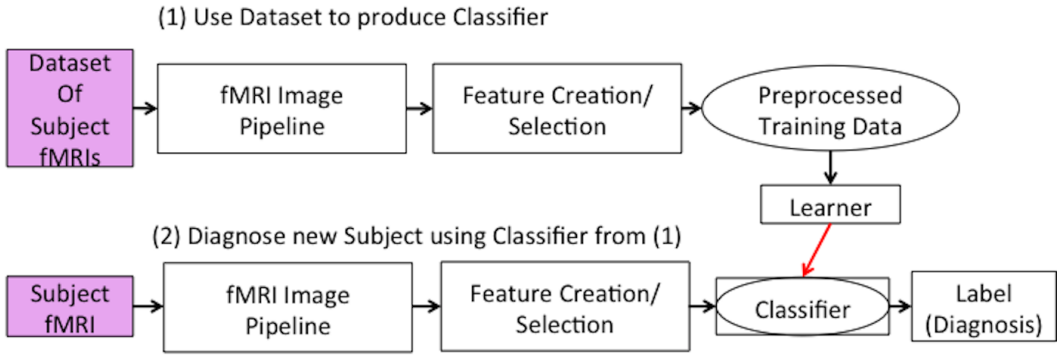

Figure 1.1: Our learning system in two stages, where the first develops the classifier responsible for diagnosing new subjects, which is then used in the second step.

diagnosed [32].

In the United States, there are other challenges associated with psychiatric diagnoses. This is exemplified by the variability in mental-health-related expertise among clinicians diagnosing neuropsychiatric disorders such as Attention-Deficit Hyperactivity Disorder (ADHD). Diagnoses for ADHD can be obtained from general practitioners, nurses, paediatricians, psychiatrists, or neuroscientists [31]. Combining the variability in mental-health-related expertise with the subjective nature of psychiatric diagnosis, there is a higher chance for a misdiagnosis.

It is desirable to decrease the subjectivity involved in diagnosing neuropsychiatric disorders, which may be possible by statistically analyzing fMRI to glean the respective subject's neurobiology. fMRI combined with machine learning / statistical analysis may provide new, objective, biologically-based measures that might assist with psychiatric diagnosis and prognosis. If we can apply Machine Learning algorithms to fMRI in order to diagnose subjects with high accuracy, future work can investigate the possibility of new diagnostic criteria that is based on fMRI.

For Machine Learning researchers, fMRI data is intriguing because it has both temporal *and* spatial dimensions, characterized by $L \times W \times H$ spatial voxels[1], each with waveforms of length T. fMRI datasets typically contain dozens to hundreds of subjects, where each subject's fMRI data contains $L \times W \times H \times T \approx 10^6$ features per subject, which may cause standard learning algorithms to over-fit – i.e., performing well on the training data, but poorly on novel instances. In Machine Learning, *dimensionality reduction* methods attempt to alleviate problems associated with high-dimensional data, either by selecting a subset of features in high dimensional space, *or* transforming this subset of features into a lower-dimensional space.

Reducing fMRI dimensionality is one part of a larger system that is responsible for learning a diagnostic classifier. Our learning system consists of two components, each of which involves many steps. At training time, the first

[1]Voxels are the three dimensional analog to pixels in two dimensions

component uses a number of subject scans to produce a *diagnostic classifier*, then at performance time, the second component uses the classifier to produce a diagnosis for a novel subject – i.e., a subject who was not used to develop the classifier. *Please note this objective – of producing a classifier – is different from the more standard associative task*, of determining, say, which specific regions of the brain are most correlated with some diagnosis [10].

As shown in Figure 1.1, both components involved first running the "fMRI Image Pipeline", followed by reducing the fMRI dimensionality in the "Feature Creation/Selection" stage. This dissertation focuses on the second "Feature Creation /Selection" stage, to show that *using non-linear mappings to reduce fMRI dimensionality over both the spatial and temporal domain will improve the classifier's discrimination between subjects with neuropsychiatric disorders*[2] *and healthy controls, in comparison to using a linear mapping that reduces over only the spatial or only the temporal domain.*

Figure 2.1 summarizes the "Feature Creation/Selection" step for the datasets. After several fixed steps (discussed in Section 2), we then consider three dimensionality reduction processes that apply some variant of Principal Component Analysis (PCA); these variants apply PCA over the temporal dimension (PCA-t) or over both the temporal and spatial dimensions (PCA-st and the kernelized variant kPCA-st).

As a "first step" towards our goal of developing a procedure to learn an effective diagnostic classifier, we opted for a *biologically naive* approach to fMRI dimensionality reduction; that is, we do not use any prior biological information, about the brain nor the fMRI signal, etc. This ensures that the diagnostic classifier's performance reflects how well the respective dimensionality reduction method discriminates patients from controls, in the absence of biological information. Once we understand which biologically naive methods work, future work (after this dissertation) will focus on extending these ideas by incorporating biological information.

kPCA-st's result on both datasets is important because it is a proof-of-concept for fMRI-based diagnosis, which will motivate future work into finding new diagnostic criteria based on fMRI. Furthermore, kPCA-st's result shows that fMRI dimensionality reduction processes should be able to discriminate patients from controls, regardless of whether these datasets contain different neuropsychiatric disorders. We use two datasets that contain healthy controls and patients with ADHD [12] and First Episode Psychosis (FEP) [25], respectively, to show that kernel principal component analysis, which uses *non-linear* mappings to reduce over both the spatial and temporal dimensions, performs statistically better than the frequently-used canonical PCA, which uses a linear mapping to reduce over only the temporal dimension. We believe that evaluating the *point-wise* differences of every voxel waveform – i.e., measuring the similarity between each point of each voxel waveform– over all subjects, should reveal differences between patients and controls for any dataset.

[2]We refer to subjects with neuropsychiatric disorder as *patients*.

The dimensionality reduction processes take as input the waveforms of each voxel, over all subjects. The PCA-t (see Figure 2.4) approach runs PCA on the matrix that includes, as rows, the original waveforms of all voxels of all subjects, which compresses each waveform into a smaller number of features, assuming that the waveform of each voxel of each subject is an independent draw; PCA-t has various applications in fMRI analysis [3]. We then learn a classifier that uses the set of all such compressed-waveforms for each subject. Unfortunately on both the ADHD and FEP datasets, PCA-t failed to perform statistically better[3] than the baseline when using only the imaging data. PCA-st (see Figure 2.5) tries to address the poor performance by treating the concatenation of the waveforms of all voxels for each subject, as independent observations. However, PCA-st performed similarly to PCA-t, suggesting that both were limited, perhaps because they both use linear compressions. We therefore applied kernel principal component analysis (kPCA-st) to introduce *non-linear compression* of the data. In general, on both the ADHD and FEP datasets, we show that kPCA-st's compression improves discrimination of patients from controls at a statistically better level than PCA-t.

Section 2 outlines the processing of subjects' raw fMRI data and overviews the methods used in our study. Sections 3 and 4 describe the results on the ADHD200[4] the and FEP datasets respectively, and Section 5 discusses potential future dimensionality reduction processes for fMRI.

[3]Throughout, we say one approach is "statistically better" than another if the paired t-test produces a confidence $p \leq 0.05$.

[4]We refer to ADHD dataset as the ADHD200 dataset.

Chapter 2

Foundations

This section presents the overall process of the diagnostic system, based on Figure 1.1. We first present the "fMRI Image Pipeline" in Subsection 2.1 before distinguishing the fMRI in the ADHD200 and FEP datasets in Subsection 2.2. We then introduce and discuss the use of the General Linear Model (GLM) on the FEP dataset. The remaining subsections summarize the "Feature Creation/Selection" step, as shown in Figure 2.1; Subsection 2.4 discusses Averaging; then Subsection 2.5 discusses Signal Normalization; Subsection 2.6 uses Fast Fourier Transforms (FFT) to process waveforms for the ADHD200 subjects; Subsections 2.7 and 2.8 present long discussions on the PCA variants and Kernel Methods, including the kernelized PCA variant, kPCA-*st*, respectively. We considered various learning algorithms but chose (linear) SVM, as this is a standard learning algorithm.

2.1 fMRI Image pipeline

Both the ADHD200 and FEP dataset's raw fMRI images were preprocessed using SPM8 [15] and personally-developed MATLAB code (as specified below). For each subject, the preprocessing pipeline involved:

1. 6 parameter rigid body motion correction (SPM8)
2. Co-registering functional scans to subject's respective anatomical scan (SPM8)
3. Spatially warping (non-linear, performed estimation and interpolation) anatomical volume to MNI T1 [11] template space at $1 \times 1 \times 1$ mm resolution (SPM8)
4. Interpolating fMRI volumes into T1 template space at $3 \times 3 \times 3$ mm spatial resolution using the same warping parameters computed in the previous step.
5. Applying 8mm full width at half maximum (FWHM) Gaussian spatial filter to fMRI volumes (SPM8).
6. Truncating all resting-state fMRI scanning data to 135 second duration (as this is the shortest time used in all hospitals), then linearly interpolating this data to a sampling rate of 2Hz. This step is exclusive to the

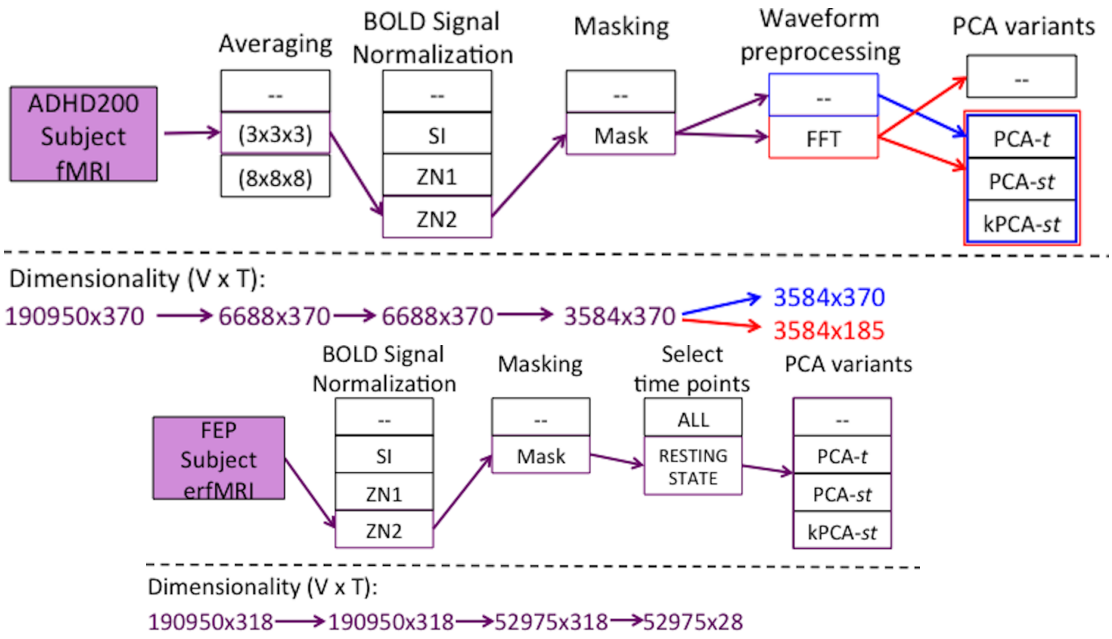

Figure 2.1: Flow of the "Feature Creation/Selection" stage of our diagnostic system for the ADHD200 (top) and FEP datasets (bottom), where the red and blue lines denote whether ADHD200 subjects' waveforms were processed with FFT or were not processed at all. For both datasets, the PCA variant step involves developing the respective variant's data matrix and processing it by one of PCA-*t* (shown in Figure 2.4), PCA-*st* or kPCA-*st* (Figure 2.5). The dimensionality is given beneath each step for the respective datasets, where the number of rows and columns represent the spatial and temporal dimensions, respectively.

ADHD200 dataset.

Afterwards, the subjects' data were aligned to MNI T1 template space. Both the ADHD200 and FEP subjects had the same spatial dimensions ($57 \times 67 \times 50$ voxels), but differed in both their sampling rates (2s and 3s volume times temporally for the ADHD200 and FEP subjects, respectively) and their temporal dimensions (370 and 318 time points for the ADHD200 and FEP subjects' voxel waveforms, respectively).

2.2 Block design fMRI and resting-state fMRI

Aside from the neurological differences between ADHD and FEP, the distinguishing property between the FEP and ADHD200 datasets is that the *FEP dataset consists of block design fMRI* [20], whereas the ADHD200 dataset consists of *resting-state* fMRI; see Figure 2.2. Resting-state fMRI does not involve any overt task or sensory stimulation, as the subject is asked to quietly rest during the scan. In block design fMRI, each event is associated with a *task block*; for the duration of each task block, subjects perform a task that involves functions associated with brain regions that are believed to be affected by FEP.

