The Pennsylvania State University

The Graduate School

College of Engineering

# HIV-ASSOCIATED NEUROPATHIC PAIN CLASSIFICATION OF MRI BRAIN IMAGES

A Thesis in

Electrical Engineering

by

Dongzhe Wang



Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

May 2014

The thesis of Dongzhe Wang was reviewed and approved* by the following:

David J. Miller
Professor of Electrical Engineering
Thesis Advisor

George Kesidis
Professor of Electrical Engineering

Kultegin Aydin
Professor of Electrical Engineering
Head of the Department of Electrical Engineering

*Signatures are on file in the Graduate School.

# ABSTRACT

HIV-associated sensory neuropathy influences over 50% of HIV patients. The clinical expression of HIV neuropathy is dramatically variable. Although many HIV patients report few symptoms, approximately half report distal neuropathic pain (DNP). To better understand how the central nervous system is associated with HIV DNP, in this thesis, an analysis of HIV-infected participants' brain structural magnetic resonance imaging (MRI) volumes was performed. Using multivariable regression analysis (involving demographic and clinical variables), the relationship between HIV DNP and the MRI results was investigated. Our study concluded that worse severity of DNP symptoms was correlated with smaller cerebral cortical gray matter [1]. According to this conclusion, we performed a statistical classification analysis on the presence of DNP symptoms in the structural MRI images. We generated three relevant feature extraction schemes, leading to three separate experiments. These three experiments will be helpful and informative for our study on clinical HIV DNP diagnosis. The novelty in this work relative to existing HIV DNP studies is the optimization of DNP classification performance based on the MRI data sets, using low dimensional features and computationally efficient models.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Dr. David J. Miller. His expertise and insight in the field of Pattern Recognition has been a great help in this research. This work is impossible without his invaluable guidance, constant support and patience. He has been the biggest source of encouragement for this work.

I would also give my thanks to the additional members of the advisory committee, Dr. George Kesidis, for his precious time in reviewing this thesis and constructive suggestions.

I would like to thank Dr. John Keltner, Assistant Professor from the Department of Psychiatry at University of California, San Diego. He is the MRI brain images provider. Without his guidance and patience from the perspective of neuroimaging, this thesis would not have been possible.

Special thanks to all my friends for their patience, love and encouragement. I would also like to express my thanks to my family for their unconditional love and constant support.

# Chapter 1

# INTRODUCTION

Pattern recognition has its origins in engineering, whereas machine learning stemmed from computer science. However, machine learning and pattern recognition could be viewed as two facets of the same field. And together they have been the focus of intense research over the past ten years. The fundamental ideas are about the discovery of pattern regularities in data through the use of automated computer algorithms. Using these learned regularities, we are able to develop some applications.

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems. Cases where the objective is to assign each input vector to one of a finite number of discrete categories are called classification problems. If the desired output consists of one or more continuous variables, the task is called regression [2].

The use of computer technology in medical diagnosis and treatments nowadays is intensively prevalent and widespread across a large range of medical areas. We have seen plenty of applications of pattern recognition in bioinformatics, such as chronic pain study, cardiac diseases, tumor detection, human motional asymmetry, etc. This study attempts to analyze bio-imaging through the use of supervised learning methods.

Persistent pain now affects so many individuals with HIV infection that it has recently been termed an “evolving epidemic” [3]. HIV-associated distal neuropathic pain is one of the most prevalent neurologic complications of HIV infection in the era of combination antiretroviral therapy (CART), affecting approximately 20% of patients. HIV DNP is typically difficult to cure by current chronic pain therapies and is associated with unemployment, impairment in activities of daily living, and significantly diminished quality of life. Despite the prevalence, persistence, and impact of HIV DNP, little is known of its neurobiological underpinnings.

A related work on HIV-associated sensory neuropathy showed that over half of HIV-infected patients have sensory neuropathy by physical examination or nerve conduction studies. About 40% of them report chronic DNP, while the remainder report only numbness or paresthesia or no symptoms at all [1]. Using procedures described in detail in a [4], the diagnosis of HIV Sensory Neuropathy (HIV-SN) was rendered by physicians and nurses trained in neurological AIDS disorders based on a standardized, neurological examination evaluating HIV-associated sensory neuropathy signs, including diminished ability to recognize vibrations and reduced sharp-dull discrimination in the feet and toes or reduced ankle reflexes. We define at least one sign of neuropathy bilaterally as evidence of HIV sensory neuropathy.

As described previously, HIV DNP was defined as a specific pattern of bilateral burning, aching, or shooting pain in a distal gradient in the lower extremities. Recognizing that this specific pattern of pain may occur in small fiber-predominant neuropathies in which clinical exam abnormalities are sometimes absent due to the

relative paucity of large fiber involvement, we included in the diagnosis of DNP those cases that did not have abnormal clinical exam findings. Indeed, some cases of HIV-SN have been shown to manifest predominantly small fiber involvement. Study clinicians classified DNP into five categories of severity: none, slight (occasional, fleeting), mild (frequent), moderate (frequent, disabling), and severe (constant, daily, disabling, requiring analgesic medication or other treatment). In the statistical model these categories were represented by ordered values: 0=none, 1=slight, 2=mild, 3=moderate, and 4=severe. In the assessments of DNP severity, additional characteristics were elicited including the continuity as well as its impact on daily activities and the need for analgesic medications.

To better understand the correlation between HIV-associated chronic DNP and central nervous system, 241 HIV-infected participants' structural MRI brain images were scanned for this study. Since magnetic resonance (MR) imaging was introduced into clinical medicine and neuroimaging, it has been widely applied in medical diagnosis and treatments. Moreover, MRI is a cutting edge medical imaging technique that has been estimated as an effective tool in the study of the human brain.

## 1.1 Statement of the Problem

Our work in this thesis consists of two separate stages. The first stage research was taken on the association between HIV DNP and the brain volume changes. It was investigated using both linear and nonlinear (Gaussian kernel support vector) multivariate regression analysis, controlling for key demographic and clinical variables.

Generally speaking, regression models are designed to determine how close are the values of the estimated observations to the ground truth values of the target variables, on the basis of training and test data. In general, a training set in machine learning applications is adopted to discover potentially predictive relationships. In regression analysis, particularly, the training set is used to build a model to predict the value of one or more continuous target variables given a vector of input variables. Comparing to the training set, in general sense, test set is used to evaluate the performance of the predictive relationships on a new distinct data set. In regression analysis, the outcomes of the test phase are usually represented by the error rates between the estimates and ground truth. Since the targets of our regression model were already available, we could utilize supervised learning methods. Regression is generally classified by linear regression models and nonlinear regression models. Linear regression models are commonly proposed based on estimation methods, i.e. least squares estimation and maximum likelihood estimation. However, nonlinear kernel regression models and graphical regression models are common nonlinear regression models, widely used to deal with complex data sets.

For medical diagnosis application, the second stage research was focus on the classification of the presence of HIV DNP based on observation of structural MRI brain imaging. Basically a classification system consists of two parts: feature extraction and classification model. An appropriate classification algorithm is definitely crucial for getting good classification results. Nevertheless, the performance of classifiers inevitably also depends on the choice of the features or characteristics of the pattern

selection. Feature extraction is introduced to reduce the dimensionality of data and transform the input data into a set of features. Moreover, it is used to draw out informative and meaningful descriptors of the data set.

However, feature extraction can be divided into dense feature extraction and sparse feature extraction. According to the properties of imaging data, a dense algorithm denotes implied voxel level extraction. In contrast, a sparse algorithm detects regions of interest for feature extraction. Our work centered on 3D MRI images and particularly in this thesis we considered sparsely extracted features.

Unlike regression analysis, which predicts continuous variable outputs, classification analysis is supposed to categorize discrete variable outputs. A classifier is used to identify to which of a set of categories a new observation belongs. In classification analysis, a training set is used for learning to predict and assign the given vector of input variables to one or more discrete, disjoint classes. And the outcomes of the test phase are commonly represented by classification error rates. Sometimes, in order to better understand the strength and utility of a classification model, true positives (sensitivity), true negatives (specificity), false positive, and false negatives are respectively obtained. Since the targets of our classification model were labeled and known, we applied a supervised learning method. Classification could be generally divided to linear classification models and nonlinear classification models. Linear classification models are commonly proposed based on linear discriminant functions and probabilistic theory. However, nonlinear kernel classification models and graphical classification models are common nonlinear classification models, i.e. Gaussian kernel

support vector machine and decision trees.

## 1.2 Organization of Thesis

In chapter 2, basic introduction of the data used in this thesis is firstly reviewed, including feature extraction algorithms regarding to the regression and classification analysis in this thesis. We then present 1) the linear regression and SVM recursive feature elimination (SVM-RFE) algorithms, which explore the correlation between HIV DNP and brain volume changes, 2) the introduction and application of linear discriminant analysis (LDA) classifiers, radial based function support vector machine (SVM) classification algorithm, Naive Bayes classifiers, and random forest classifier in the classification analysis. Chapter 3 presents experimental results for regression analysis and classification analysis and chapter 4 provides a summary, conclusion, and the suggestions for future studies.

# Chapter 2

# METHODOLOGY

In this chapter, basic introduction of the data and some prior research relating to this thesis are reviewed. In addition, feature extraction schemes for regression and classification analysis are carefully discussed. These give us a starting point for the analysis of structural brain imaging. Then, in terms of the methodology in this thesis, key supervised learning concepts such as linear regression, support vector regression (SVR), linear discriminant analysis (LDA), radial based function support vector machine (SVM) classification, Bayes decision classifier, and decision tree algorithm are reviewed.

## 2.1 Introduction of Data

In this section, we briefly discuss the data used in this thesis. Section 2.1.1 firstly demonstrates the subjects and MRI database we obtained. Section 2.1.2 explains the schemes and algorithms we used to acquire the features required by our machine learning models.

### 2.1.1 Subjects and MRI Data

We used T1-weighted CHARTER images[1] that were collected by seven MRI scanning machines of five US academic medical centers participating in the CHARTER. Of 1,556 HIV patients at these five US academic medical centers, 241 underwent structural MRI. The sites performing MRI included: one scanner from Johns Hopkins University (Baltimore, MD, n=47); two scanners from Mount Sinai School of Medicine (New York, NY, n=48); two scanners from University of California at San Diego (San Diego, CA, n=70); one scanner from University of Texas Medical Branch (Galveston, TX, n=46); and one scanner from University of Washington (Seattle, WA, n=30). In terms of the diagnosis of HIV-SN and HIV DNP, we have discussed it in chapter one. Of the 241 participants in this sub-study, 175 had no signs of DNP, 18 had occasional or fleeting DNP, 22 had frequent symptom of DNP, 10 had frequent and disabling DNP signs, and 4 had severe signs of DNP.

### 2.1.2 Features for Regression

In our multivariate regression study of the correlation between HIV DNP and brain volume changes, we investigated the significance of features in the change of cerebral cortical gray matter volume. The target variables of the regression model were

[1] The original data in this study was the structural MRI brain images in the case of 241 HIV-infected participants involved in the CNS HIV Antiretroviral Treatment Effects Research Study (CHARTER). The CNS HIV Anti-Retroviral Therapy Effects Research (CHARTER) study was first funded in September 2002 in response to NIMH RFA 00-AI-0005 to explore the changing presentation of HIV neurological complications in the context of emerging antiviral treatments such as highly active antiretroviral therapy (HAART).

241 participants' log cortical gray matter volumes. The features of the regression model consisted of the demographic and clinical characteristics of the study participants, such as age, ethnicity, gender, cerebral-vault, scanner, history of D-drug use, history of inhalant abuse, history of methamphetamine abuse, Global Deficit Score, DNP, etc.

Table 2.1 shows all the covariates we used in the regression analysis. 41 variables in total were used to estimate structural log cortical gray matter volumes. However, four out of 41 of them were categorical variables. In order to make use of these variables, we converted them into computable variables by creating new binary features from them. For example, variable "Ethnicity" consists of four categories: Caucasian, African American, Hispanic and other, which correspond to numerical values {"1", "2", "3", "4"}. We therefore split "Ethnicity" into three binary feature sets. For a single data point, it has value "1" of only one of three feature sets and "0" otherwise. Similarly, variable "Scanner" represents six scanner machines, which correspond to numerical values {"1", "2", "3", "4", "5", "6"}. We therefore split "Scanner" into six binary feature sets. For a single data point, it has value "1" of only one of three feature sets and "0" otherwise. Variable "D-drug ever" consists of three categories: current, past, and never, which correspond to numerical values {"1", "2", "3"}. We therefore split "D-drug ever" into two binary feature sets. For a single data point, it has value "1" of only one of three feature sets and "0" otherwise. Variable "Antiretroviral Regimen" consists of five categories: None, PI-based[2], NNRTI-based[3], PI+NNRTI-based and other, which

[2] Protease Inhibitor

[3] Non-nucleoside reverse-transcriptase

correspond to numerical values {"1", "2", "3", "4", "5"}. We therefore split "Antiretroviral Regimen" into four binary feature sets. For a single data point, it has value "1" of only one of three feature sets and "0" otherwise. Consequently, 52 features in total were obtained to establish the regression model. We named this data set *regression.dat.*

Table 2.1 List of the explanatory variables used for the multivariable models of the association between DNP and log brain volumes. The types of each explanatory variable are listed.

| FEATURE | TYPE | FEATURE | TYPE |
|---|---|---|---|
| Distal Neuropathic Pain | Continuous | Alcohol Abuse Ever | Binary |
| Age | Continuous | Alcohol Dep Ever | Binary |
| Education | Continuous | Cannabis Abuse Ever | Binary |
| Ethnicity | Categorical | Cannabis Dependence Ever | Binary |
| Gender | Binary | Cocaine Abuse Ever | Binary |
| Log-cerebralVault | Continuous | Cocaine Dep Ever | Binary |
| Scanner | Categorical | Halucinogen abuse | Binary |
| HCV status | Binary | Halucinogen Dependence | Binary |
| Plasma HIV RNA | Binary | Inhalant Abuse Ever | Binary |
| CSF HIV RNA | Binary | Inhalant Dependence Ever | Binary |
| Sqrt CD4 Nadir | Continuous | Methamphetamine Abuse | Binary |
| Sqrt CD4 Current | Continuous | MethamphetaimeDependence | Binary |
| Current D-drug Exp | Continuous | Opiate Abuse Ever | Binary |
| On D-drugs (Y/N) | Binary | Opiate Dependence Ever | Binary |
| D-drug ever | Categorical | Sedative Abuse Ever | Binary |
| Total D-drug Exposure | Continuous | Sedative Dependence Ever | Binary |

| | | | |
|---|---|---|---|
| # protease inhibitors | Continuous | Global Deficit Score | Continuous |
| Antiretroviral Regimen | Categorical | Opiate Pain Treatment | Binary |
| Beck Depression | Continuous | TAP Treatment | Binary |
| Major Depression | binary | Anticonvulsant Pain Treatment | Binary |
| Major Depression Ever | binary | | |

### 2.1.3 Features for Classification

Given the structural MRI images from 241 HIV study participants, the following classification study involved the presence or absence of HIV DNP analysis. Unlike the demographic and clinical features in regression models, the classification models were obtained based on extracting features from MRI brain images. So the reliability of feature extraction inherently is need to capture the intrinsic information in the MRI imaging. After clinical evaluation by bioimaging specialists, eight out of 241 HIV participants' brain scanning images were diagnosed as abnormal brains[4] and they were carefully excluded and removed by the neuroimaging experts.

We then used only 233 investigable HIV-infected patients' standard space MRI brain images in classification analysis. And the target variables in the classification models were certainly intended to be the presence of DNP. Based on the data profile, the ground truths of 233 patients' pain conditions were categorized into {"0", "1", "2", "3", "4"} with respect to the severity of DNP. Intuitively, pain condition "0" was automatically labeled as "0", and all pain conditions other than "0" in general were

[4] Abnormalities in this case include the presence of brain tumors, brain lesions, enlarged ventricles, etc.

labeled as “1” in our experiments. Considering this class separation principle, among these 233 patients, there are 169 patients with DNP symptoms and 64 patients with no DNP symptoms.

Using these 233 MRI subjects as described above, three separate experiments were performed with a goal of classifying/diagnosing binary DNP conditions. The fundamental idea of the three experiments was the same. However, what distinguished them was the feature subsets (models) we selected to perform the classification analysis. In order to investigate and obtain more effective and significant information from the image data, we tended to gradually increase the dimensionality of the feature space in our experiments. We started from volume space feature extraction, and then extended to area space feature extraction and to sub-regional area space feature extraction. We therefore proposed three separated classification experiments, with the feature dimensionality escalated, from one to the next.

#### 2.1.3.1 Features for Volume Space Classification

To begin with, the first experiment is a study regarding volume space features of brain images and how they predicted the presence of DNP in individuals diagnosed as HIV-infected. By the CHARTER study, the MRI brain images were manually identified to cerebral cortex[5] and cerebral sub-cortex. Regardless of the structure of human brain,

[5] The cerebral cortex is the outer covering of gray matter over the hemispheres. This is typically 2- 3 mm thick, covering the gyri and sulci. Certain cortical regions have somewhat simpler functions, termed the primary cortices. These include areas directly receiving sensory input (vision, hearing, somatic sensation) or directly involved in production of limb or eye movements [5].

cerebrum could be segmented into gray matter (grey matter), white matter and cerebrospinal fluid (CSF) from the perspective of anatomical components. CHARTER study segmented gray matter, white matter and CSF by hand in FSLView[6]. In addition, the structural volumes of the three components were computed in both cortical brain and sub-cortical brain and eventually six volumes were chosen for the study. Furthermore, all structural volumes were log transformed to symmetrize the distributions and stabilize the variances. Table 4.2 shows the names of six log-volumes and their mean, and standard deviation values.