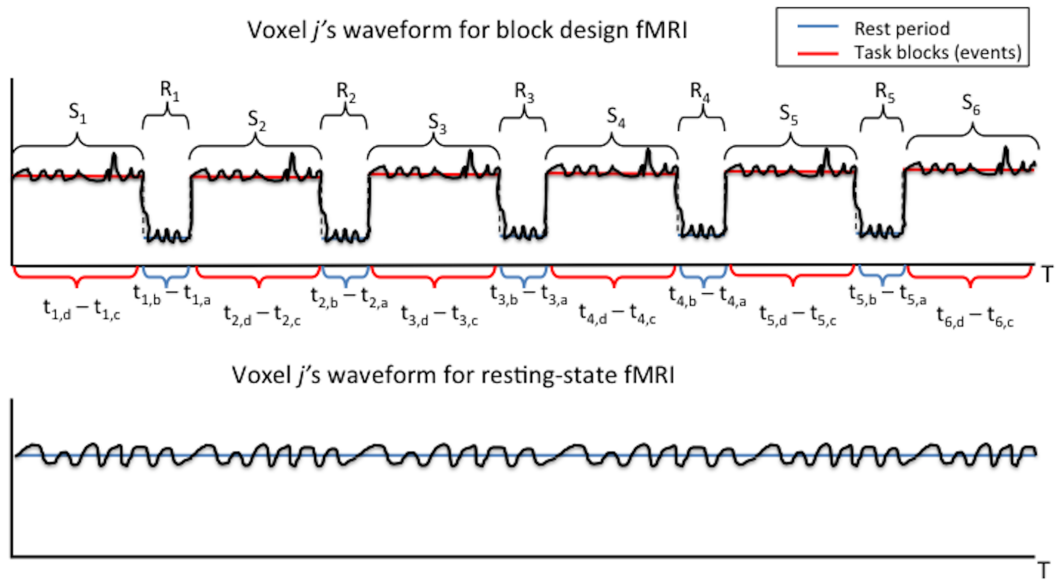

Figure 2.2: Illustrating a hypothetical difference between a voxel's waveform for block design fMRI (top) and resting-state fMRI (bottom), assuming that this voxel is involved in all of the block design fMRI's task blocks. R_i denotes the i^{th} rest period, and S_i denotes the i^{th} task block.

The large difference between the experimental designs of the ADHD200 and FEP data manifests as differences in the voxel waveforms generated by the two designs: the FEP dataset's voxel waveforms are composed of activations during and between[1] events. Since the events evoke activation in certain brain regions associated with FEP, some of these regions will exhibit differences in activation between FEP patients and controls.

2.3 General Linear Model (GLM)

Earlier, we mentioned that voxel waveforms consist of many time points, where each time point represents the BOLD-signal intensity. The General Linear Model (GLM) explains the voxel's BOLD-signal activation level at time t in terms of the *explanatory variables* that are related to the conditions under which volume t was collected [14]. The j^{th} voxel in volume t is expressed as

$$x_{t,j} = g_{t,1}\beta_{1,j} + \ldots + g_{t,N}\beta_{N,j} + \epsilon_{t,j} \tag{2.1}$$

where $[g_{t,1}, \ldots, g_{t,N}]$ is an N-dimensional row vector containing the N explanatory variables believed to affect the BOLD-signal activation levels of *all* voxels in the volume collected at time t, $[\beta_{1,j}, \ldots, \beta_{N,j}]^T$ are the N parameters relating voxel j's BOLD-signal activation level at time t to the N explanatory variables, and $\epsilon_{t,j} \sim \mathcal{N}(0, \sigma_j^2)$ is the independently and identically distributed (i.i.d) error for voxel j at time t from the standard normal distribution. Equation 2.2

[1]The time between events is referred to as a *rest* period.

shows that the GLM models *every* voxel at *every* time point separately.

Let $\mathbf{X}$ be a $T \times V$ matrix representing a subject's fMRI scan that consists of T volumes, where each volume contains V voxels. Define $\mathbf{G}$ as the $T \times N$ *design matrix*, where each column represents one condition over T time points, and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 | \ldots | \boldsymbol{\beta}_V]$ as the $N \times V$ parameter matrix, where column j contains the N parameters $\boldsymbol{\beta}_j$ for voxel j. The matrix form for Equation 2.2 is given by

$$\mathbf{X} = \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

where $\boldsymbol{\epsilon}$ is the $T \times V$ error matrix containing the i.i.d noise from the standard normal distribution.

The GLM calculates the Ordinary Least Squares (OLS) estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, which is an unbiased estimator uniquely determined by

$$\hat{\boldsymbol{\beta}} = (\mathbf{G}^\intercal \mathbf{G})^{-1} \mathbf{G}^\intercal \mathbf{X} \tag{2.3}$$

With this framework established, the remaining question is how to construct the design matrix $\mathbf{G}$ such that it is representative of the experimental design. Such a matrix $\mathbf{G}$ contains column vectors that model interesting effects, such as the activation pattern for each event in a block design fMRI study, and uninteresting effects, such as motion, scanner noise, and other effects that might impact the quality of an fMRI scan.

The 318×34 design matrix used in for the FEP dataset models 2 interesting and 32 uninteresting effects, where each interesting effect is a column containing indicator variables that model the voxels' BOLD-signal activation pattern over all 318 time points for the respective event type[2]. We use these events' indicator variables to:

I. Produce the 2 rows of predictor variables in $\hat{\boldsymbol{\beta}}$ that relate the voxels' BOLD-signal intensity to the respective event's activation pattern.

II. Select the resting-period time points from each FEP subject. This was done by selecting the time points where both events' indicator variables were zero, which gave a total of 28 resting state time points for each subject.

In the first case, we give the $2 * 57 * 67 * 50$ parameters, which relate the voxels' BOLD-signal intensity to the events' activation pattern, as input to the learner to determine if this relationship is diagnostic. It follows that these parameters reduce the temporal dimensionality of the FEP subjects' fMRI data.

In the second case, we acknowledge that, for block design fMRI, the BOLD-signal is returning to the baseline activation from the previous block-related activation level during the rest periods, which means that the FEP subjects' rest periods are different from the ADHD subjects' resting-state fMRI. However, the indicator variables for each event type are zero *only* for the last 2

[2]There are two types of events in the FEP subjects' fMRI.

time points of every rest period, which suggests that these time points are *after* the BOLD-signal returns to its baseline activation level. We therefore believe that using these time points for FEP subjects' fMRI homogenizes our comparisons across datasets as much as possible.

2.4 Averaging

We describe each subject using $V = L \times W \times H$ spatial voxels, each with a waveform of length T, meaning each subject scan has dimensionality $L \times W \times H \times T$; here, this corresponds to $57 \times 67 \times 50 \times 370$ and $57 \times 67 \times 50 \times 28$ real values for the ADHD200 and FEP datasets, respectively.

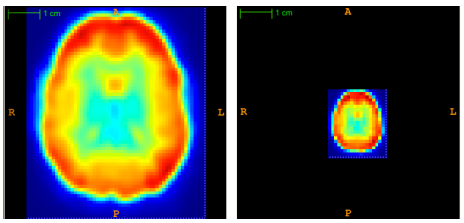

Figure 2.3: Temporal view of subject's original (left) and averaged (right) volume at t=1. In 2-D, this produced a factor of 3x3 compression.

Hence, each subject image in the ADHD200 dataset requires roughly 282MB of memory to hold this single-precision, four dimensional matrix. Applying PCA (Subsection 2.7) to the ADHD200 dataset, which consists of 668 subjects, would require 188.78GB in memory, which strains most computers, both directly and indirectly (by thrashing).

Given the high dimensionality of the data, we first reduced the ADHD200 subjects' spatial dimensionality by representing each $k \times k \times k$ subvolume by its average, for each time $t = 1, \ldots, T$, which reduces the data size by a factor of k^3. Figure 2.3 illustrates the subject fMRI before and after averaging.

All subjects' fMRI data in the ADHD200 dataset were averaged by taking the mean over $3 \times 3 \times 3$ subvolumes, resulting in $\lfloor 57/3 \rfloor \times \lfloor 67/3 \rfloor \times \lfloor 50/3 \rfloor = 19 \times 22 \times 16 = 6688$ voxel waveforms per subject. We considered averaging the FEP subjects' data over $3 \times 3 \times 3$ subvolumes, but observed a decrease in performance (results not shown), and we therefore did not average the FEP subjects' fMRI.

2.5 BOLD-signal Normalization and Masking

After preprocessing the ADHD200 and FEP datasets, and only averaging the ADHD200 subjects, we use signal normalization to normalize differences between waveform magnitudes that can arise from either scanner differences or image registration.

We considered three different signal normalization methods. Two of these methods [8] normalized voxel waveforms according to their mean, $\mu_{\mathbf{x}_i}$, and/or

standard deviation, $\sigma_{\mathbf{x}_i}$, for waveform $\mathbf{x}_i$ associated with a single voxel. Here, for each $i = 1, \ldots, V$:

Percent Signal Change:

$$\mathbf{x}_i = \frac{\mathbf{x}_i - \mu_{\mathbf{x}_i}}{\mu_{\mathbf{x}_i}} \times 100$$

Z-Score Normalization 1:

$$\mathbf{x}_i = \frac{\mathbf{x}_i - \mu_{\mathbf{x}_i}}{\sigma_{\mathbf{x}_i}}$$

The classification accuracies when using these methods were no better than the baseline for both the ADHD200, which considers seven dimensionality reduction processes (PCA Variants only, FFT only, FFT then PCA Variants), and FEP datasets, which considers three dimensionality reduction processes (PCA Variants only). This suggested that normalizing waveforms according to their *local* properties – i.e., the voxel waveform's mean and/or standard deviation – was not useful for discriminating patients from controls.

We tried a third method: normalizing voxel waveform values by using the global mean, $\mu_{\mathbf{x}}$, and global standard deviation, $\sigma_{\mathbf{x}}$, over the waveforms from the *entire* fMRI scan for that subject:

Z-Score Normalization 2 (ZN2) Each subject's image intensities were set to the z-scores computed using mean and standard deviation over the *entire* image:

$$\mathbf{x} = \frac{\mathbf{x} - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}}$$

After performing ZN2 signal normalization, we identified voxels inside the brain for the ADHD200 and FEP datasets using $19 \times 22 \times 16$ and $57 \times 67 \times 50$ volumes, respectively, and applied a mask that removed voxels outside of the brain. This left 3584 and 52975 voxel waveforms for every subject in the ADHD200 and FEP datasets, respectively.

2.6 Fourier Transforms[3]

Here, we view each voxel waveform as an observation, with T time points for each waveform. An FFT transforms these voxel waveforms from the time domain to the frequency space, and produces T Fourier Components, which allow us to consider the magnitude of the complex-valued Fourier coefficients (i.e. "amount" of signal) corresponding to each specific range of frequencies.

In many situations, the amount of signal that lies in a few specific frequency bands may distinguish one class of signals from another. We considered using only a pre-determined subset of the frequencies (called "bandpass filtering"), however we found that the resulting accuracies (for the biologically motivated bands we considered) were well-below the baseline (results not shown).

[3]FFT is only used on the ADHD200 dataset.

After performing FFT on each ADHD200 subject's 3584 waveforms, we selected only the first 185 components because the remaining components are reflections of the first half about the Nyquist frequency. We extracted the magnitude and discarded the phase for each frequency component, thus each subject was described using $3584 \times 185 = 663{,}040$ features. As shown in Figure 2, we can use these values as input to the next step – either a PCA variant, or the learner itself.

We did not apply FFT to the FEP dataset after extracting the 28 resting-state time points because the collection of these time points cannot be viewed as a time series or waveform, as they are not contiguous.

2.7 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that computes the linear combination of features – e.g. the intensities of a specific voxel at a single time, over the set of data points – that have high variance [26]. Instead of representing a data point using the original features, we can "project" those original features onto a smaller number of principal components, and know that this re-encoding "captures" a large proportion of variance in the data. We first introduce *linear similarity measures*, which are used in Subsection 2.7.2 to illustrate the relevant theoretical properties of PCA.

2.7.1 Linear Similarity Measures

We want the learners to produce classifiers that can *generalize* to unseen data points. Assuming that there are c classes in the data, when a new data point x_{N+1} is provided, the classifier will assign x_{N+1} a label belonging to one of these c classes. The class assigned by the learner for data point x_{N+1} should contain data points similar to x_{N+1}, based on some notion of *similarity*. Before defining such a similarity measure, Schölkopf and Smola [30] formalize the problem setting as follows:

Let $\mathcal{X}$ be the nonempty set containing the empirical data, and $\{1, \ldots, c\}$ be the set containing the class labels. Here the i^{th} data point with the associated label is $(x_i, y_i) \in \mathcal{X} \times \{1, \ldots, c\}$ for $i = 1, \ldots, N$. We assume that $\mathcal{X} = \mathbb{R}^p$ – i.e., data points are p-dimensional. In general, we use

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \tag{2.4}$$

for some similarity function k that outputs a *real value* that characterizes the similarity between a pair of data points. Here, we can use the *dot (inner) product*, also called the linear kernel, as a similarity function

$$k_{linear}(x_i, x_j) = \langle x_i, x_j \rangle = \sum_{\ell=1}^{p} x_{i,\ell} x_{j,\ell} \tag{2.5}$$

where $x_i = [x_{i,1}, \ldots, x_{i,p}]$ and $x_j = [x_{j,1}, \ldots, x_{j,p}]$ are data points in $\mathbb{R}^p$.

Interpreted geometrically, the inner product between two data points x_i and x_j is the cosine of the angle, assuming that both data points are normalized to length 1, where the length, or norm, of a data point x is defined as $||x||_2 = \sqrt{\langle x, x \rangle}$. Note that this dot product is largest when $x_i = x_j$ (i.e., are extremely similar) and is zero when x_i is orthogonal to x_j.