Table 2.2 Classification features in volume space. The mean and standard deviation values are listed.

| Volume (log transformed) | Num. Subj. | Mean | Stdev |
|---|---|---|---|
| Cortical Gray Matter | 233 | 13.299 | 0.139 |
| Subcortical Gray Matter | 233 | 10.462 | 0.104 |
| Abnormal White Matter | 233 | 8.935 | 0.363 |
| Total White Matter | 233 | 13.077 | 0.147 |
| Ventricular CSF | 233 | 9.857 | 0.554 |
| Sulcal CSF | 233 | 11.175 | 0.603 |

[6] FMRIB Software Library (FSL) [6] [7] is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging data. FSL is available as both precompiled binaries and source code for Apple and PC (Linux) computers. It is freely available for non-commercial use. FSLView is one of the self-installed tools in FSL, which provides visual viewing of MRI data.

Using 6 log-volume variables from the data profiles as the feature subset, three variables were controlled in this experiment: Age, Ethnicity, and Gender. The individuals diagnosed with HIV in this research were both male and female with the youngest age being 23 and the oldest age being 67 from 4 ethnicities. Male and female were separated and classified by age with other patients within roughly a decade of the same age (below age 34, ages 35-44, ages 45-55, and ages 55 above). And then African American patients were separated from other races. These 16 groupings' brain region volumes were averaged and then the group-averages were subtracted from the individual volume estimations of each group to arrive at a set of detrended (by group averaging) values. However, the feature set version without detrending was used as a comparison experiment. We entitled the data set with detrending process *volume_norm.dat* and the data set with no detrending process *volume.dat*. In addition, instead of detrending the feature vectors using group averaging, we also inserted variable age, variable gender, and variable ethnicity into the data set and created the third volume space data set entitled *volume_plus.dat.*

### 2.1.3.2 Features for Area Space Classification

In the second experiment, we increased the dimensionality of the feature subsets. To do so, some image processing techniques were required on 233 standard-space $T_1$-weighted images. Although the most sufficient information in the standard-space $T_1$-weighted images implicitly exists in the voxel space features, high dimensional feature sets will bring about considerably low computational efficiency. As a result, we

compromised to conduct classification studies based on area space features by slicing the MRI brain images. In addition, some recent neuroimaging studies have revealed several correlated brain regional activities based on the changes of brain volumes. They provided us convincing evidence to remove uncorrelated brain anatomy from the whole brain images. In such a way, the dimensionality of features was further reduced.

Over 30 independent studies in distinct chronic pain conditions have suggested gray matter decrease positively correlates with chronic pain [8]. Neuropathic pain, as one type of chronic pain conditions, was usually diagnosed based on clinical signs. A recent study demonstrated that patients exhibited a gray matter decrease in the insula, ventromedial prefrontal cortex, and nucleus accumbens [9]. In contrast, neuronal activation patterns in primary headache studies showed the areas known to be generally involved in pain processing: the cingulate, insular cortex, and thalamus. Since most of the subcortical regions were proved to be uncorrelated to the presence of DNP, we decided to eliminate the effects of brainstem and cerebellum in classification analysis.

Image preprocessing was performed using the FSL 4.1 in Ubuntu 13.04 system and MATLAB 8.0 in Windows. The input to our processing was a standard-space $T_1$-weighted image in NifTI[7] format. First, using a precise gray matter mask carefully created by fslmaths in FSLutils[8], we roughly removed 99% of tissues other than gray matters. Next, we tended to remove two useless subcortical brain tissues (cerebellum and

[7] NIfTI is adapted from the widely used ANALYZE™ 7.5 file format. The primary goal of NIfTI is to provide coordinated and targeted service, training, and research to speed the development and enhance the utility of informatics tools related to neuroimaging.

[8] FSLUTILS is a set of useful command-line utilities which allow the conversion, processing etc. of Analyze and Nifti format data sets.

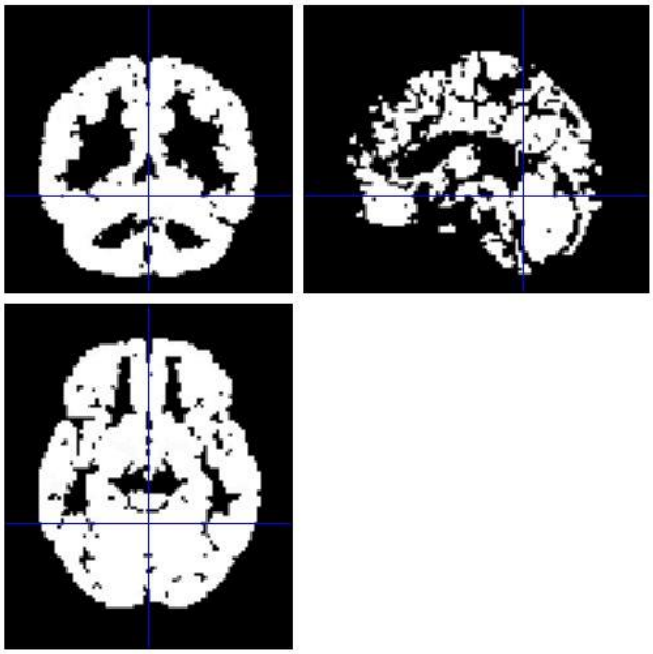
brainstem) from the whole brain image, using the FSL masking algorithm. In FSLView, the electronic atlas MNI structural atlas[9] has cerebellum and brainstem masks for standard space brain images. We therefore used these automated masking techniques and gained a segmented cerebellum image and a brainstem image.

In order to strip cerebellum and brainstem from the whole brain, we outputted the whole brain image, cerebellum image and brainstem image to the processing in MATLAB. First, the program MRIcro[10] was chosen to convert the NifTI file into an equivalent file type pair '.img' and '.hdr'. Next, we used an MRI toolbox to import the MRI '.img' data to our image processing algorithm. We then implemented a linear subtraction algorithm to wipe out cerebellum and brainstem regions, such that only the regions we are interested in remained in the image. Figure 2.1 shows the original standard-space $T_1$ - weighted image. Figure 2.2 shows the image preprocessing outcomes in MATLAB.

[9] MNI structural atlas is one of the eleven templates and atlases with FSL. See: http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases

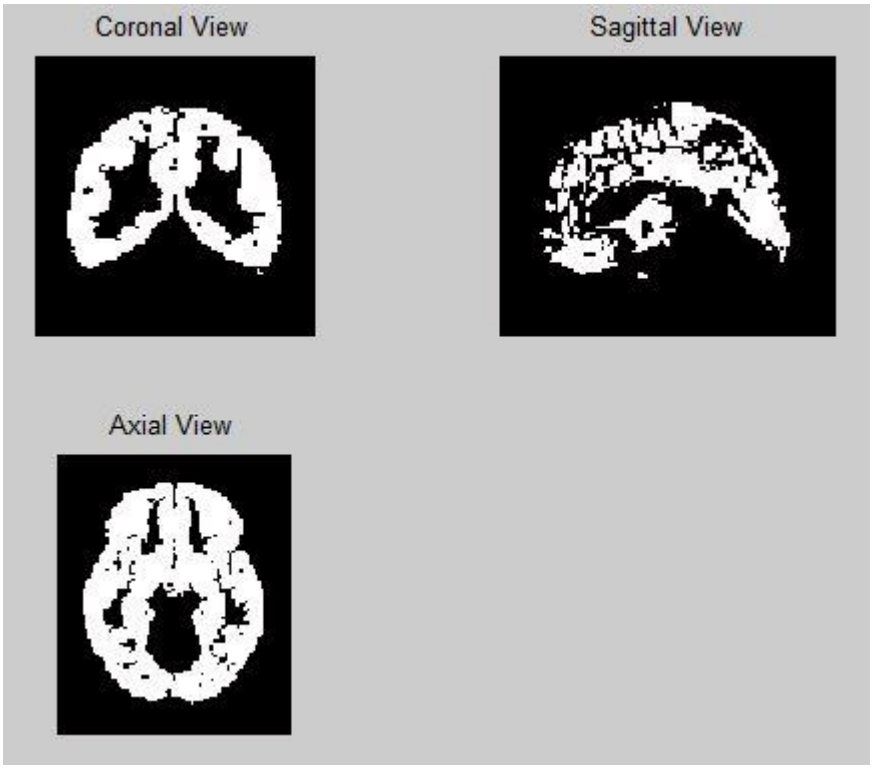
[10] MRIcro allows Windows and Linux computers view medical images. It is a standalone program, but includes tools to complement SPM (software that allows neuroimagers to analyse MRI, fMRI and PET images). MRIcro allows efficient viewing and exporting of brain images. In addition, it allows neuropsychologists to identify regions of interest (ROIs, e.g. lesions). MRIcro can create Analyze format headers for exporting brain images to other platforms. More information here

Figure 2.1 An example of original standard-space $T_1$- weighted image (patient ID: RA029006, image dimensionality: 91x109x91) displayed by SPM8[11]. Upper left: coronal view (front view of head); Upper right: sagittal view (right side view of head); Lower left: axial view (top view of head). Cursor position: X: 46; Y: 40; Z: 31.



[11] SPM (Statistical Parametric Mapping) refers to the construction and assessment of spatially extended statistical processes used to test hypotheses about functional imaging data. SPM8 (SPM version 8) provided a relatively simple interface while also being a software package written for use with MATLAB.

Figure 2.2 An example of processed standard-space $T_1$- weighted image (patient ID: RA029006, image dimensionality: 91x109x91) displayed by MATLAB. Cerebellum and brainstem tissues in the brain have been successfully stripped out (Cursor position: X: 46; Y: 40; Z: 31, the same position as Figure 2.1).




In terms of the anatomy of brain, MRI images are displayed according to three different 3-D views of the target tissues (see Figure 2.1), each a sequence of 2-D images. Using these processed three-view MRI images (see Figure 2.2), we intensively sliced the standard-space $T_1$- weighted images in three directions. For the coronal view, we had a sequence of 109 2-D images/slices. For all 109 2-D slices, the areas of cortex regions were computed and the 87 non-zero values were used as features for each patient/sample. For the sagittal view, we had a sequence of 91 2-D images/slices. For all 91 2-D slices,

the areas of cortex regions were computed and the 74 non-zero values were used as features for each patient/sample. For the axial view, we had a sequence of 91 2-D images/slices. For all 91 2-D slices, the areas of cortex regions were computed and the 65 non-zero values were used as features for each patient/sample. As a result, 226 features of slices area have been acquired. Among them, all the cortical gray matter regions and the region of interest (ROI) in the subcortical gray matter were collected.

With the same idea of group-averaging scheme in the first experiment (section 2.1.3.1), a feature set with detrending values was used in the second experiment. However, the feature set version without detrending were used as comparison experiment. We entitled the data set with detrending process *slicearea_norm.dat* and the data set with no detrending process *slicearea.dat*. In addition, instead of detrending the feature vectors using group averaging, we also inserted variable age, variable gender, and variable ethnicity into the data set and created the third area space data set entitled *slicearea_plus.dat.*

### 2.1.3.3 Features for Sub-regional Area Space Classification

In the third experiment of the classification analysis, we further expanded the dimensionality of the area-space feature subset. The idea was straight forward based on the feature extraction algorithm in section 2.1.3.3. Sub-regional slices areas were retrieved using a grid segmentation on the 2-D images/ slices. As showed in Figure 2.2, we used the stripped images without cerebellum and brainstem brain. We then equally cut the images crosswise into four sub-regions in all three views. Therefore, instead of

an individual area space feature extracted from a slice, four area space features were computed and acquired from that. As a result, roughly four times larger size of feature set has been processed to filter out the zero area values, and we intuitively obtained 878 features of the sub-regional slices areas.

With the same idea of group-averaging scheme in the first experiment (section 2.1.3.1), a feature set with detrending values was used in the second experiment. However, the feature set version without detrending were used as comparison experiment. We entitled the data set with detrending process *subslicearea_norm.dat* and the data set with no detrending process *subslicearea.dat*. In addition, instead of detrending the feature vectors using group averaging, we also inserted variable age, variable gender, and variable ethnicity into the data set and created the third sub-regional area space data set entitled *subslicearea_plus.dat.*

## 2.2 Regression by Linear Discriminant Functions

The simplest form of linear regression models are linear functions of the input variables. However, we can obtain a much more useful class of functions by taking linear combinations of a fixed set of nonlinear functions of the input variables, known as basis functions. Such models are linear functions of the parameters, which gives them simple analytical properties. Although linear models have significant limitations as practical techniques for pattern recognition, particularly for problems involving input spaces of

high dimensionality, they have nice analytical properties and form the foundation for more sophisticated models.

### 2.2.1 Linear Basis Function Models

The simplest multiple variables regression model is one that involves a linear combination of the input variables

$$y(\mathbf{x}, \boldsymbol{\omega}) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n$$

where $\mathbf{x} = \begin{bmatrix} x_0 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \epsilon\, R^{n+1}$, and $\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_2 \\ \vdots \\ \omega_n \end{bmatrix} \epsilon\, R^{n+1}$. Let $n$ to be the number of features involved in the regression model, we then have a linear combination

$$y(\mathbf{x}, \boldsymbol{\omega}) = [\omega_0\ \omega_1 \cdots\ \omega_n] * \begin{bmatrix} x_0 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \boldsymbol{\omega}^{\mathbf{T}}\mathbf{x} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_{n.}$$

For convenience of notation, we usually define $x_0 = 1$. Given an input data sample $\mathbf{x}$ with $n + 1$ features, we estimate the parameter matrix $\boldsymbol{\omega}$ and target variable $y(\mathbf{x}, \boldsymbol{\omega})$. This is often simply known as linear regression. Now we extend this simple regression model to multiple training sample space. Such that, $m$ dimensional input variables with $n$ features in the form

$$\mathbf{X} = \begin{bmatrix} x_1^0 & \cdots & x_1^n \\ \vdots & \ddots & \vdots \\ x_m^0 & \cdots & x_m^n \end{bmatrix} \epsilon\, R^{m\times(n+1)}$$

where $x_j^{(i)}$ denotes value of feature $i$ in $j^{th}$ training sample. In order to figure out the best fit estimate the target variable $t = [t_1\ t_2 \cdots\ t_m]^{\mathrm{T}}$, we investigate the parameter matrix $\boldsymbol{\omega}$ by minimizing a sum of squares error function. However, this error function could be motivated as the maximum likelihood and least squares approach. In terms of Sum-of-Square error (SSE), given the equation of 1this error as,

$$E_D(\boldsymbol{\omega}) = \frac{1}{2}\|t - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{X}^{\mathbf{T}}\|^2.$$

Given this SSE function, we can take gradient of that with respect to $\boldsymbol{\omega}$. We then set this gradient to zero and solve for $\boldsymbol{\omega}$ to obtain

$$\boldsymbol{\omega} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}t.$$

Given a new set of test data $\widehat{\mathbf{X}}$ to this regression model with the parameter vector $\boldsymbol{\omega}$, the outcomes $\hat{y}$ can be linearly estimated.

### 2.2.2 Linear Models for Feature Ranking

In statistical analysis, linear regression models are prevalently used to investigate the practical statistical problems, for instance, statistical hypothesis tests [10]. A t test statistic as a type of hypothesis test, can be used to determine if two sets of data are significantly different from each other. In particular, in terms of the parameters ω's, $\omega_i$ can be defined as the change of target value when $x^i$ increases by 1 unit, all other features already fixed. A t-test statistic on $\omega_i$ is a measurement of how much evidence we have to reject the null hypothesis[12]. Alternatively, there is another measurement to identify whether the null hypothesis holds. In statistical significance testing, the *p*-value is the probability of seeing something more extreme than the given observation if the null hypothesis is true. If this p-value is less than the significance level previously set, it rejects the null hypothesis and supports the alternative hypothesis. Apparently, the smaller the *p*-values are, the more evidence we have against $H_0$.

[12] It can be considered as a null hypothesis $H_0: \omega_i = 0$ and an alternative hypothesis $H_a: \omega_i \neq 0$. We interpret the hypothesis as a test measuring the linear relationship between the target variable and feature $x^i$. The null hypothesis indicates that there is no linear relationship between them.

We used the concept of t-statistic and *p*-values to conduct a linear regression model selection/ refinement. Regardless of sufficient information provided by high dimensional feature subsets, more is not always better according to the curse of dimensionality. This term was originally coined by Richard E. Bellman when considering problems in dynamic optimization. When the number of predictors is large relative to the data sample size, we tend to observe either one or more of these predictors are not significantly related to the target variable or two or more of the predictors are related to each other. In either case, we would prefer to drop unnecessary predictors.

The methods commonly used in the linear regression area are stepwise procedures [13]. Accordingly, we used stepwise backward elimination in our linear regression analysis. Generally, we assumed that we had one target/ response variable $t$ (the log cortical gray matter volume) and a pool of the demographic and clinical characteristics covariates $\mathbf{x} = [x_1\ x_2 \cdots\ x_m]^{\mathrm{T}}$. And an elimination criterion was set by specifying a level of significance $\alpha$. First, we fitted the model with all predictors/ covariates (we show the regression results in chapter 4). If all predictors are significant (*p*-values less than $\alpha$), stop and retain all predictors. However, if any are not significant, remove the least significant one (the predictor indexed by the largest *p*-values). Second, if one predictor was removed in the first step, we refit the model without that predictor and repeat the filtering procedure in the first step. We then repeatedly replicate the elimination steps until all the *p*-values in the regression model are less than $\alpha$. We

[13] Stepwise model selection procedures include stepwise backward selection/ elimination, stepwise forward selection/ elimination and the combination of stepwise forward and backward elimination.

discussed the features used for the regression analysis in section 2.1.2. If we manually controlled the elimination steps, we stop high at top 10 significant features remaining in the model. However, without human intervention, the stepwise backward elimination algorithm automatically stopped at top 13 significant features remaining in the model, which means 13 features can be considered to have correlations with the target variable. We implemented the linear regression model selection processing using Minitab[14].