2.7.2 Theory of PCA

Let $\mathbf{X}$ be the $N \times p$ matrix, where each row corresponds to a subject and each column is a feature, where the mean over each feature is zero. We can define a matrix of similarities

$$\begin{aligned} \mathbf{K} &= \mathbf{X}\mathbf{X}^T \\ \mathbf{K}(i,j) = k_{linear}(x_i, x_j) &= \langle x_i, x_j \rangle = \sum_{\ell=1}^{p} x_{i,\ell} x_{j,\ell} \end{aligned} \tag{2.6}$$

where x_i and x_j are rows of $\mathbf{X}$. Then

$$\mathbf{K}\mathbf{e} = \lambda \mathbf{e} \tag{2.7}$$

holds for each eigenvalue/eigenvector pair $(\lambda, \mathbf{e})$, where *eigenvector* $\mathbf{e} \in \mathbb{R}^N$ has the corresponding *eigenvalue* $\lambda \in \mathbb{R}$. Since $\mathbf{K}$ is symmetric, it will always have non-negative eigenvalues $\lambda \geq 0$. Assume that the eigenvalues of $\mathbf{K}$ are sorted in descending order – i.e. $\lambda_i \geq \lambda_{i+1}$. The eigenvectors of a matrix are *orthogonal* to every other eigenvector of this matrix. Note there are at most N non-zero eigenvalues with the corresponding eigenvector matrix $\mathbf{E}_N = [\mathbf{e}_1, \ldots, \mathbf{e}_N] \in \mathbb{R}^{N \times N}$.

Notice that if there many more data points than features – i.e., $N \gg p$ – then $\mathbf{K}$ can be very large. The *dual* of PCA (or the dual trick) provides an easy way to compute the eigenvalue/eigenvector pairs of such a high-dimensional similarity matrix, by using the eigenvalue/eigenvector pairs of the lower-dimensional covariance matrix $\mathbf{S} = \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{p \times p}$ [19]:

$$\mathbf{S}\mathbf{e} = \lambda \mathbf{e} \tag{2.8}$$

$$\begin{aligned} \mathbf{X}(\mathbf{X}^T\mathbf{X})\mathbf{e} &= \mathbf{X}\lambda\mathbf{e} \\ \mathbf{K}(\mathbf{X}\mathbf{e}) &= \lambda(\mathbf{X}\mathbf{e}) \end{aligned} \tag{2.9}$$

which proves that the eigenvalue/eigenvector pairs $(\lambda, \mathbf{e})$, computed using Equation 2.8, of the relatively-small $p \times p$ matrix $\mathbf{S}$, correspond to the eigenvalue/eigenvector pairs $(\lambda, \mathbf{X}\mathbf{e})$ of the much larger $N \times N$ similarity matrix $\mathbf{K}$. That is, each (λ, $\mathbf{e}$) eigenvalue/eigenvector pair of $\mathbf{S}$ corresponds to the eigenvalue/eigenvector pair of $\mathbf{K}$: (λ, $\mathbf{X}\mathbf{e}$). As $N \gg p$, it follows from Equation 2.7 that there are at most p eigenvalue/eigenvector pairs with corresponding eigenvector matrix $\mathbf{E}_p = [\mathbf{e}_1, \ldots, \mathbf{e}_p] \in \mathbb{R}^{N \times p}$.

Note that $\mathbf{E}_p$ can also be viewed the matrix that uses the *low-dimensional* eigenvector matrix $\mathbf{E}_S = [\mathbf{e}_1, \ldots, \mathbf{e}_p] \in \mathbb{R}^{p \times p}$ of $\mathbf{S}$ to redefine the coordinate system of $\mathbf{X}$, which produces the *principal component score matrix*

$$\mathbf{Z}_S = \mathbf{X}\mathbf{E}_S \in \mathbb{R}^{N \times p} \tag{2.10}$$

where rows of $\mathbf{Z}_S$ are the data points in the new coordinate system, and columns are the *scores* of the N data points, where the i^{th} column contains the N scores from the i^{th} eigenvector of $\mathbf{S}$.

Although Equation 2.10 shows that the score matrix is *numerically* equivalent to $\mathbf{E}_p$, *its interpretation depends on the application*: When we use the dual trick, we view $\mathbf{E}_p$ as the eigenvector matrix containing p N-dimensional eigenvectors that allow us to *project* the dataset onto these eigenvectors, which produces the principal component matrix

$$\mathbf{Z} = \mathbf{E}_p^T\mathbf{X} \in \mathbb{R}^{p \times p} \tag{2.11}$$

where we omit this step when interpreting $\mathbf{E}_p$ as the score matrix.

The i^{th} principal component is viewed as the data matrix *projected* onto (that is, multiplied by) the eigenvector $\mathbf{e}_i$. If there are a total of N principal components, the proportion of variance captured in the dataset by the first (largest) $m < N$ principal components is given by:

$$\text{proportion of variance}(\lambda_1, ..., \lambda_\text{m}) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{N} \lambda_j} \tag{2.12}$$

For the ADHD200 dataset, we select the m for each dimensionality reduction process as the number of components needed to capture over 99% of the variance. For the FEP dataset, m is defined as the number of components that capture 98% of the variance. We provide our reason for the selection of m on the FEP dataset in Section 4.2.

2.7.3 PCA-*t*

Here we describe PCA-t, which is a standard approach to reducing fMRI dimensionality; its purpose is to capture the variance over waveforms by selecting the top m components – i.e., the ones that are responsible for the largest proportion of the variance over voxel waveforms. It is believed that PCA-t's principal components may represent regular activation patterns across voxels [18].

Andersen *et al.* use primates' fMRI data to show that PCA-t's largest principal components captured the systematic structure – i.e. voxel activation patterns – while relegating effects of random noise to the smaller T - m principal components. However, Andersen *et al.* also state that determining the usefulness of PCA-t's projections is *largely subjective* [3], presumably because their article *introduces* the different ways of applying PCA to fMRI. Note

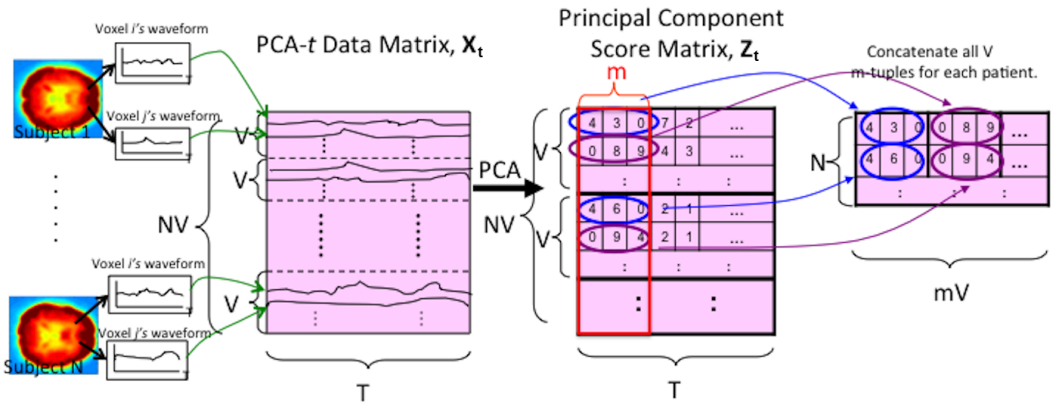

Figure 2.4: The PCA-t process: The subject's voxel waveforms are rows in the data matrix $\mathbf{X}_t$ used by PCA-t. After PCA, we then extract the first m principal components scores for all of the subject's voxel waveforms, and use these as the reduced-dimension imaging features. For some studies (Figure 2.1), the "waveform" used to form the PCA-t matrix is FFT of those waveforms.

that our supervised learning framework provides an *objective* way to evaluate various dimensionality reduction techniques, based on the performance "down-stream", of the resulting classifier. We use this supervised learning framework to evaluate whether PCA-t the systematic structure, captured by PCA-t's largest principal components, allows us to learn a classifier that discriminates ADHD patients from controls.

As shown in Figure 2.4 PCA-t takes, as input, a data matrix, $\mathbf{X}_t$, that treats the waveforms of each voxel of each subject as a data point, to produce a matrix $\mathbf{Z}_t$ whose principal components capture over 99% of the variance in $\mathbf{X}_t$ (Equation 2.12). Representing every subject's waveforms as data points involved reshaping their averaged, BOLD-signal normalized, and masked 4-D fMRI into 2-D (done by vectorizing the spatial dimensions), which produces a $V \times T$ matrix for each subject, followed by vertically concatenating all subjects' $V \times T$ matrices, resulting in an $NV \times T$ data matrix, $\mathbf{X}_t$.

Using Equation 2.10, $\mathbf{X}_t$ is projected onto the low-dimensional eigenvector matrix $\mathbf{E} \in \mathbb{R}^{T \times m}$ to produce the $NV \times m$ principal component scores matrix $\mathbf{Z}_t$, where the m NV-dimensional principal components (columns) contain the N subjects' principal component scores. Note: If we use the dual trick to recover the T NV-dimensional eigenvectors of $\mathbf{K}_t$, then Equation 2.11 would produce a $T \times T$ principal component matrix, where each principal component is a dot product between the volumes of *all* subjects at one of the T time points, and one of the T NV-dimensional eigenvectors of $\mathbf{K}_t$. Our empirical studies (not shown) demonstrate that this principal component matrix does not assist in our goal of developing an effective diagnostic classifier.

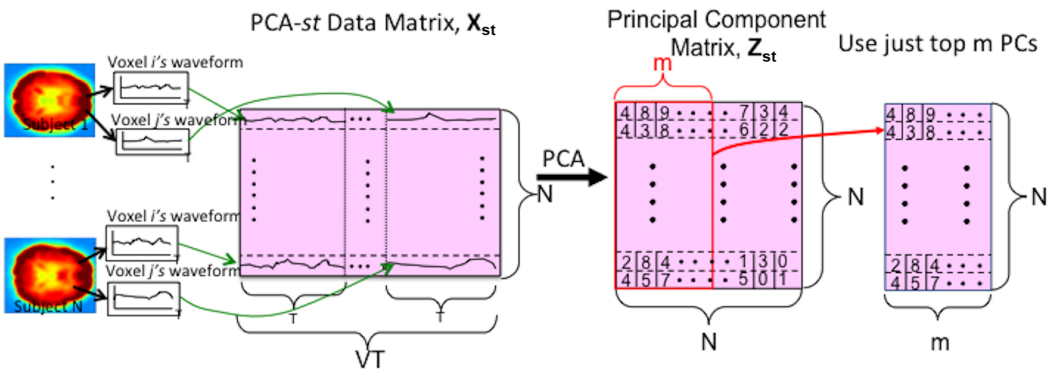

Figure 2.5: This figure shows how the fMRI spectra are assembled into the data matrix $\mathbf{X}_{st}$ used by PCA-*st* / kPCA-*st*. For some studies (Figure 2.1), the "waveform" used to form the PCA-*st* matrix is the Fourier Transform of those waveforms.

2.7.4 PCA-*st*

Each voxel waveform's BOLD signal is confounded by noise from various sources [4], which can impact the pattern of voxel waveforms. We believe that the *point-wise* comparison of V waveforms over all subjects, represented as a covariance matrix between VT points, can provide eigenvalue/eigenvector pairs, and consequently principal components, that are more informative about differences between patients and controls than the principal components produced using the eigenvalue/eigenvector pairs of $\mathbf{S}_t$, as the former, basically, relates the waveforms at a spatial location only to the other waveforms at this location, across all subjects, while the latter relates the volumes at *each* time point (over *all* subjects) to each other. This motivates our formulation of PCA-*st*'s data matrix, where each subject's averaged, BOLD-signal normalized and masked fMRI were represented as rows, by horizontally concatenating all of the respective subject's voxel waveforms into a row vector, resulting in an $N \times VT$ data matrix, $\mathbf{X}_{st}$.

As shown in Figure 2.5, PCA-*st* takes as input a data matrix $\mathbf{X}_{st}$ that contains each subject's (vectorized) fMRI as a data point, to produce a matrix $\mathbf{Z}_{st}$, whose principal components capture over 99% of the variance over the waveforms (Equation 2.12).

PCA-*st* uses the dual trick to recover the high-dimensional eigenvectors of the covariance matrix $\mathbf{S}_{st}$ from the low-dimensional eigenvectors of $\mathbf{K}_{st}$ (Equation 2.9). We then use Equation 2.11 to project $\mathbf{X}_{st}$ onto the $VT \times N$ eigenvector matrix; this produces an $N \times N$ principal component matrix $\mathbf{Z}_{st}$ that contains N principal components for every subject, where each principal component is the dot product between the respective subject's VT-dimensional vectorized imaging data, and one of the N VT-dimensional eigenvectors of $\mathbf{S}_{st}$. We only use the top m principal components as features for each subject.

2.8 Kernel Methods

Non-linear Similarity Measures

Again, we follow Schölkopf and Smola [30] to show that a *non-linear* similarity measure can still exist in an inner product space, and that the different values produced by using a nonlinear similarity measure (in comparison to the values produced by the linear similarity measure) may improve the discrimination of patients from controls.

In order to use the inner product as a measure of similarity, the data points are mapped into vectors in an inner product space, $\mathcal{H}$:

$$\Phi : \mathcal{X} \to \mathcal{H} \tag{2.13}$$

Φ's embedding into $\mathcal{H}$ allows the introduction of different similarity measures that can be expressed as inner products:

$$k_\Phi(x_i, x_j) := \langle x_i, x_j \rangle_\Phi = \langle \Phi(x_i), \Phi(x_j) \rangle \tag{2.14}$$

Mapping subject images to an inner product space through $\Phi(\cdot)$ allows us to investigate whether the enriched geometric relationship, afforded by the mapping $\Phi(\cdot)$, can improve the discriminatory power of a classifier.