## 2.3 Classification by Linear Discriminant Functions

In the previous section, we explored a class of linear regression models having particularly simple analytical and computational properties. We now discuss an analogous class of linear models for solving classification problems. The objective in classification is to take an input vector **x** and to allocate it to one of K discrete, disjoint classes $C_k$, where $k = 1, \ldots, K$. with each input assigned to one and only one class.

### 2.3.1 Linear Discriminant Analysis (LDA)

The input space is thereby divided into decision regions whose boundaries are called decision boundaries or decision surfaces, by which we mean boundaries of decision regions in the input space. In linear classification analysis, data sets whose classes can be separated exactly by linear decision surfaces are said to be linearly separable. Similar to the definition of target variables in linear regression models, we use

[14] Minitab 16 statistical software, Minitab Inc., 2010, www.minitab.com. Minitab is a statistics package developed at the Pennsylvania State University by researchers Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner in 1972. The newest version is Minitab 17.

target values to represent class labels. Moreover, in the linear regression models, considered in section 2.2, the model prediction $y(\mathbf{x}, \boldsymbol{\omega})$ was given by a linear function of the parameters $\boldsymbol{\omega}$**.** Returning to the simplest case, the model is also linear in the input variables and therefore takes a simplest linear discriminant function $y(\mathbf{x}, \boldsymbol{\omega}) = \boldsymbol{\omega}^{\mathbf{T}}\mathbf{x} +$ b, where $\boldsymbol{\omega}$ is called a weight vector and b is a bias parameter. In particular binary classification case, the decision boundary in this case is $y(\mathbf{x}, \boldsymbol{\omega}) = 0$, an input vector $\mathbf{x}$ is assigned to class $C_1$ if $y(\mathbf{x}, \boldsymbol{\omega}) \geq 0$ and to class $C_2$ otherwise.

There are three methods to learning the parameters of linear discriminant functions, based on least squares, Fisher's linear discriminant, and the perceptron algorithm. The ideas of least squares or maximum likelihood were already mentioned in section 2.2. Intuitively, we saw that the minimization of a sum of squares error function led to simple closed-form solution for the parameter values. Thus, it is a straightforward attempt to apply the same formalism to classification problems. The advantage of least squares approach is this is an effective and efficient approximation algorithm to implement. Yet, the problems of least squares approach are they lack robustness to outliers and Maximum likelihood, e.g. a Gaussian conditional distribution may or may not be a good model assumption for the data.

The basic idea of Fisher's linear discriminant is to maximize a function that will give a large separation between the projected classes means while also give small variance within each class. By contrast, the least-squares approach to the determination of a linear discriminant was to make the predictions as close as possible to the target values. Actually, the Fisher criterion can be obtained as a special case of least squares.

In terms of maximizing class separation, we attempt to maximize inter-class projected mean and simultaneously minimize the intra-class variance. Hence the solution of that is to maximize the ratio of them. Considering a two-class problem in which there are $N_1$ points of class $C_1$ and $N_2$ points of class $C_2$, so that the mean vectors of the two classes are given by

$$\mathbf{m_1} = \frac{1}{N_1}\sum_{n\in C1} x_n, \qquad \mathbf{m_2} = \frac{1}{N_2}\sum_{n\in C2} x_n.$$

Suppose we take the input vector **x** and project it down to one dimension using $y = \boldsymbol{\omega}^{\mathbf{T}}\mathbf{x}$**.** The simplest measure of the separation of the classes, when projected onto $\boldsymbol{\omega}$, is the separation of the projected class means.

$$m_1 - m_2 = \boldsymbol{\omega}^{\mathbf{T}}(\mathbf{m_1} - \mathbf{m_2}).$$

The projection function $y = \boldsymbol{\omega}^{\mathbf{T}}\mathbf{x}$ transforms the set of labelled data points in **x** into a labelled set in the one-dimensional space *y*. The within-class variance of the transformed data from class $C_k$ is therefore given by

$$S_k^2 = \sum_{n\in Ck} \{y_n - m_k\}^2.$$

Now the Fisher criterion is defined to be maximizing the ratio of the between-class variance to the within-class variance

$$J(\boldsymbol{\omega}) = \frac{(m_1 - m_2)^{\mathbf{2}}}{{S_1}^2 + {S_2}^2}.$$

By some basic rules of matrix computation, we could rewrite the Fisher criterion in the form

$$J(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^{\mathbf{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathbf{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{\omega}}$$

where $\boldsymbol{S}_{\mathrm{B}}$ is the between-class covariance matrix with the form $\boldsymbol{S}_{\mathrm{B}} = (\mathbf{m_1}-\mathbf{m_2})(\mathbf{m_1}-\mathbf{m_2})^{\mathrm{T}}$ and $\boldsymbol{S}_{\mathrm{w}}$ is the within-class covariance matrix , denoted by

$$\boldsymbol{S}_{\mathrm{w}} = \sum_{n\in C1} (\mathbf{x_n}-\mathbf{m_1})(\mathbf{x_n}-\mathbf{m_1})^{\mathrm{T}} + \sum_{n\in C2} (\mathbf{x_n}-\mathbf{m_2})(\mathbf{x_n}-\mathbf{m_2})^{\mathrm{T}}.$$

We take the derivation of $J(\boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ and $J(\boldsymbol{\omega})$ is maximized when $(\boldsymbol{\omega}^{\mathbf{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{\omega})\boldsymbol{S}_{\mathrm{w}}\boldsymbol{\omega} = (\boldsymbol{\omega}^{\mathbf{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{\omega})\boldsymbol{S}_{\mathrm{B}}\boldsymbol{\omega}$. We then simply it by multiplying $\boldsymbol{S}_{\mathrm{w}}{}^{-\mathbf{1}}$ on both sides and such that $\boldsymbol{\omega}$ is obtained as a proportion of the difference of the class means

$$\boldsymbol{\omega} \propto \boldsymbol{S}_{\mathrm{w}}{}^{-\mathbf{1}}(\mathbf{m_2}-\mathbf{m_1}).$$

### 2.3.2 LDA for Feature Selection

In terms of the application of LDA in this thesis, we made use of a Fisher's LDA classifier to implement the feature selection for the classification experiments of DNP diagnosis. In such a way, we linearly reduced the feature subset dimensionality and complexity. Feature subset selection techniques are commonly designed to find a reliable range of tradeoffs between accuracy and data complexity [12]. The basic idea of this subset selection is to select a subset of existing features based on some pre-defined criteria. "Filtering" methods select features independent to classifier training, based on evaluation of discrimination power for individual features or small feature groups, e.g. information theory based measurements, variance ratio (VR) and average variance ratio (AVR). "Wrapper" methods are generally more reliable to solve the discrimination problems using some classifiers based on some criterion such as classification rate. Regardless of the statistical backward feature selection we discussed in section 2.2, practical wrapper methods inherently involve greedy heuristic search such as sequential

feature selection and classifier design steps, with features sequentially selected to maximize the current subset's joint discrimination power. Two basic components in classical feature selection algorithms are a selection criterion and a stopping criterion. In our study, we used a LDA classifier (Fisher's LDA) which we obtained an evaluation to measure the goodness of the feature subsets. To be precise, the optimization criterion exclusively measured in our HIV DNP study was focus on the rate of true positive (sensitivity) due to the unbalanced characteristic of target values. The stopping condition was governed by a sequential backward selection algorithm, which starts from the full set and removes features sequentially. We implemented 5-fold cross validation [13] as a quantitative validation to improve the feature elimination accuracy by remove the features indexed by the worst mean sensitivity results.

## 2.4 Classification by Nonlinear Kernel Method and SVMs

In section 2.2, we considered linear parametric models for regression in which the form of the mapping $y(\mathbf{x}, \boldsymbol{\omega})$ from the input $\mathbf{x}$ to the output $y$ is fitted by a vector $\boldsymbol{\omega}$ of adaptive parameters. This approach can be also used in nonlinear parametric model. The linear parametric models can be reformulated in term of an equivalent dual representation [11] in which the kernel function intrinsically evaluated. Although a couple of decades passed, the kernel idea was reintroduced by Boser, Guyon and V. Vapnik in 1992 giving rise to the technique of support vector machines (SVM). Since then SVM became widely prevalent many years ago for solving problems in regression and classification analysis.

### 2.4.1 Kernel Method and SVM

The kernel function is given by the linear combination of the feature space mapping function $\Phi(\mathbf{x})$

$$k(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{x}').$$

The kernel concept was introduced into the field of pattern recognition by Aizerman et al in 1964. The kernel function is a generalization of the distance (similarity) metric; it measures the similarity between two expression vectors as the data are projected into a higher-dimensional space. The simplest case of a kernel function is referring to the linear kernel function, such that $\Phi(\mathbf{x}) = \mathbf{x}$**.** In terms of more complex specialization involves radial basis function/ Gaussian kernel, the nonlinear kernel method is introduced in which is commonly used in the form of

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right).$$

The support vector machine approaches this problem through the concept of the margin. Margin as an extremely important concept, is defined as the smallest/ minimum distance between the decision boundary and any of the samples. Maximizing the margin causes a particular choice of decision boundary (hyper-plane). The location of this boundary is determined by a subset of the data points, known as support vectors. The hyper-plane is increasingly dominated by the nearby data points. In the limit, the hyper-plane becomes independent of data points that are not support vectors. Generally, support vector machine constructs a hyper-plane which maximizes the margin. Intuitively, we return to the binary classification problem with the linear model in the form of

$$\text{Hyper-plane: } y(\mathbf{x}) = \boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x}) + \mathrm{b}$$

where $\Phi(\mathbf{x})$ denotes a fixed feature-space transformation and $b$ the bias parameter. We then recall the concept of target variables $t$ ($t_m = \pm 1$ in binary classification models, where $m$ denotes a certain training data point). Start by assuming the training data is linearly separable in the feature space, i.e. there exists one choice of parameters $\boldsymbol{\omega}$ and b such that the hyper-plane function satisfies $y(\mathbf{x_m}) > 0$ for points having $t_m = +1$. And on the other hand, the hyper-plane function satisfies $y(\mathbf{x_m}) < 0$ for points having $t_m = -1$. Therefore, $t_m y(\mathbf{x_m}) > 0$ for all training data points, in which we normalize it as $t_m(\boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x_m}) + \mathrm{b}) \geq 1$.

Recall that the perpendicular distance of a point $\mathbf{x_m}$ from a hyper-plane defined by $y(\mathbf{x_m}) = 0$ where $y(\mathbf{x_m})$ takes the hyper-plane function is given by $|y(\mathbf{x_m})|/\|\boldsymbol{\omega}\|$ Furthermore, we are only interested in solutions for which all data points are correctly classified, so that $t_m y(\mathbf{x_m}) > 0$ for all data points. Thus the distance of a point $\mathbf{x_m}$ to the decision surface is given by

$$\frac{t_m y(\mathbf{x_m})}{\|\boldsymbol{\omega}\|} = \frac{t_m(\boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x_m}) + \mathrm{b})}{\|\boldsymbol{\omega}\|}.$$

The margin is given by the perpendicular distance to the closest point $\mathbf{x_m}$ from the data set, and we wish to optimize the parameters $\boldsymbol{\omega}$ and b in order to maximize this distance. Thus the maximum margin solution is found by solving

$$\mathrm{argmax}_{\boldsymbol{\omega},\mathrm{b}}\left\{\frac{1}{\|\boldsymbol{\omega}\|}\min[t_m(\boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x_m}) + \mathrm{b})]\right\}$$

where $t_m(\boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x_m}) + \mathrm{b})/\|\boldsymbol{\omega}\|$ is the distance between a data point $\mathbf{x_m}$ to the decision surface. The optimization problem simply requires that we maximize $\|\boldsymbol{\omega}\|^{-1}$, which is equivalent to minimizing $\|\boldsymbol{\omega}\|^2$, and so we have to solve the optimization problem

$$\text{argmax}_{\boldsymbol{\omega},\mathrm{b}} \frac{1}{2}\|\boldsymbol{\omega}\|^2.$$

In order to solve this constrained optimization problem, we introduce Lagrange multipliers $a_m \geq 0$, giving the Lagrange function

$$\boldsymbol{L}(\boldsymbol{\omega}, \mathrm{b}, \mathbf{a}) = \frac{1}{2}\|\boldsymbol{\omega}\|^2 - \sum_{m=1}^{M} a_m\{t_m(\boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x_m}) + \mathrm{b}) - 1\}$$

where $\mathbf{a} = (\mathrm{a}_1, \ldots, \mathrm{a}_m)^{\mathrm{T}}$**.** Even though there is a minus sign in front of the Lagrange multipliers, we could still solve for the optimization function by maximizing the values of the Lagrange multipliers. Now we take the derivation of the Lagrange function with respect to $\boldsymbol{\omega}$ and b and set it to zero. We then obtain that

$$\boldsymbol{\omega} = \sum_{m=1}^{M} a_m t_m \Phi(\mathbf{x_m}), \qquad \sum_{m=1}^{M} a_m t_m = 0.$$

So when we eliminate $\boldsymbol{\omega}$ and b using two representations we obtained, we just need to maximize

$$\tilde{L}(\mathbf{a}) = \sum_{m=1}^{M} a_m - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M} a_m a_n\, t_m t_n k(\mathbf{x_m}, \mathbf{x_n})$$

here we introduce the kernel function $k(\mathbf{x_m}, \mathbf{x_n})$ is defined to be $k(\mathbf{x_m}, \mathbf{x_n}) = \Phi(\mathbf{x_m})^{\mathrm{T}}\Phi(\mathbf{x_n})$. To classify new data points using the training model, we reformulate $y(\mathbf{x})$ and replace $\boldsymbol{\omega}$ by expressing $a_m$ and the kernel function

$$y(\mathbf{x}) = \sum_{m=1}^{M} a_m t_m k(\mathbf{x}, \mathbf{x_m}) + \mathrm{b}.$$

Recall the constraints $t_m(\boldsymbol{\omega}^{\mathbf{T}}\Phi(\mathbf{x_m}) + \mathrm{b}) \geq 1$, we now demonstrate a constrained optimization of this form satisfies the *Karush-Kuhn-Tucker* (KKT) conditions with three properties as following:

$$a_m \geq 0$$

$$t_m y(\mathbf{x_m}) - 1 \geq 0$$

$$a_m(t_m y(\mathbf{x_m}) - 1) = 0.$$

It means for all data points, either $a_m = 0$ or $t_m y(\mathbf{x_m}) - 1 = 0$ holds. Regardless of how large the number of $a_m = 0$ points is, they are literally not supposed to be constructive predictions for new data points. On the other hand, those data points who satisfies $t_m y(\mathbf{x_m}) = 1$ are called support vectors lie on the maximum margin hyperplanes in feature space. SVM therefore is a type of sparse kernel machine.

For the purpose of maximizing the margin while softly penalizing points that lie on the wrong side of the margin boundary, we introduce slack variables $\xi_m$ to provide a formalism where data points are allowed to be on the 'wrong side' of the margin boundary, but with a penalty that increases with the distance from that boundary. We therefore minimize

$$C \sum_{m=1}^{m} \xi_m + \frac{1}{2} \|\boldsymbol{\omega}\|^2$$

where the cost parameter $C > 0$ balances the trade-off between the slack variables penalty and the margin. Moreover, if we include the Lagrange multipliers $a_m$ and KKT conditions analysis, we obtain a box constraint $0 < a_m < C$. We therefore denote a soft margin SVM called C-SVM.

### 2.4.2 Radial Basis Function Kernel C-SVM for Classification

In terms of the application of C-SVM classifier in this thesis, we implemented a weighted RBF kernel SVM classifier to determine the DNP presents or not. Recall that a Gaussian kernel is defined to be represented as

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right).$$

$\|\mathbf{x} - \mathbf{x}'\|^2$ should be recognized to be the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$. We thereby conclude that Gaussian kernel is a RBF kernel. Moreover, we replace $1/2\sigma^2$ by a kernel hyper-parameter $\gamma$, such that the RBF (Gaussian) kernel could be rewritten as

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2\right).$$

When we applied RBF kernel in C-SVM in our experiments, we had two hyper-parameters: regularization constant *C* and kernel parameter $\gamma$. They are required hyper-parameters optimization, which led to a grid search algorithm. Hence, to perform a grid search, we selected a set of reasonable values for each, where $C \in [2: 25]$ and $\gamma \in [-23: 0]$. Grid search trained an C-SVM with a pair of hyper-parameters ($C$, $\gamma$) in the cross product of these two sets. We investigated the highest sensitivity scores from the grid search on the training data set only. Finally, we test how well the methods generalize to new data using the unbiased test data.

## 2.5 Regression by Nonlinear Kernel Method and $\epsilon$-SVR

Now we extend support vector machine to regression problems. As we known, instead of taking an input vector **x** and assign it to K discrete and disjoint classes, we estimate the target values and minimize the error function in the regression. A version of SVM for regression was proposed in 1996 by Vladimir N. Vapnik, et al [15]. This method is called support vector regression (SVR) [16].

### 2.5.1 $\epsilon$-Support Vector Regression

The model generated by support vector classification (as described above) accounts only for a subset of the training data, because the cost function for building the model is not correlated to the training points that are not the support vectors. In a simple linear regression model, we minimize a regularized error function[15] given by

$$E_D(\boldsymbol{\omega}) = \frac{1}{2}\sum_{j}^{m}\{t_j - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{x}^{\mathbf{T}}\}^2 + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2.$$

In order to acquire sparse solutions, Vapnik replaced the quadratic error function term by an $\epsilon$-insensitve error function, which actually insert a threshold term $\epsilon$ to zero out the difference between the prediction $y(\mathbf{x})$ and the target vector *t*. Therefore, we produce and minimize a regularized error function given by

$$E_D(\boldsymbol{\omega}) = C\sum_{j}^{m}E_\epsilon(t_j - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{x}^{\mathbf{T}}) + \frac{1}{2}\|\boldsymbol{\omega}\|^2$$

where $E_\epsilon\big(t - y(\mathbf{x})\big) = \begin{cases} 0, & \text{if } |t - y(\mathbf{x})| < \epsilon \\ |t - y(\mathbf{x})| - \epsilon, & \text{othewise} \end{cases}$.