For the *Radial Basis Function* (RBF) kernel, defined below, the range of the $\Phi(x)$ mapping is infinite dimensional; this is also a non-linear transform, which is difficult, if not impossible, to explicitly represent [17]. The *Kernel Trick efficiently* computes the kernel function, by avoiding explicitly mapping the data point x to the higher-dimensional $\Phi(x)$, and so can compute each kernel value, $k(x_i, x_j)$, in a way that depends on the dimensionality of the (original) data points. The RBF kernel value of two data points x_i and x_j, in our original space, is

$$\tilde{\mathbf{K}}_{RBF}(i,j) = k_{RBF}(x_i, x_j) = \exp\left(\frac{-\langle (x_i - x_j), (x_i - x_j) \rangle}{2\sigma^2}\right) \tag{2.15}$$

where x_i and x_j represent two subjects' voxel waveforms viewed as a vector, and σ is a user-determined parameter. Note that the matrix containing the kernel values between *all* pairs of the data points is denoted as $\tilde{\mathbf{K}}$ where, for the remainder of this article, the subscript (eg, the *RBF* of $\mathbf{K}_{RBF}$) denotes the mapping.

When performing kPCA-*st* (Section 2.8.1), we use the RBF Kernel because the kernel matrix, $\tilde{\mathbf{K}}_{RBF}$, is *strictly positive definite*, which guarantees the recovery of N strictly positive eigenvalues. Therefore we will always have N principal components to capture the variance in the data; having fewer than N principal components could hurt the result, as there would be fewer eigenvectors to capture the variance in the data.

2.8.1 kernel Principal Component Analysis (kPCA-*st*)[4]

kPCA-*st* consists of four basic steps:

1. Compute kernel matrix $\tilde{\mathbf{K}}_{RBF} \in \mathbb{R}^{N \times N}$ over the training data using Equation 2.15.
2. Center $\tilde{\mathbf{K}}_{RBF}$ using $\mathbf{K}_{RBF} = \tilde{\mathbf{K}}_{RBF} - \mathbf{1}_N\tilde{\mathbf{K}}_{RBF} - \tilde{\mathbf{K}}_{RBF}\mathbf{1}_N + \mathbf{1}_N\tilde{\mathbf{K}}_{RBF}\mathbf{1}_N$, where $\mathbf{1}_N$ is an $N \times N$ dimensional matrix of ones.
3. Compute the eigenvector matrix $\mathbf{E}_{RBF}$ of $\mathbf{K}_{RBF}$.
4. Project $\mathbf{K}_{RBF}$ onto the eigenvector matrix using Equation 2.11, where $\mathbf{X}_{st}$ and $\mathbf{E}$ are replaced with $\mathbf{K}_{RBF}$ and $\mathbf{E}_{RBF}$, respectively.

Figure 2.5 illustrates the process for kPCA-*st*, assuming PCA is replaced with kPCA. Similar to PCA-*st*, kPCA-*st* produces N reduced imaging features for each subject.

We emphasize that *every element in kPCA-st's kernel matrix represents the point-wise similarity between two subjects' entire fMRI image (after applying mapping Φ) because it measures similarity of all T points for every voxel location.*

Using kPCA-*st* as a dimensionality reduction process for fMRI is appealing because (1) the result is different from using the standard linear methods, and therefore might be good, and if so (2) the kernel matrix, $\mathbf{K}_{RBF}$, *grows with the square of the data points, instead of the dimensionality of the range of* Φ. All of kPCA-*st*'s reported results use the RBF kernel with the kernel parameter $\sigma = 150$.

2.8.2 kPCA in only spatial or only temporal domains

Applying kPCA-*st* to fMRI data non-linearly reduces dimensionality over *both* spatial and temporal dimensions. Some might consider applying kPCA in *only* the spatial or temporal dimensions. In the following paragraphs, we show that kPCA in only the spatial or temporal dimensions defeats the purpose of using kPCA over canonical PCA.

Performing kPCA over the temporal dimension, using PCA-*t*'s data matrix $\mathbf{X}_t$, produces an $NV \times NV$ kernel matrix, which faces the same computational issues as $\mathbf{K}_t$ because of its large size. If we average over larger subvolumes to reduce the computational strain of calculating the kernel matrix, the loss of spatial information can cripple performance: We considered averaging over $8 \times 8 \times 8$ subvolumes (result not shown), but observed poor performance in comparison to averaging over $3 \times 3 \times 3$ subvolumes, which supports our claim. When averaging over $8 \times 8 \times 8$ subvolumes, we believe that the large size of the kernel matrix *and* the poor result of kPCA-*st* are sufficient to dismiss applying kPCA in the temporal dimension.

If kPCA-*st* is applied to the *spatial* matrix – i.e. subject volumes at a single time point – then the data matrix $\mathbf{X}$ given as input to kPCA-*st* has di-

[4] A full theoretical description is provided in Appendix C.

mensionality $N \times V$. Given that there are T resting-state volumes per subject, it is difficult to justify selection of a specific time point's volume because all volumes were collected under the same conditions.

Chapter 3

ADHD200 Dataset

This section deals with the ADHD200 data; see Section 4 for the FEP dataset.

Attention Deficit Hyperactivity Disorder (ADHD) is a disability that befalls an estimated 2-9% [24] of school-age children in the United States, often leading to substantial lifelong impairment for these children. Assuming a prevalence rate of 5%, the indirect cost associated with ADHD in the United States exceeds over 36 billion dollars annually [24]. While ADHD is listed as a disorder in the DSM, its underlying neurobiology is not thoroughly understood.

3.1 Methodology

We only used 668 of the 776 subject scans, as we removed the 108 scans that either failed image registration (in our fMRI pipeline), or their respective hospital's fMRI quality check (given as a binary value for every subject). This 668-subject dataset, summarized in Table 3.1, contained the age, gender, handedness, IQ scores, and the scanning site, as well as the resting-state fMRI scans for 429 healthy controls, 141 ADHD-combined (ADHD-1), and 98 ADHD-inattentive subtypes (ADHD-3), primarily over adolescents and some children and young adults[1]. We evaluated the dimensionality reduction processes over two settings, depending on whether the label ranged over two classes (ADHD [both subtypes] vs control) or three classes (ADHD-1 vs ADHD-3 vs control).

For each setting, we grouped the dimensionality reduction processes into *three* categories, depending on how each process transformed the averaged, BOLD-signal normalized, and masked original waveforms (called "waveform" below; see Figure 2.1). The number in parentheses in each category is the number of inputs associated with this class of dimensionality reduction processes.

For the first category, we gave each subject's waveforms as:

I. PCA Variant only input to PCA-*t*, PCA-*st* and kPCA-*st*. (3)

[1]For readers with a clinical background, Appendix A.1 provides additional information about the ADHD dataset.

For the two remaining categories, we first compute the FFT for each subject's waveforms, then use the magnitudes from each subject's FFTed waveforms as:

II. FFT Only the reduced-dimensionality imaging features. (1)

III. FFT then PCA Variant input to the PCA variants. We delineate these processes from those in the first category by referring to these variants as FFT+PCA-*t*, FFT+PCA-*st*, and FFT+kPCA-*st*. (3)

which gives a total of *seven* dimensionality reduction processes on the ADHD200 dataset.

We used three feature sets for our evaluation:

Phenotypic Data consisted of subject age, gender, scanning site and three IQ scores: Verbal, Performance, and the Full4 IQ. For each type of IQ, we replaced any missing value with the average, over all subjects who had a non-missing value for that type of IQ.

Imaging Data contained only the reduced-dimensionality imaging features, produced by one of the 7 dimensionality reduction processes.

Imaging and Phenotypic Data appended the phenotypic data to the imaging data returned by one of the 7 respective dimensionality reduction processes.[2]

which gives a total of 15 feature sets for each setting: The "Imaging Data" and "Imaging and Phenotypic Data" feature sets *for each of the seven dimensionality reduction processes*, and the phenotypic feature set.

We give each of the resulting feature sets as input to a linear kernel Support Vector Machine (SVM)[3], which produces a classifier. We then evaluate the feature sets in terms of the accuracy of the resulting classifier; this accuracy is based on 10-fold Cross Validation [16], where we use the same folds throughout.

We compared every method's performance to the baseline simply to determine whether its results were statistically better than simply guessing the majority class. We also compare kPCA-*st*'s results to PCA-*t*'s; each time kPCA-*st* produces a statistically better accuracy than PCA-*t* supports our hypothesis that kPCA-*st* is superior to canonical PCA for dimensionality reduction. To further test the results – i.e. to make sure they were not a consequence of overfitting – we used a holdout set to evaluate the performance of each method. At the close of the competition, the most accurate classifier was based on only the phenotype data [13]. Therefore, we also compare our results to that system.

We discuss the performance of all dimensionality reduction processes on the original dataset, followed by the holdout set (Sections 3.2 and 3.3). For each of the original and holdout datasets, we initially discuss the performance of the PCA variants without using FFT (Sections 3.2.1 and 3.3.1); we then

[2]This feature set allows us to determine whether the combination of the imaging and phenotypic feature will have a *synergistic* effect.

[3]This is a standard learning algorithm. We considered other learners, but found none worked better.

Hospital	# of Subjects	Control Group	ADHD Combined	ADHD Inattentive
Kennedy Kriegler Institute (KKI)	78	58	15	5
NeuroIMAGE	38	16	12	0
Peking University	194	116	29	49
Oregon Health Science University (OHSU)	64	36	17	11
New York University (NYU)	188	91	64	33
University of Pittsburgh	66	66	0	0
Washington University	40	40	0	0

Hospital	# of Subjects	Control Group	ADHD Combined	ADHD Inattentive
Kennedy Kriegler Institute (KKI)	11	8	3	0
NeuroIMAGE	25	14	11	0
Peking University	51	27	10	14
Oregon Health Science University (OHSU)	34	28	5	1
New York University (NYU)	41	12	22	7
University of Pittsburgh	9	5	0	4
Washington University	0	0	0	0

Table 3.1: Distribution of ADHD patients and control subjects contained in the original (left) and holdout (right) datasets.

discuss the performance when using the magnitudes of the FFTed waveforms' frequency components to learn a classifier (Sections 3.2.2 and 3.3.2), and conclude each dataset's results subsection by discussing the performance using these magnitudes as input to the PCA variants (Sections 3.2.2 and 3.3.2).

For each of these cases, we first discuss the results in the two-class setting before the three-class setting, where in each setting we first consider using only the (reduced) imaging features without the phenotypic data and then consider the combination of these (reduced) imaging features with the phenotypic data. Section 3.4 discusses some of the issues we encountered when working with the ADHD200 data.

3.2 Original Dataset Results

3.2.1 PCA Variants Only

For PCA-t, we only use the m=2 largest components. For PCA-st[4] and kPCA-st we use the m=667 and m=668 largest principal components.

When using only the reduced imaging features from PCA-t or PCA-st in the two-class setting, the accuracies were not statistically better than the baseline. In contrast, using only the reduced imaging features from kPCA-st produced an accuracy of 70.3% that is statistically better than the baseline, PCA-t and PCA-st (p=2.41e-3, p=1.44e-2, p=2.22e-2), but was not statistically better than using only the phenotypic data.

When combining the imaging and phenotypic features in the two class setting, each of PCA-t, PCA-st and kPCA-st produced classification accuracies that is statistically better than the baseline (p=1.28e-2, p=1.59e-2 and p=3.18e-4). Even though PCA-t and PCA-st's classification accuracies were statistically better than the baseline when using both imaging and phenotypic data in the two-class setting, neither surpassed the phenotypic classification accuracy. This result suggested that, in the two-class setting, using only the

[4]The smallest eigenvalue was zero for PCA-st.

# of classes	Baseline	Phenotypic Only	FFTed Waveforms	PCA Variant	Imaging Only	Imaging & Phenotypic data
2	64.22	72.9	-	PCA-t	65.69 (7.16)	**70.51** (6.91)
			-	PCA-st	65.57 (5.51)	**69.89** (6.46)
			-	kPCA-st	**70.35***(5.21)	**73.20** (4.79)
			+	–	**68.41** (5.50)	**70.95** (7.66)
			+	PCA-t	**69.60** (5.36)	**70.06** (4.83)
			+	PCA-st	**69.30** (5.82)	**70.06** (5.08)
			+	kPCA-st	**68.70** (5.53)	**76.04*** (4.92)
3	64.22	66.77	-	PCA-t	58.82 (6.26)	62.86 (6.55)
			-	PCA-st	59.82 (6.16)	63.30 (6.55)
			-	kPCA-st	64.06* (3.74)	66.0*(7.56)
			+	–	63.92* (5.88)	64.06 (4.37)
			+	PCA-t	59.56 (5.87)	61.23 (5.19)
			+	PCA-st	60.76 (5.17)	61.07 (4.97)
			+	kPCA-st	64.36*(5.19)	**68.55***(6.61)

Table 3.2: Two and three-class classification accuracies on the original, 668 subject, dataset, where the reduced features were obtained by performing FFT and/or PCA-*t*, PCA-*st* and kPCA-*st* on the imaging data. Standard deviations are provided in parenthesis; an accuracy is in **bold** if it is statistically better than baseline, and is asterisked (*) if it is statistically better than PCA-*t* in the same setting.

phenotypic data was better than combining it with the reduced imaging features from PCA-*st* and PCA-*t*. In contrast, kPCA-*st* produced an accuracy that was superior, but not statistically better (p=0.88), than the phenotypic data. In general for the two-class setting, kPCA-*st*'s results suggest that the reduced imaging features improved the discriminatory power of the classifier better than PCA-*t* and PCA-*st*.