As the soft margin SVM idea we discussed in section 2.4.1, we reformulate the optimization problem by introducing two slack variables $\xi_m$ and $\hat{\xi}_m$. The purpose to do so is that $\epsilon$ and two slack variables establish a $\epsilon$-tube for a target point, such that $y(\mathbf{x}) - \epsilon - \hat{\xi}_m \le t_m \le y(\mathbf{x}) + \epsilon + \xi_m$, where $\xi_m$ and $\hat{\xi}_m$ are nonzero and positive unless the prediction $y(\mathbf{x})$ lies outside the tube region. We therefore are able to rewrite the regularized error function as

$$E_D(\boldsymbol{\omega}) = C\sum_{j}^{m}\big(\xi_m + \hat{\xi}_m\big) + \frac{1}{2}\|\boldsymbol{\omega}\|^2.$$

---

[15] Recall the non-regularized error function with respect to the Sum of Square error (SSS) showed in section 2.2.1, a penalty term controlled by a regularization term $\lambda$ is added to the equation of this error.

To minimize this error function with respect to two slack variables, as we showed in section 2.4.1, Lagrange multipliers are introduced and we optimize the Lagrangian function. In *C*-SVM, we obtain the box constraints $0 < a_m < C$ and $0 < \hat{a}_m < C$. To estimate new data points using the training model, we reformulate $y(\mathbf{x})$ and replace $\boldsymbol{\omega}$ by expressing $a_m$, $\hat{a}_m$, and the kernel function

$$y(\mathbf{x}) = \sum_{m=1}^{M} (a_m - \hat{a}_m) k(\mathbf{x}, \mathbf{x_m}) + \text{b}.$$

The corresponding KKT conditions, which states that the product of the dual variables and the constraints mush vanish. In other words, for all data points, either $a_m = 0$ or $\xi_m + \epsilon + y(\mathbf{x_m}) - t_m = 0$ holds. Regardless of how large the number of $a_m = 0$ points is, they are literally not supposed to be constructive predictions for new data points because $a_m \neq 0$ data points either lies on or above the upper boundary of the $\epsilon$-tube. On the other hand, regardless of how large the number of $\hat{a}_m \neq 0$ points is, they are literally not supposed to be constructive predictions for new data points because $\hat{a}_m \neq 0$ either lies on or below the lower boundary of the $\epsilon$-tube. We again have a sparse solution and we call this soft margin SVR $\epsilon$-SVR.

### 2.5.2 $\epsilon$-SVR-RFE for Feature Ranking

We presented a nonlinear regression model to validate the feature ranking performance in the section 2.2.2, which involves a feature subset selection technique called SVR-Recursive Feature Elimination (SVR-RFE) [17]. We implemented RBF kernel SVR wrapped with RFE to nonlinearly rank the input features we discussed in the section 2.1.2. SVM-RFE removes the feature with least weight magnitude in the SVM

solution. Intuitively, we ran 50 different random 5-fold splits of the data. For each split (trial) we selected the two SVR hyper-parameters ($\epsilon$ and $\gamma$ for $\epsilon$-SVR) using grid search to minimize the average held-out fold mean-squared prediction error. The sample mean and standard deviation of the RFE feature ranks over the 50 trials were used to compare with that of the feature ranking results we acquired in 2.2.2. All the $\epsilon$-SVR models were implemented in the LIBSVM toolbox compatible with MATLAB 8.0 [14].

## 2.6 Classification by Bayesian Decision Theory

Bayes decision theory is a fundamental statistical approach to the classification task. Unlike to the discriminant functions classification analysis we discuss, the probabilities play an important role in the parametric Bayes classifiers. The form of the input distributions is assumed to be known and parameters of the distributions are estimated from the design samples [18].

### 2.6.1 Bayes Decision Rule

Although the Bayes classifier is optimal its implementation is often difficult in practice due to its complexity, particularly when the data dimensionality is high. The traditional Bayes classifier characterizes classes by their probability density functions (pdfs) on the input features and uses Bayes decision rule to form decision regions form these densities. The Bayes decision rule generalizes an optimal classifier, i.e. classification error is minimal, when the model assumptions are correct.

In mathematical way, we express **x** to be the input vector and $C_i$ represents one of the possible classes that are of interest. Let $p(x|C_i)$ represent the class-conditional pdf for **x** and $p(C_i)$ represents the priori probability that class $C_i$ occurs. Then the Bayes rule is given by a conditional a posteriori probability of a class

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)}$$

where $p(x) = \sum_i p(x|C_i)p(C_i)$.

### 2.6.2 Naïve Bayes for Classification

Naïve Bayes classifier is a probabilistic classifier based on applying Bayes' theorem with assumptions of feature independence, by which we assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Due to the independence assumptions, we call this Bayes probabilistic model naïve Bayes. In terms of Naïve Bayes classifier, the idea is pretty simple. Using Bayes' theorem, we achieve a classifier as

$$\text{classify}(f_1, \ldots, f_n) = \underset{c}{\text{argmax}}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c).$$

The estimation of parametric naïve Bayes commonly uses the maximum likelihood method as showed above. To be more precise, this classifier stems from the maximum a posterior (MAP) decision rule. Regardless of the oversimplified assumptions, naive Bayes classifiers perform very well in many complex real-world situations.

In this thesis, we used Gaussian kernel model of this Naïve Bayes classifier to predict the presence or absence of HIV DNP, by computing the probability of each class and the product of the likelihood functions. The Naïve Bayes classifier results were used

to compare with that of Gaussian kernel SVM (section 2.4.2) and random forest classifier, which we are going to discuss in the next section.

## 2.7 Classification by Decision Tree Theory

As one of the directed graphical models, decision trees are useful for expressing relationships between random variables in the decision analysis. In the case of directed graphs, a tree is defined such that there is a single node, called the root, which has no parents, and all other nodes have one parent.

Random Forest [19] is an ensemble classifier that comprises of many decision trees. It outputs the class that is the mode of the class's output by individual trees. It deals with “small n large p”-problems, high-order interactions, correlated predictor variables. In terms of the algorithm of random forest, we assume that the user knows about the construction of single classification trees. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is grown as follows: If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant

during the forest growing. Each tree is grown to the largest extent possible. There is no pruning. The Random forest algorithm can be summarized to be the following steps:

Let Ntrees be the number of trees to build for each of *Ntrees* iterations

1. Select a new bootstrap sample from training set

2. Grow an un-pruned tree on this bootstrap.

3. At each internal node, randomly select *mtry* predictors and determine the best split using only these predictors.

4. Do not perform cost complexity pruning. Save tree as is, alongside those built thus far.

Output overall prediction as the majority vote (classification) from all individually trained trees.

In this thesis, we implemented a random forest algorithm for the classification of the presence or absence of HIV DNP analysis. The Random forest classifier results were used to compare with that of Gaussian kernel SVM (section 2.4.2) and Naïve Bayes classifier (section 2.6.2).

# Chapter 3

# EXPERIMENTAL RESULTS

In chapter 2, linear regression, epsilon-support vector regression (SVR), linear discriminant analysis (LDA) models, Gaussian kernel support vector machine (SVM) classifiers, Naïve Bayes classifiers, and random forest classifiers and learning algorithms were described. In this chapter, we perform the experimental comparisons of our two regression algorithms, and three classification techniques. First, the objectives of the experiments in this thesis are demonstrated. Then, we make several different comparisons of our regression and classification analysis methods. We will first pairwise compare our regression structures and learning methods. Then, we compare SVM classifier with the other two classifiers in three parallel experiments based on three different data sets. Third, we "vertically" compare the experimental significance of feature resolutions in the classification analysis. In addition, we make comparisons between the experiments with detrending and that without detrending.

The goals of the study are as follows: First, the study of regression analysis seeks the correlation between HIV DNP and cortical brain gray matter volume. Then, the classification study should firstly be able to achieve not only accuracy, but sensitivity and specificity rates as high as possible. Furthermore, more than one quantitative

validation should be implemented in the study, in order to analyze the advantages of different classifiers. In particular, we implemented nonlinear weighted support vector machine (SVM), repeated random subsampling (RRS) random forest and RRS Naïve Bayes classifiers.

## 3.1 Experimental Comparisons

In the case of regression, we simply compare the feature ranking results of two regression models. For each of the classifiers, two types of comparisons may be indicated. The first relates to the performance achieved by three distinct classifiers, which demonstrates how well the classification methods generalize to new data. The second and the more important comparison relates to the influence of feature dimensionality in classification analysis. In addition, we also measure the necessity of removing trends (age, gender and ethnicity) from the data samples by comparing the supplementary results using detrended data with that using non-detrended data.