In the three-class setting, PCA-*t*, PCA-*st* and kPCA-*st* failed to produce classification accuracies that were statistically better than the baseline, using either the combination of imaging or phenotypic data or only the imaging data. While all of the processes performed below the baseline when using only the imaging data in the three-class setting, kPCA-*st* performed statistically better than PCA-*t* and PCA-*st* (p =1.8e-2 and p=1.8e-2). When combining the imaging and phenotypic data in the three-class setting, kPCA-*st* produced accuracies that were not statistically better than baseline (p=3.8e-1), PCA-*t* (p=6.3e-2), or PCA-*st* (p=1e-1).

Combining the phenotypic data with kPCA-*st*'s reduced imaging features in the three class setting seems to have a *synergistic* effect because using the phenotypic or reduced-imaging features separately fail to outperform the baseline, but their combination approaches a result that was statistically better result than the baseline. This suggested that using only phenotypic data was insufficient to distinguish between ADHD subtypes.

3.2.2 FFT Only

In both the two and the three class settings, the FFTed waveforms were used to explicitly reduce the temporal dimensionality of the imaging data.

In the two-class setting when using the 663, 040 imaging features or com-

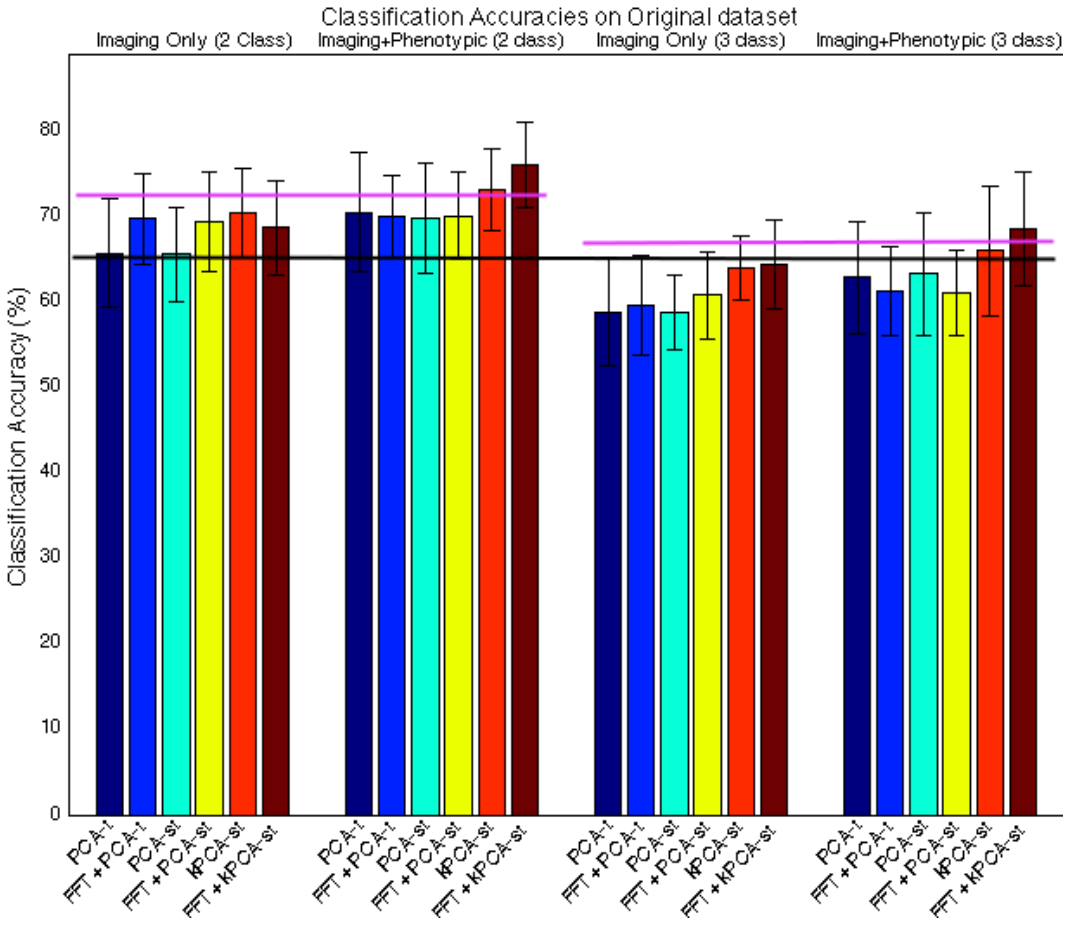

Figure 3.1: Visualization of the results on the original dataset (Table 3.2), where the error bar is ± the standard deviation. The purple lines denote the phenotypic data classification accuracy in the two-class (left) and three-class (right) setting, and the black line denotes the baseline in both the two and three-class settings, where the error bars are the added and subtracted standard deviations.

bining them with the phenotypic data, FFT produced accuracies that were statistically better than the baseline (p = 5.53e-4, p=2.14e-3). However, in the three-class setting FFT failed to produce accuracies equal to, or above, the baseline. This suggested that dimensionality reduction afforded by extracting the magnitude of the frequency components produced features that discriminated patients from controls in the two-class setting.

We show that FFT is more effective as a preprocessing step in the next subsection, and substantiate our claims with the results of FFT+PCA-*t*, FFT+PCA-*st* and FFT+kPCA-*st*.

3.2.3 FFT then PCA Variant

Instead of applying PCA-*t*, PCA-*st* and kPCA-*st* to the waveforms directly, as done in the previous and current sections, we first applied FFT as a *preprocessing step* to determine whether using magnitudes of the FFTed waveforms

could improve the results. The magnitude represents the amplitude, or height, of each frequency component, which may be more informative than the raw BOLD intensities. This transformed the imaging data's temporal dimensionality from 370 time points to 185 frequency components. If we let T equal the number of frequency components' magnitudes, then the process outlined in Figures 2.4 and 2.5 still applies.

For FFT+PCA-*t*, we only use the m=2 largest components; for both PCA-*st* and kPCA-*st*, we use all m=668 principal components.

In the two-class setting using only the imaging data, FFT+PCA-*st*, FFT+PCA-*st* and FFT+kPCA-*st* produced classification accuracies that are statistically better (p=1.13e-3, p = 7.10e-3 and p = 0.049) than the baseline. Combining FFT+PCA-*t* and FFT+PCA-*st*'s reduced imaging features with the phenotypic data resulted in a classification accuracy that was statistically better accuracy than the baseline (p =1.05e-3 and p=1.44e-3). However, this performance is nearly identical to using only FFT+PCA-*t* and FFT+PCA-*st*'s reduced imaging features. This suggests that combining the phenotypic data with the reduced-imaging features produced by FFT+PCA-*t* and FFT+PCA-*st* do not improve the discriminatory power of the classifier.

In the two class setting, combining the phenotypic data with the reduced imaging features of FFT+kPCA-*st* resulted in a classification accuracy that was statistically better than the baseline, PCA-*t*, PCA-*st*, *and* the phenotypic data (p=1.06e-7, p=4.76e-3, p=5.59e-3 and p=4.71e-2). FFT+kPCA-*st*'s results suggest that using the imaging features with the phenotypic data substantially improved the discriminatory power of our classifier.

In the three class setting, FFT+PCA-*t* and FFT+PCA-*st* performed similar to the case where FFT is not used as a preprocessing step – i.e., they failed to produce classification accuracies that were statistically better than the baseline. Each of FFT+PCA-*t*, FFT+PCA-*st* and FFT+kPCA-*st* fails to outperform the baseline when using imaging data only. Even here, FFT+kPCA-*st* produced a classification accuracy that was statistically better than FFT+PCA-*t* (p=3.2e-2), but not FFT+PCA-*st*. Combining the phenotypic data with FFT+kPCA-*st*'s reduced imaging features produced a classification accuracy that was statistically better than the baseline (p=4.7e-3), but was not statistically better than using only the phenotypic data.

When using phenotypic and imaging data in the two-class setting, FFT+kPCA-*st*'s statistically better performance than the phenotypic data suggests that FFT preprocessing improves the level of agreement between the reduced imaging features and phenotypic data. FFT+kPCA-*st* is the only method that produced a statistically better result than the phenotypic data in either the two or three-class settings, which also suggests that FFT+kPCA-*st*'s reduced imaging features, when combined with the phenotypic data, *improve* discriminatory power of the classifier.

The results suggest that taking the magnitudes of all voxel waveforms' frequency components returned by FFT could be a useful *preprocessing* step for voxel waveforms.

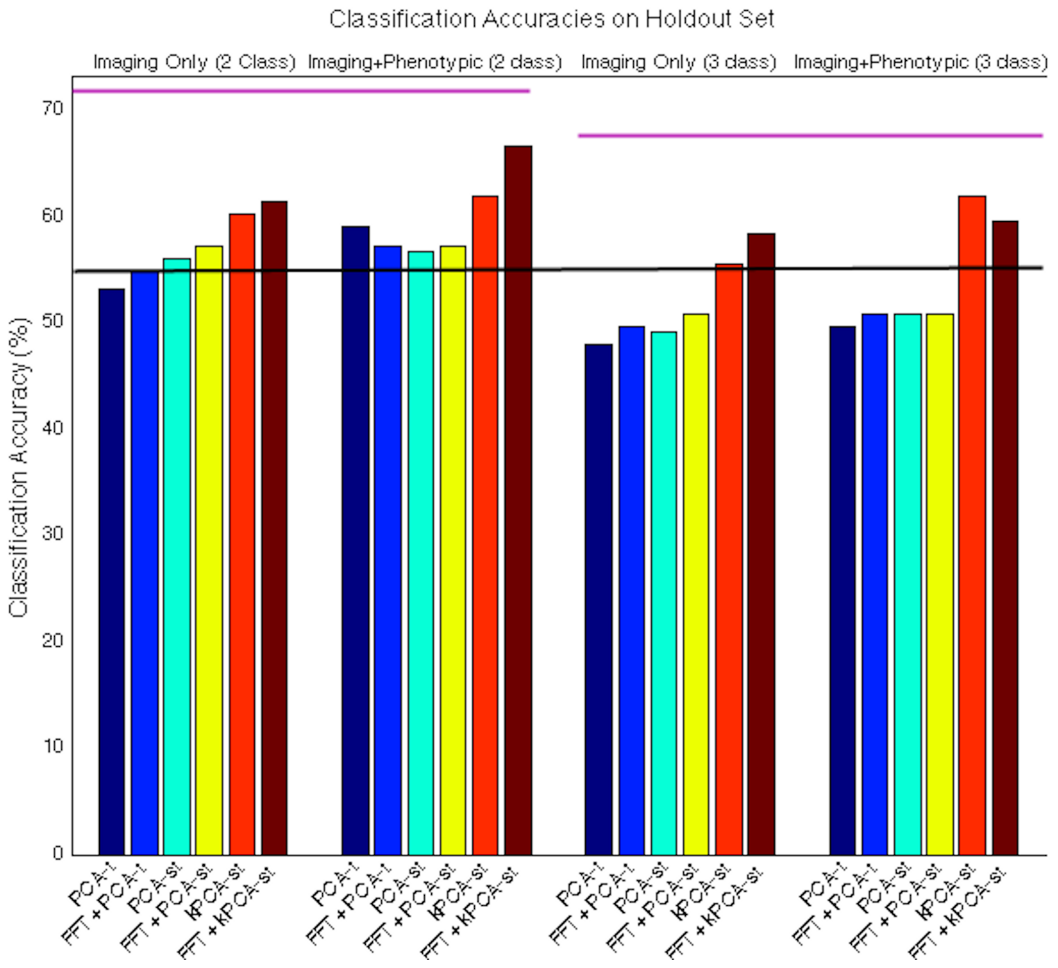

Figure 3.2: Visualization of the results on the holdout dataset (Table 3.3). The purple lines denote the phenotypic data classification accuracy in the two-class (left) and three-class (right) setting, the black line denotes the baseline in both the two and three-class settings where the error bars are the added and subtracted standard deviations.

3.3 Holdout Set Results

The ADHD200 Competition used a test set to evaluate the submissions; the organizers released this data after the competition had concluded. We used 171 of the 197 subjects in the test set, as 26 subjects were collected from a hospital that did not authorize the release of these 26 subjects' diagnoses. For both the two-class and three-class settings, the baseline accuracy on this dataset is 54.97%.

3.3.1 PCA Variants Only

In the two class setting, using only the reduced imaging features, PCA-*t*, PCA-*st* and kPCA-*st* produced classifiers with accuracies of 53.22%, 56.14% and 60.23%; note only PCA-*t*'s accuracy is below the baseline accuracy of 54.97%. Using phenotypic data only resulted in a classification accuracy of 71.35%.

# of classes	Baseline	Phenotypic Only	FFTed Waveforms	PCA Variant	Imaging Only	Imaging & Phenotypic data
2	54.97	71.35	-	PCA-t	53.22	59.1
			-	PCA-st	56.14	56.73
			-	kPCA-st	60.23	61.99
			+	–	56.73	56.73
			+	PCA-t	54.97	57.31
			+	PCA-st	57.31	57.31
			+	kPCA-st	61.4	66.67
3	54.97	67.25	-	PCA-t	47.95	49.71
			-	PCA-st	49.12	50.88
			-	kPCA-st	55.56	61.99
			+	–	50.88	53.22
			+	PCA-t	49.71	50.88
			+	PCA-st	50.88	50.88
			+	kPCA-st	58.48	59.65

Table 3.3: Two and three-class classification accuracies on the hold out set, where the reduced features were obtained by performing FFT and/or PCA-*t*, PCA-*st* and kPCA-*st* on the averaged imaging data.

These results are consistent with the cross-validation results from the original dataset: the phenotypic data outperforms any method that used only the reduced imaging data, and kPCA-*st* outperforms both PCA-*t* and PCA-*st*. When combining the phenotypic data to the imaging data in the two-class setting, PCA-*t*, PCA-*st* and kPCA-*st* perform better than the baseline, but not the phenotypic data.

In the three-class setting, both PCA-*t* and PCA-*st* fail to produce classification accuracies that are above the baseline. In contrast, the classification accuracy for kPCA-*st* using only the reduced imaging features is 55.56%, which is slightly better than the baseline in the three-class setting. When these reduced imaging features are combined with the phenotypic data, a classification accuracy of 61.99% is produced.

3.3.2 FFT Only

Interestingly, FFT's performance on the hold out dataset is not consistent with the results on the original dataset; see the result in the two-class setting using the phenotypic and imaging data. FFT fails to produce accuracies that are well-above baseline in either the two or three-class settings. Similar to what was done on the original dataset, we used FFT as a preprocessing step so that the voxel waveform frequency component magnitudes could be given as input to PCA-*t*, PCA-*st* and kPCA-*st*. As we show in the next subsection, this preprocessing step improves the result of each method.

3.3.3 FFT then PCA Variant

In the two-class setting using only the imaging data, FFT+PCA-*t* performs equal to the baseline whereas FFT+PCA-*st* and FFT+kPCA-*st* outperform the baseline by different margins, with FFT+kPCA-*st*'s margin being larger

than FFT+PCA-*st*'s. When combined with the phenotypic data, FFT+PCA-*t*, FFT+PCA-*st* and FFT+kPCA-*st* all outperform the baseline, with FFT+kPCA-*st*'s performance being superior to FFT+PCA-*t* and FFT+PCA-*st*. Even with FFT as a preprocessing step, combining the phenotypic data with FFT+kPCA-*st*'s reduced-imaging features fails to outperform the phenotypic data in the two-class setting.

In the three-class setting when using only the imaging data, FFT+PCA-*t* and FFT+PCA-*st* fail to outperform the baseline. In comparison, FFT+kPCA-*st* produces an accuracy of 58.48%, which is better than the baseline when using only the reduced-imaging features. When combining the phenotypic data with the reduced imaging features, both FFT+PCA-*t* and FFT+PCA-*st* fail to outperform the baseline whereas FFT+kPCA-*st* produces an accuracy of 59.65% that is better than the baseline, but is slightly inferior to kPCA-*st*'s accuracy of 61.99% in the same setting.

In general, FFT as a preprocessing step improves each method's classification accuracy. kPCA-*st* still produces the best overall result and consistently outperforms PCA-*st* and PCA-*t*. We see that FFT+kPCA-*st* benefits the most from using FFT as a preprocessing step, as it produces classification accuracies that are better than the baseline *and* kPCA-*st* when using only the imaging data, in both the two and three-class settings.

The results suggest that FFT should be used a preprocessing step to extract the magnitudes of voxel waveforms' frequency components before applying dimensionality reduction. Furthermore, results on the holdout set are consistent with those achieved on the original dataset, which substantiates our hypothesis that reducing dimensionality using kPCA-*st*, which reduces over both the spatial and temporal dimensions, is better than methods that reduce only the temporal or only the spatial dimensions, such as PCA-*t*. However, there is more to be desired because the phenotypic data outperforms our methods in either setting on the holdout set.

3.4 Potential Limitations of the Data

3.4.1 Unstandardized Scanning protocol

The ADHD200 data was collected from multiple hospitals across the world. One possible problem with this large data release is that the scanning protocols across all hospitals could differ. If each site used different scanning protocol, it is likely that each site's fMRI quality control criteria are also different. If the criteria were different for each hospital then, the different fMRI quality standards could potentially hurt our result.

3.4.2 Truncating scan time in preprocessed data

Every subject's scan length was different in different hospitals. In our preprocessed data, subject images containing longer scan lengths were truncated

to the shortest scan length in the data. To evaluate the consequence of such truncation, we considered learning a classifier within each hospital, using the ADHD200 Competition's preprocessed data.

Here, we only considered performance *within sites* because the temporal and spatial dimensions were not consistent across sites. Using subjects from the NeuroIMAGE hospital [12], we compared kPCA-*st*'s performance on the Competition's preprocessed data, to the result that used our preprocessed data; we observed that kPCA-*st* produced similar results for each dataset. This suggests that truncating temporal information in our preprocessed data did not substantially impact performance.

Masking

In Subsection 2.5, we mentioned that a mask was used to exclude the waveforms of the voxels outside of the brain. Applying the mask *before* BOLD-signal normalization dramatically impacted the classification accuracy, significantly decreasing the classification accuracy. We observed that all voxels outside of the brain had negative values and all voxels inside the brain had positive values. The magnitude of the voxels outside of the brain depended on the scanning site, and applying the mask prior to BOLD-signal normalization ignored information that we believe should be included. Thus, all images were masked *after* BOLD-signal normalization was applied over the averaged image.

We believed that normalizing the signal for *all* 6688 spatial locations *prior* to applying the mask allowed the BOLD signal to be retain information about the site it was from. Future work should thoroughly investigate the impact of masking prior to BOLD-signal normalization after performing image registration using the fMRI data from different sites and/or scanners.

3.5 Discussion

This article shows that applying kernel Principal Component Analysis (kPCA-*st*) leads to classifiers that are statistically better than canonical PCA (PCA-*t*) in every case using only the imaging data, except when FFT is used as a preprocessing step in the two-class setting.

Without using FFT to preprocess voxel waveforms, our results show that kPCA-*st* is statistically better than canonical PCA (PCA-*t*) when only using imaging data in both the two-class (p=0.043) and three-class (p=0.018) settings. It is also statistically better than PCA-*t* in the three-class setting when using both imaging and phenotypic data (p=0.0468).

When FFT is used as a preprocessing step, FFT+kPCA-*st* is statistically better than FFT+PCA-*st* (p=4.76e-3 and p=9.26e-3) and FFT+PCA-*t* (p=5.59e-3 and p=7.36e-3) when using the imaging and phenotypic data in the two and three-class settings; it is also statistically better than PCA-*t* (p=3.21e-2), but not PCA-*st* (p=1.02e-1), when only using the imaging data

in the three-class setting.

kPCA-*st*'s dominant performance over PCA-*t* in either the two-class or three-class settings without using FFTed waveforms suggests that using *non-linear* mappings to reduce over the spatial and temporal dimensions is superior to using linear mappings to reduce over the temporal dimension. However, FFT+kPCA-*st* fails to produce a result that is statistically better than FFT+PCA-*t*, which suggests that using the FFTed waveforms greatly benefits PCA-*t* and marginally impacts kPCA-*st*. A possible explanation for this result is that kPCA-*st*'s advantage over PCA-*st* and PCA-*t* is mitigated by introducing a preprocessing step that reduces the temporal dimensions of the imaging data.

The results show that non-linear mappings of subjects' fMRI data to a high-dimensional inner product space, as kPCA-*st* does, can increase discriminatory power of a classifier when compared to methods that do not, such as PCA-*t* and FFT. This inner product space allowed the spatial and temporal dimensionality to be preserved unlike PCA-*t*, which reduced over the temporal dimensions.

Interestingly, FFT+kPCA-*st* outperforms the phenotypic data in *both* the two and three-class setting, but *only the result in the two-class setting was statistically better than the phenotypic data* (p=4.7e-2). We believe that FFT+kPCA-*st*'s statistically insignificantly improvement over the phenotypic data in the three-class setting is not a large issue, because the phenotypic data itself was not statistically better than the baseline.

There are two potentially large consequences of our results:

1. Combining the imaging and phenotypic data improves the discrimination of ADHD subtypes from healthy controls. This is substantiated by the fact that separately using either FFT+kPCA-*st*'s features or the phenotypic data, produces accuracies that are not statistically the baseline, but their combination produces an accuracy that is statistically better than the baseline.
2. The imaging data is only sufficient to discriminate ADHD patients from controls when assigning ADHD subtypes to the same class.

To elaborate on the second statement: the failed distinction between subtypes and controls may be a consequence of only having 141 and 98 subjects for the combined type and inattentive subtypes, but 429 subjects for healthy controls. Given that there is over three times as many images for control subjects than ADHD combined-type patients, the second largest class, providing additional ADHD patient scans may improve accuracy in the three-class setting.

An alternative explanation is that ADHD subtypes are not distinguishable in the three-class setting when using only the imaging data. Our results support this claim because aggregating ADHD combined and inattentive types into a single class improved the discriminatory power of kPCA-*st* in the two-class setting. One possible explanation for this result is that the number of features that distinguish ADHD subtypes from controls (or subtypes from each

other) is very small in comparison to those that discriminate *all* ADHD patients from healthy controls.

Chapter 4

FEP Dataset

First Episode Schizophrenia (or First Episode Psychosis, FEP) is defined as the first psychotic episode experienced by an individual. While only having a lifetime prevalence of about 1%, the associated cost and social debilitations are very large in comparison to this proportion. As of 10 years ago, it was estimated that schizophrenia costs about 2.35 billion dollars CAD [7].

FEP usually occurs between late adolescence and the early twenties, and is often undetected until two to three years after clearly diagnosable symptoms manifest themselves. This delayed identification is a consequence of most subjects possessing other pre-existing conditions involving cognition, language or behavior [7]. There is a significant push by the psychiatric community to improve FEP detection in order to decrease the subject's stigma, as it often hinders their ability to re-integrate into their community. Thus, earlier detection of FEP can lead to a subject re-integrating into the community as a productive citizen, which helps to offset the "cost" incurred to treat these patients.

Since kPCA-*st*'s on the ADHD200 involved resting-state fMRI, we use the FEP dataset, which contains block design fMRI for subjects with First Episode Schizophrenia, to determine if dimensionality reduction processes' performance will generalize to fMRI datasets that differ in experimental design *and* neuropsychiatric disorders.

4.1 Results

We use the same methodology as the ADHD200 data for applying PCA-t, PCA-st and kPCA-st to the FEP dataset with one exception: m was chosen to capture over 98% of the variance; see Section 4.2. The data matrices, $\mathbf{X}_t$ and $\mathbf{X}_{st}$, processed by PCA-t and PCA-st/kPCA-st have dimensions $(52975 * 34) \times 28$ and $(52975 * 28) \times 34$ respectively. The results were obtained by performing 17-fold cross validation, where each fold contained an FEP patient and healthy control, resulting in a baseline accuracy of 50%.

For PCA-t we only use the m=1 largest principal component, which produces classification accuracy that were equal to the baseline. For PCA-st and

kPCA-*st*, we use the m=33 and m=32 largest principal components, respectively. Using PCA-*st*'s 33 largest principal components produced a classification accuracy that was also equal to the baseline, suggesting that both PCA-*t* and PCA-*st*'s reduced imaging features were unable to discriminate FEP patients from controls. Using kPCA-*st*'s 32 largest principal components produced a classification accuracy of 70.6% that is statistically better than the baseline (p=0.049) and PCA-*t* (p = 2.4e-2).

Using the GLM's $2 * 57 * 67 * 50$ parameters[1] as features produces a classification accuracy of 73.5% that is statistically better than the baseline (p = 3.67e-2) and PCA-*t* (p = 2.7e-7), but is not statistically better than kPCA-*st*. To eliminate the effect of the parameters that correspond to voxels outside of the brain, we select the 52975 parameters that correspond to voxels inside the brain; learning a classifier over the $2 * 52975$ parameters produced a classification accuracy of 73.5% that is statistically better than baseline(p = 4.11e-8) and PCA-*t* (p = 2.8e-2), but is not statistically better than kPCA-*st*.

4.2 Comparing the FEP and ADHD200 Dataset Results

Since we select m to capture 99% and 98% of the variance in the ADHD200 and FEP datasets respectively, we show that kPCA-*st* produces a similar result on the ADHD200 dataset when using the m=641 principal components that capture 98% of the variance: In the two-class setting, classification accuracies of 70.1% (p=3.52e-3) and 73.7% (p=1.08e-4) were statistically significantly better than baseline when using the imaging data only, or combining it with the phenotypic data.

In comparison to using all 668 principal components on the ADHD200 dataset, kPCA-*st* performs slightly worse with m=641 principal components, which shows that including the 27 smallest principal components improves the performance by a small amount. However, for the FEP dataset we noticed that using kPCA-*st*'s m=32 largest principal components, which capture 98% of the variance, performed statistically better than baseline and PCA-*t*, whereas using m=34 principal components, which capture over 99% of the variance, did not perform not statistically better than baseline or PCA-*t*. This is an interesting phenomenon when considering that the ADHD200 dataset's performance *improved* when the smallest 27 principal components were included.

After further investigation, we saw that kPCA-*st*'s smallest principal component captured approximately 2e-17% of the variance in the FEP dataset. We conjecture that this principal component is produced solely because of our choice in kernel, as the RBF kernel produces N eigenvectors whose eigenvalues are strictly positive; we feel this is a legitimate explanation when considering that PCA-*st*, which uses the linear kernel, only produced 33 eigenvectors with strictly positive eigenvalues. Furthermore, we believe that having only 34 eigenvectors affected the proportion of variance captured by each eigenvector

[1] See Section 2.3.

in the FEP dataset, as the ADHD200 data showed a *decrease* in performance when using fewer principal components, but the FEP data showed an *increase* in performance. We therefore use only the m=32 largest principal components produced by kPCA-*st* to exclude any erratic phenomena that might be contained in the two smallest principal components.

We considered using the FEP subjects' *full* voxel waveforms as input to the PCA variants, but found that none of the PCA variants performed statistically better than the baseline. We also considered applying the Fourier Transform to the FEP subjects' voxel waveforms before they were given as input to the PCA variants, however we observed that none of these processes performed statistically better than baseline. It is possible that incorporating biological information is pivotal for discriminating patients from controls in block design fMRI when using subjects' entire voxel waveform, as the voxels' activation patterns during each event type is presumably different.

We believe that kPCA-*st*'s statistically significantly better results than PCA-t on both datasets is a consequence of using a non-linear mapping to measure the *point-wise* similarity between two subjects' waveform values at the same voxel location. When only the imaging data is used in the two-class setting without FFT as a preprocessing step, kPCA-*st*'s classification accuracies on *both* the FEP and ADHD200 datasets were roughly $\approx 70\%$. kPCA-*st*'s similar performance on two datasets that differed in neuropsychiatric disorder and experimental design suggests that kPCA-*st*'s performance could generalize to different fMRI datasets.

4.3 Discussion

Dealing with block design fMRI data is difficult because it contains both rest periods and task blocks. We have shown that using the resting period time points for each subject's block design fMRI produces statistically significantly better results than the baseline and PCA-t (p=0.049) and PCA-t (p = 2.4e-2). Such a result strengthens our claim that kPCA-*st* is superior to PCA-t, especially considering that *both* datasets were resting-state.

We think that analyzing block design fMRI is a larger challenge than resting-state fMRI because the variance of *specific* voxel waveforms–i.e. the voxels that exhibit elevated activations during the event– will increase, while others will not be affected. We showed that one way to alleviate such discrepancies introduced by block design fMRI is to incorporate only the resting-state time points.

Chapter 5

Future Work

kPCA-*st*'s results suggest that it can discriminate patients from controls on datasets with different neuropsychiatric disorders and experimental designs. This result is interesting because kPCA-*st* is biologically naive, suggesting that there is ample room to improve our result.

It is reasonable to suggest that incorporating biological information when performing dimensionality reduction of fMRI could improve the result. The hemodynamic response function (HRF), defined as the function that models the response of the system after brief and intense period of neural stimulation, is an example of biological information that could be helpful in reducing fMRI dimensionality. Previous work has shown intra-subject differences in activation levels for different brain regions [5][29] and implies that such differences may arise from hemodynamic factors. Nakai et al. show that there are intra-subject differences between the HRF for the primary sensorimotor (SM1) and supplementary motor area (SMA) [22].

When observing brain regions using fMRI, each region can be viewed as a collection of neighbouring voxels. Since each waveform represents a voxel's BOLD-signal intensity over the scan duration, it is possible for voxels *within* a brain region– i.e., voxels that are in close proximity – to have a similar HRF, which would be reflected by their waveform, because of their similar vascular properties. To test this assertion, we treat each voxel's three-dimensional index as a three-dimensional point and compute the euclidean distance between voxels, which is a measure of *physical proximity*; if the assertion is true, then we can compute the k-Nearest Neighbours (k-NN) for every voxel, which are the k nearest voxels in three-dimensions for the respective voxel, to compare this voxel's waveform to its neighboring voxels' waveforms, which allows us to observe if the waveforms are similar, as shown in Figure 5.1.

If we construct a $V \times T$ matrix for each subject, where each row is a voxel waveform, we can perform Locally Linear Embedding (LLE) [28] on each subject's $V \times T$ matrix, which would reconstruct each waveform in terms of its K nearest neighboring voxel waveforms using d points, where $d \leq T$ is a user-specified parameter.

The appealing property of applying LLE to fMRI data is that it reconstructs a voxel waveform in terms of its k nearest neighbors, which does not

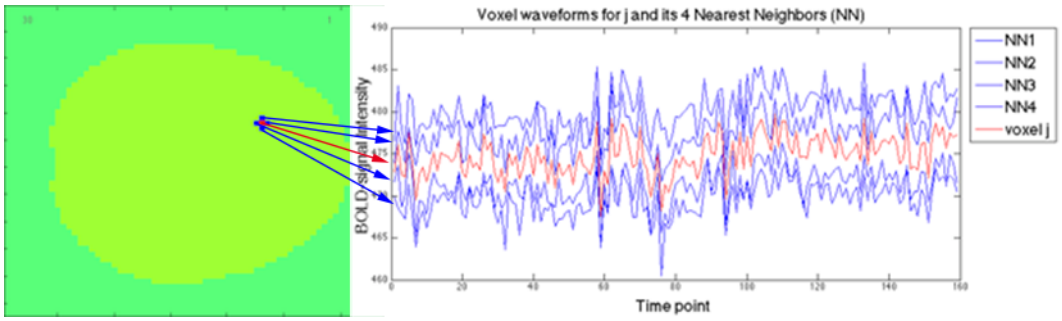

Figure 5.1: Illustrating (right) voxel j (red) and its 4 nearest neighbors (blue), which are determined by the euclidean distance between voxels in 3-D space, and their respective waveforms (right). These waveforms belonged to a healthy control from the FEP dataset.

require the BOLD-signal to be normalized, as the weights used to reconstruct the respective voxel waveform in terms of its neighboring waveforms are *rotation, shift, and translation invariant*. Furthermore, LLE is unaffected by subjects in the same class having different waveforms for the *same* voxel location, as the weights used to reconstruct a voxel waveform in terms of its neighbors are *relative* to the respective waveform. Thus, LLE exploits the property of *physical proximity* to reduce dimensionality. Preliminary results suggest that LLE is a dimensionality reduction method that may generalize to *all* fMRI data, but we are still in the process of evaluating performance on the ADHD200 dataset.

The above example illustrates how to encode very general biological information into a graphical model that is used to reduce fMRI dimensionality. The unifying intuition behind the dimensionality reduction methods that reduce fMRI to produce a statistically better result than the baseline, is that they focus on comparing the voxel waveforms across all subjects in the dataset. We believe that future work should focus on encoding biological information that is *specific* to a neuropsychiatric disorder, because these graphical models may improve the discrimination of patients from controls; one way to encode this information would be to assign *priors* for voxel locations that correspond to brain regions that are associated with a specific neuropsychiatric disorder.

Chapter 6

Conclusion

This dissertation investigated dimensionality reduction of fMRI data using two separate datasets that differed in neuropsychiatric disorder and experimental design. We showed that kPCA-*st* produced accuracies that were statistically significantly better than PCA-*t* on both the FEP and ADHD200 datasets when using the resting-state data only.

When FFT was used as a preprocessing step, as was done for the ADHD200 dataset, FFT+kPCA-*st* produced a statistically significantly better result than FFT+PCA-*t* in every setting except the two-class setting when using only the imaging data. Given that FFT+kPCA-*st*, FFT+PCA-*t* and FFT+PCA-*st* had very similar accuracies in the two-class setting when using only the imaging data, the reduced temporal dimensionality afforded by FFT as a preprocessing step improved the performance of PCA-*st* and PCA-*t*, but had negligible impact on kPCA-*st*.

Our results support the statement that dimensionality reduction methods, such as kPCA-*st*, that use non-linear mappings to reduce over *both* the spatial and temporal dimensions improve the diagnostic system's discrimination of group differences in comparison to methods that reduce only the spatial or only the temporal dimensionality, such as PCA-*t*. This statement is strengthened when considering that it was true for two datasets that differed in neuropsychiatric disorders and experimental design.

While the performance of kPCA-*st* is not at the level where fMRI-based diagnosis is feasible, *the results are a proof-of-concept for fMRI-based diagnosis.* kPCA-*st* reduces fMRI dimensionality by measuring the point-wise similarity between two subjects entire fMRI image (after applying mapping Φ) in an inner-product space. The diagnostic classifier learned over the reduced imaging features produced by kPCA-*st* distinguishes patients from controls at a statistically better level than guessing the majority class. Preliminary results from our recent work (LLE) gave accuracies of 82.3% on the FEP dataset and 68.6% on the ADHD200 dataset when using only the NYU subjects from the original dataset[1]. These results substantiate the claim that fMRI has diagnostic value, but the real question is: how much?

[1]The baseline for this dataset is 51.5%.

Bibliography

[1] *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision)*. Amer Psychiatric Pub, 07 2000.

[2] Ethem Alpaydin. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004.

[3] Anders H. Andersen, Don M. Gash, and Malcolm J. Avison. Principal component analysis of the dynamic response measured by fmri: a generalized linear systems framework. *Magnetic Resonance Imaging*, 17(6):795 – 815, 1999.

[4] Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995.

[5] Randy L. Buckner, Wilma Koutstaal, Daniel L. Schacter, Anders M. Dale, Michael Rotte, and Bruce R. Rosen. Functionalanatomic study of episodic retrieval: Ii. selective averaging of event-related fmri trials to test the retrieval success hypothesis. *NeuroImage*, 7(3):163 – 175, 1998.

[6] Heather J Cordell. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404, 06 2009.

[7] T J Crow. How to manage the first episode of schizophrenia. authors did not take account of systemically collected information. *BMJ*, 322(7280):234–5, 2001.

[8] John E Desmond and Gary H Glover. Estimating sample size in functional mri (fmri) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, 118(2):115 – 128, 2002.

[9] Nico U. F. Dosenbach, Binyam Nardos, Alexander L. Cohen, Damien A. Fair, Jonathan D. Power, Jessica A. Church, Steven M. Nelson, Gagan S. Wig, Alecia C. Vogel, Christina N. Lessov-Schlaggar, Kelly Anne Barnes, Joseph W. Dubis, Eric Feczko, Rebecca S. Coalson, John R. Pruett, Deanna M. Barch, Steven E. Petersen, and Bradley L. Schlaggar. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

[10] Naomi I. Eisenberger, Matthew D. Lieberman, and Kipling D. Williams. Does rejection hurt? an fmri study of social exclusion. *Science*, 302(5643):290–292, 2003.

[11] A. C. Evans, D. L. Collins, S. R. Millst, E. D. Brown, R. L. Kelly, and T. M. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. pages 1813–1817, 1993.

[12] Mennes Martin Fair, Damien and Michael Milham. ADHD200 Global Competition. *http://www.fcon_1000.projects.nitrc.org/indi/adhd200/*.

[13] Mennes Martin Fair, Damien and Michael Milham. ADHD200 Global Competition Results. *http://fcon_1000.projects.nitrc.org/indi/adhd200/results.html*.

[14] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.*, 2(4):189–210, 1994.

[15] The FIL Methods Group. "Statistical Parametric Mapping". *http://www.fil.ion.ucl.ac.uk/spm/*.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2nd ed. 2009. corr. 3rd printing 5th printing. edition, September 2009.

[17] Ralf Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Cambridge, MA, USA, 2001.

[18] S.A. Huettel, A.W. Song, and G. McCarthy. *Functional magnetic resonance imaging*. Sinauer Associates, 2009.

[19] Ian Jolliffe. *Principal component analysis*. Springer Verlag, New York, 2002.

[20] O. Josephs, R. Turner, and K.J. Friston. Event-related fMRI. *Human Brain Mapping*, 5:243–248, 1997.

[21] Nikos K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 06 2008.

[22] Toshiharu Nakai, Kayako Matsuo, Chikako Kato, Yasuo Takehara, Haruo Isoda, Tetsuo Moriya, Tomohisa Okada, and Harumi Sakahara. Post-stimulus response in hemodynamics observed by functional magnetic resonance imagingdifference between the primary sensorimotor area and the supplementary motor area. *Magnetic Resonance Imaging*, 18(10):1215 – 1219, 2000.

[23] S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[24] William E. Pelham, E. Michael Foster, and Jessica A. Robb. The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. *Ambulatory Pediatrics*, 7(1, Supplement):121 – 131, 2007. ¡ce:title¿Measuring Outcomes in Attention Deficit Hyperactivity Disorder¡/ce:title¿.

[25] Woodward ND Wilman AH Tibbo PG Purdon SE, Waldie B. Procedural learning in first episode schizophrenia investigated with functional magnetic resonance imaging. *Neuropsychology*, 25(2):147–158, 2011.

[26] Alvin C. Rencher. *Methods of multivariate analysis*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. J. Wiley, 2002.

[27] Lee N. Robins and John E. Helzer. Diagnosis and clinical assessment: The current state of psychiatric diagnosis. *Annual Review of Psychology*, 37(1):409 – 432, 1986.

[28] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 4:119–155, December 2003.

[29] Daniel L. Schacter, Randy L. Buckner, Wilma Koutstaal, Anders M. Dale, and Bruce R. Rosen. Late onset of anterior prefrontal activity during true and false recognition: An event-related fmri study. *NeuroImage*, 6(4):259 – 269, 1997.

[30] Bernhard Schölkopf and Alexander J. Smola. Learning with kernels : support vector machines, regularization, optimization, and beyond. Hardcover, December 2002.

[31] Ilina Singh. Beyond polemics: science and ethics of adhd. *Nat Rev Neurosci*, 9(12):957–964, 12 2008.

[32] Daniel J Smith and Nassir Ghaemi. Is underdiagnosis the main pitfall when diagnosing bipolar disorder? yes. *BMJ*, 340, 2 2010.

[33] Ian Wallace and Angela Wood. *The sandwich*. Kids Can Press, Toronto, 1985.

[34] JB Williams, M Gibbon, MB First, RL Spitzer, M Davies, J Borus, MJ Howes, J Kane, HG Pope, and B Rounsaville. The structured clinical interview for dsm-iii-r (scid). ii. multisite test-retest reliability. *Archives of general psychiatry*, 49(8), 08 1992.

[35] Qing-Hai Ye, Lun-Xiu Qin, Marshonna Forgues, Ping He, Jin Woo Kim, Amy C. Peng, Richard Simon, Yan Li, Ana I. Robles, Yidong Chen, Zeng-Chen Ma, Zhi-Quan Wu, Sheng-Long Ye, Yin-Kun Liu, Zhao-You Tang, and Xin Wei Wang. Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nat Med*, 9(4):416–423, 04 2003.

Appendix A

Supplementary Information for Clinical Researchers

A.1 ADHD200 Dataset

A.1.1 Demographics

The ADHD200 Global Competition was an initiative which focused on expediting the scientific community's understanding of ADHD by publicly sharing a dataset consisting of 776 fMRI images collected at eight different hospitals, where each hospital provided various amounts of phenotypic data for the respective subject's fMRI. All hospitals provided the age, gender and handedness of the subject. We briefly discuss the distribution of the phenotypic data for each hospital. Table 3.1 provides the distribution of diagnoses for each hospital.

Kennedy Krieger Institute (KKI)

The Kennedy Krieger Institute (KKI) contained 78 subjects, where 42 were male and 36 were female. The subjects' average age in this dataset is 10.27 years, where the oldest and youngest subjects were 12.99 and 8.02 years of age, respectively. All of the KKI subjects had Verbal, Performance and Full4 IQ scores from the Wechsler Intelligence Scale for Children (WISC-IV) and ADHD Index, Inattentive and Hyperactivity/Impulsivity scores obtained from the Connors' Parent Rating Scale-Revised, Long Version (CPRS-LV).

NeuroIMAGE

The NeuroIMAGE sample contained 38 subjects, where 24 were male and 14 were female. The subjects' average age was 17.42 years, with the oldest and youngest subjects being 21.74 and 11.05 years of age, respectively. No IQ or ADHD Index scores were provided for these subjects.

Peking

The Peking hospital contained 194 subjects, where 142 were male and 52 were female. The subjects' average age was 11.98 years, with the oldest and youngest subjects being 17.33 and 8.42 years of age. All of the Peking subjects had Verbal, Performance and Full4 IQ scores from the Wechsler Intelligence Scale for Chinese Children (WISCC-R). 172 subjects also had ADHD Index scores obtained from the ADHD Rating Scale IV (ADHD-RS).

New York University (NYU)

The New York University (NYU) hospital contained 188 subjects, where 118 were male and 70 were female. The subjects' average age was 11.86 years, with the oldest and youngest subjects being 17.96 and 7.17 years of age, respectively. All of the NYU subjects had Verbal, Performance and Full4 IQ scores from the Wechsler Intelligence Scale for Children (WISC-IV) and ADHD Index, Inattentive and Hyperactivity/Impulsivity scores obtained from the Connors' Parent Rating Scale-Revised, Long Version (CPRS-LV).

University of Pittsburgh

The University of Pittsburgh hospital contained 66 subjects, where 31 were male and 35 were female. The subjects' average age was 15.67 years, with the oldest and youngest subjects being 20.45 and 10.11 years of age, respectively. Only 34 of these subjects had Verbal, Performance and Full4 IQ scores from the Wechsler Intelligence Scale for Children (WISC-IV), and none had ADHD Index scores.

Washington University in St. Louis

The Washington University in St Louis hospital contained 40 subjects, where 22 were male and 18 were female. The subjects' average age was 12.53 years, with the oldest and youngest subjects being 21.83 and 7.09 years of age, respectively. All of these subjects had a Full4 IQ that was obtained from the Two subtest Wechsler Abbreviated Scale of Intelligence (WASI), but did not have any other IQ or ADHD Index scores.

Oregon Health and Science University (OHSU)

The Oregon Health and Science University sample contained 64 subjects, where 35 were male and 29 were female. The subjects' average age was 8.96 years, with the oldest and youngest subjects being 11.92 and 7.33 years of age, respectively. All of these subjects had a Full4 IQ that was obtained from the Wechsler Abbreviated Scale of Intelligence (WASI) and ADHD Inattentive and Hyperactivity/Impulse scores obtained from the Connors' Rating Scale-3rd Edition.

A.1.2 Diagnosis/Labels

As we mentioned in the introduction, there is substantial variability in the mental-health-related expertise of clinicians that diagnose ADHD in the United States. We feel that this variability could be reflected in the subject labels for each hospital. For example, New York University contains 97 ADHD patients and 91 healthy controls. These 97 ADHD patients account for 40% of the ADHD subjects in the dataset. Given the variability of the mental-health-related expertise of the clinicians, it is possible that these labels are incorrect.

We see one very real limitation with fMRI-based diagnosis: It requires expertly labelled data, where labelling this data is expensive. Given the comprehensiveness of psychiatric evaluation for many neuropsychiatric disorders, the labels are usually reliable. In the United States, however, neuropsychiatric disorders such as ADHD are not labelled exclusively by psychiatrists. It would be desirable to use fMRI to *learn* a diagnosis because it would substantially reduce the expense to diagnose subjects through comprehensive evaluations. In order to learn a diagnosis, we still need an expertly-labelled dataset to learn features that distinguish patients from healthy controls. To address this challenge, we first need to reduce fMRI dimensionality such that we can use the reduced features to *reproduce* an expert's diagnosis.

A.1.3 Rest Periods of Block Design fMRI vs resting-state fMRI

We use the last 2 time points of every rest period in the FEP subjects' block design fMRI data to homogenize the comparison of the PCA Variants. We acknowledge that the rest period time points are not the same as those from resting-state fMRI. This is a consequence of voxels' BOLD-signal activation returning to the baseline between the task blocks. Since the volume time for the FEP dataset is 3 seconds and the rest period is 18 seconds long, there are 6 time points for every rest period. For the first two-thirds of this rest-period, the voxels' BOLD-signal activation is returning to baseline. For the last third (2 time points), the voxels are at their baseline activation level.

It is clear that resting-state fMRI involves subjects sitting idly during the scan, and rest-periods in a block design fMRI also involve subjects sitting idly between task blocks. Since we account for the voxels' BOLD-signal activation returning to their baseline by selecting the last 2 time points of every rest period, we feel that these time points are *similar but not identical* to resting-state fMRI time points. It is possible that resting-state fMRI is different from the volumes collected at last 2 time points of a rest period in block design fMRI, however more data and investigation is required.

Appendix B

Support Vector Machine

Given N labeled data points $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$ for all *i=1,...,N*, the dataset consists of N pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \in \mathbb{R}^p$. Per Tibshirani *et al.* [16], define a hyperplane by

$$\{x : f(x) = x^T w + w_0 = 0\}$$

where w is a unit vector: $||w|| = 1$, and $f(x)$ gives the signed distance from a data point x to the hyperplane defined by $f(x) = x^T w + w_0 = 0$. Assuming that the classes are separable, we can find a function $f(x) = x^T w + w_0$ that satisfies $y_i f(x_i) > 0$ for all $i = 1, \ldots, N$. This hyperplane creates the largest *margin* between the data points for class 1 and -1. To recover this hyperplane, we solve the optimization problem

$$\begin{aligned} &\min_{w, w_0} ||w|| \\ &\text{subject to } \; y_i(x_i^T w + w_0) \geq 1, i = 1, \ldots, N \end{aligned} \tag{B.1}$$

where w is the vector that produces this hyperplane on the N pairs that are given as input, and $\frac{1}{||w||}$ is the size of the *margin* on *each* side of the hyperplane, as shown in Figure B.1.

Support Vector Machines (SVM) are a supervised learning method in Machine Learning commonly used for classification. SVM seeks to find the $(p-1)$-dimensional *maximum margin hyperplane*, which is defined as the hyperplane that best separates the data. When the data points for each class do not overlap– i.e., they are separable – then the formulation above will find the maximum margin hyperplane.

In the case where the data is not separable, which is often the case when dealing with real-world datasets, the maximum margin hyperplane that contains the least number of *misclassified* data points is selected. Define ξ_i as the *slack variable* representing the proportional amount by which the prediction for $\mathbf{x}_i$, which is given by $f(x_i) = x_i^T w + w_0$, is on the wrong side of its margin. Then the maximum margin hyperplane can be found from solving the following optimization problem:

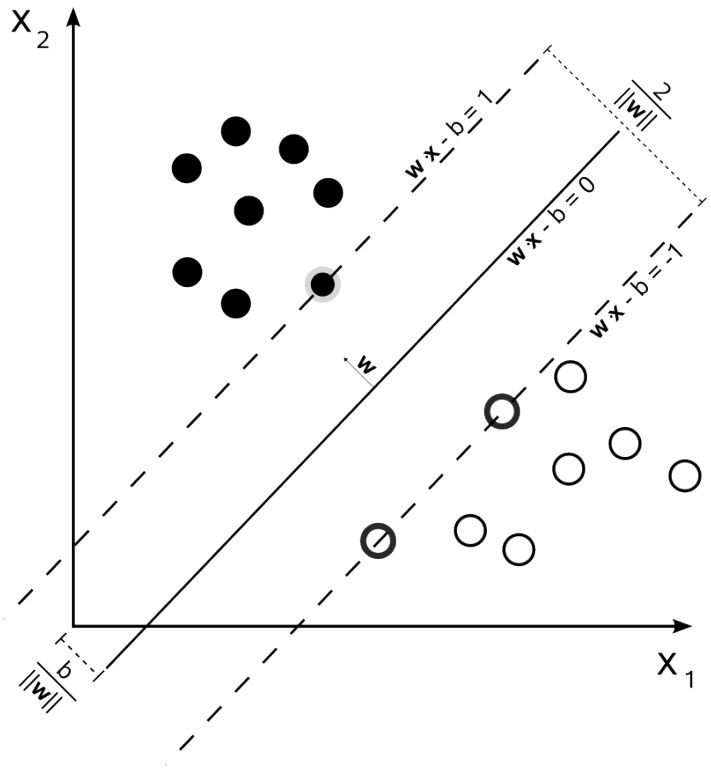

Figure B.1: A maximum-margin hyperplane produced from SVM on separable data.

$$\begin{aligned} &\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to} \begin{cases} y_i(x_i^\intercal w + w_0) \geq 1 - \xi_i \ \forall i \\ \xi_i \geq 0, \ \sum \xi_i \leq \text{constant} \end{cases} \end{aligned} \tag{B.2}$$

This optimization problem can be expressed as a quadratic programming solution that can be solved using the method of Lagrange Multipliers. LibSVM toolbox's $C-SVC$ classification algorithm solves the dual, and is the algorithm used to produce the classification results with the shrinking parameter disabled ('-h 0').

Appendix C

Kernel Principal Component Analysis

We select a transformation $\Phi(\mathbf{x}_i) : \mathbb{R}^n \rightarrow \mathbb{R}^p, n < p$ that is applied to the covariance matrix. This produces a covariance matrix $\mathbf{S} = \frac{1}{n}\sum_i \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^\intercal$, which only applies if the data is centered over its origin. For any eigenvector of $\mathbf{S}$ we will have:

$$\frac{1}{n}\sum_i \Phi(\mathbf{x}_i)\big(\Phi(\mathbf{x}_i)^\intercal \mathbf{e}\big) = \lambda \mathbf{e} \tag{C.1}$$

We want to find the [eigenvalue, eigenvector] pairs, $[\lambda, \mathbf{e}]$. The naive approach would solve this using a $p \times p$ matrix, where it is possible for $p \gg n$. We can avoid solving in a higher-dimensional space by recognizing that for any $\lambda \neq 0 \implies \mathbf{e} \in span(\{\Phi(\cdot)\})$, which means that we can write the eigenvector as the linear combination $\mathbf{e} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$ which can be substituted into Equation C.1:

$$\frac{1}{n}\sum_i \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^\intercal \sum_i \alpha_i \Phi(\mathbf{x}_i) = \lambda \sum_i \alpha_i \Phi(\mathbf{x}_i) \tag{C.2}$$

Note that all of the $\Phi(\mathbf{x}_i)$ expressions are in an inner product in equation C.2, so take the inner product with all data points $\Phi(\mathbf{x}_\ell)$, where $\ell = 1 : n$

$$\lambda \sum_i \alpha_i \Phi(\mathbf{x}_\ell)^\intercal \Phi(\mathbf{x}_i) = \lambda \sum_j \alpha_i \sum_i \big(\Phi(\mathbf{x}_\ell)^\intercal \Phi(\mathbf{x}_i)\big)\big(\Phi(\mathbf{x}_i)^\intercal \Phi(\mathbf{x}_j)\big) \tag{C.3}$$

For many classes of kernel functions, we have

$$K_{i,j} = \Phi(\mathbf{x}_i)^\intercal \Phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$$

thereby allowing us to rewrite Equation C.3 as

$$\lambda K \alpha = \frac{1}{n} K^2 \alpha$$

The kth principal component can be extracted using the kth eigenvector $\mathbf{e}_k$ as follows:

$$\mathbf{y}_k = \mathbf{e}_k^\intercal \mathbf{K}$$

where $\mathbf{y}_k = [y_{1,k}, \ldots, y_{n,k}]$, and $y_{i,k}$ corresponds to the k^{th} principal component for the i^{th} data point.

Note that kernel matrix $\mathbf{K}$ only grows with the number of samples instead of the dimensionality, which was the case in standard PCA. This article only uses the following kernel:

Radial Basis Function (RBF) kernel :

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\langle(\mathbf{x}_i - \mathbf{x}_j), (\mathbf{x}_i - \mathbf{x}_j)\rangle}{2\sigma^2}\right)$$

where σ is a parameter given as input, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the i^{th} and j^{th} subjects imaging data reshaped into one dimensional vectors.