### 3.1.1 Linear Regression vs. $\epsilon$-SVR

The objective of this study is to prove our assumption on the correlation between HIV DNP and the brain volume of cortical gray matter. Due to this primary objective, we used linear regression to explore how significant a role HIV DNP plays in the cortical gray matter brain volume changes. Furthermore, since our focus in this regression analysis is on the significance of features (feature ranking), we used all the data points to train the regression models, without any test data points. Here, we firstly implemented

a simple linear regression model and computed statistical *p*-values for each of the input variables in the regression model. However, the features remaining in the final selected model are considered as significantly influencing features in the cortical gray matter brain volume change. As the definition and properties of *p*-values we described in chapter 2, we should be able to investigate the significance of feature HIV DNP in the final selected model by sorting those *p*-values in ascending order. The higher order a feature ranks the more significant it is supposed to be.

The RBF $\epsilon$-SVR is specified by a set Gaussian basis functions $\exp\left(-\gamma\|\mathbf{x}-\boldsymbol{\mu}_j\|^2\right)$, and by a set of scalar weights $\{\lambda_j\}$, where $\mathbf{x}$ is the input vector and $j = 1,2,\ldots,M$. The weight vector indicates the significance of a feature in the model. The model selection algorithm we built was called recursive feature elimination (RFE). Statistically, we ran 50 different random 5-fold splits of the data (held-out cross validation) in order to obtain statistically meaningful results. Instead of setting a stopping condition for the algorithm, we recursively removed the minimum weight value and retrained the model until all features were eliminated from the model. Meanwhile, the indexes of the eliminated features were recorded, such that we should be able to investigate the significance of variable HIV DNP in the model by averaging the ranks of the index vector for 50 trials.

Table 3.1 gives the linear regression model analysis obtained by Minitab 16. For each of those 52 features, we calculated the *p*-values during the backward feature elimination process (the significance level was chosen to be $\alpha = 0.2$) and reported them in the figure. Table 3.2 shows top 10 ranked features selected by 50 trials SVM-RFE

algorithm. We made the ranking based on 50 trials average ranks a certain feature appeared to be. The mean and standard deviation of top 10 ranked features, as well their names were given. Note that the standard deviations presented to identify some negligible statistical significant features, which inferred to the high instability in the regression model due to high standard deviation.

Table 3.1 List of the explanatory features used for the multivariable models of the association between DNP and log brain volumes. The feature names along with the *p*-values (uncorrected for multiple comparisons) were listed. And the features less than 0.20 were the features in the final selected model.

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 1.7898 | 0.7227 | 2.48 | 0.014 |
| age | -0.0042232 | 0.0006868 | -6.15 | 0.000 |
| education | 0.001501 | 0.002100 | 0.71 | 0.476 |
| Gender@Birth | 0.04870 | 0.01505 | 3.24 | 0.001 |
| Log_cerebralVAULT | 0.82936 | 0.05202 | 15.94 | 0.000 |
| HCV Status | -0.01296 | 0.01232 | -1.05 | 0.294 |
| P1 HIV RNA UD/D | 0.00675 | 0.01340 | 0.50 | 0.615 |
| CSF HIV RNA UD/D | 0.00786 | 0.01392 | 0.56 | 0.573 |
| sqrt cd4 nadir | -0.0009626 | 0.0008946 | -1.08 | 0.283 |
| sqrt cd4 current | -0.0004850 | 0.0009369 | -0.52 | 0.605 |
| Current Ddrug exp | -0.0002378 | 0.0004350 | -0.55 | 0.585 |
| On Ddrugs | 0.05209 | 0.03158 | 1.65 | 0.101 |
| Total Ddrug exposure | 0.0000720 | 0.0001622 | 0.44 | 0.658 |
| num of PIS | 0.00110 | 0.01723 | 0.06 | 0.949 |
| bdi | 0.0002043 | 0.0004415 | 0.46 | 0.644 |
| GDS of demo | -0.02727 | 0.01077 | -2.53 | 0.012 |
| opiate | 0.01005 | 0.01274 | 0.79 | 0.432 |
| TCA | -0.00303 | 0.01542 | -0.20 | 0.844 |
| anticonvuls | -0.00692 | 0.01821 | -0.38 | 0.704 |
| DNP | -0.012869 | 0.004630 | -2.78 | 0.006 |
| ethnicity_1 | -0.04346 | 0.01354 | -3.21 | 0.002 |
| ethnicity_2 | -0.00919 | 0.01685 | -0.55 | 0.586 |
| ethnicity_3 | -0.01162 | 0.02949 | -0.39 | 0.694 |
| Scanner_1 | 0.01543 | 0.01852 | 0.83 | 0.406 |
| Scanner_3 | 0.01893 | 0.01649 | 1.15 | 0.253 |
| Scanner_4 | 0.05799 | 0.01855 | 3.13 | 0.002 |
| Scanner_5 | 0.08324 | 0.01882 | 4.42 | 0.000 |
| Scanner_6 | 0.06581 | 0.03179 | 2.07 | 0.040 |
| Scanner_7 | -0.00167 | 0.01955 | -0.09 | 0.932 |
| Ddrug Ever_1 | -0.01955 | 0.02875 | -0.68 | 0.497 |
| Ddrug Ever_2 | 0.01806 | 0.01331 | 1.36 | 0.176 |
| egimen Type 1=nnrti, 2=other, _0 | 0.02760 | 0.04223 | 0.65 | 0.514 |
| egimen Type 1=nnrti, 2=other, _1 | 0.03048 | 0.03945 | 0.77 | 0.441 |
| egimen Type 1=nnrti, 2=other, _2 | 0.05394 | 0.04307 | 1.25 | 0.212 |
| egimen Type 1=nnrti, 2=other, _3 | 0.02309 | 0.02385 | 0.97 | 0.334 |

S = 0.0625617  R-Sq = 82.4%  R-Sq(adj) = 78.9%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 34 | 3.178280 | 0.093479 | 23.88 | 0.000 |
| Residual Error | 174 | 0.681030 | 0.003914 | | |
| Total | 208 | 3.859311 | | | |

Table 3.2 The top 10 features by their mean ranked over 50 experimental trials which predict total cortical gray matter volume. Excluding the three bolded features (4th, 7th, and 8th) with negligible statistical significance owing to high sample standard deviation, DNP ranks seventh.

| **mean feature ranks** | **Features** | **Mean** | **Stdev** |
|---|---|---|---|
| 1 | Log_cerebralVAULT | 1 | 0 |
| 2 | GDS | 2.42 | 0.55 |
| 3 | Age | 2.61 | 2.54 |
| **4** | **Sqrt cd4 current** | **4.36** | **8.18** |
| 5 | Gender | 5.37 | 1.47 |
| 6 | Ethnicity | 6.79 | 0.16 |
| **7** | **Education** | **7.66** | **8.04** |
| **8** | **Sed abuse** | **7.68** | **6.59** |
| 9 | Scanner | 8.34 | 0.50 |
| 10 | DNP | 9.52 | 1.57 |

Our results in Table 3.1 and Table 3.2 together proved the assumption we made regarding the correlation between HIV DNP and cerebral cortical gray matter volume. Table 3.1 shows that DNP ranked sixth for the multivariable model, by which it means worse severity of DNP symptoms is significantly correlated with smaller cerebral

cortical gray matter volume. Correspondingly, Table 3.2 shows that the nonlinear regression model validated the same conclusion made by linear regression model. DNP ranked seventh in 50 trials feature ranking made by SVM-RFE. All the features ranked top 10 in term of statistical significance should be considered to be significantly correlated with cerebral cortical gray matter volume change.

### 3.1.2 Classifiers Comparisons

Due to the imbalance of the data samples (the ratio of active[16] data points and inactive[17] data points is 1:2.64), the sensitivity rates in the study would be considerably improved by equilibrating the ratio of the active and inactive data samples. Therefore, the algorithms used in the study should be able to reduce the imbalanced data impacts as much as possible.

The hyper-parameters optimization algorithm is necessarily crucial to our C-SVM classification analysis. It is because firstly, grid search algorithm essentially generalizes the optimal decision separation between two classes. Secondly, this automated optimization algorithm actually enhanced the data imbalance tolerance by adjusting the regularization hyper-parameter *C*. Therefore, our C-SVM algorithm with RBF kernel is called weighted-SVM algorithm. We then evaluated the classification performance using 50 different random 10-fold splits of the data, followed by a held-out cross validation algorithm. For each of the split (trial), we obtained the average

[16] Active data, in particular, stands for those samples with the presence of HIV DNP.

[17] Inactive data, in particular, stands for those samples with the absence of HIV DNP.

performance scores in the cross validation procedure. And eventually we took the average scores in 50 trials as the final outcomes from the weighted-SVM classifiers. All the SVM classification models were implemented in the LIBSVM [14] toolbox compatible with MATLAB 8.0.

Similarly, weighted random forest (WRF) algorithm [20] is also a good solution to data imbalance problems when we implemented random forest classifier. [21] compared WRF and balanced random forest (BRF) on six different and highly imbalanced datasets. In WRF, they tuned the weights for every data set, while in BRF, they changed the votes cutoff for the final prediction. They concluded that BRF is more computationally efficient than WRF for imbalanced data. They also found that WRF is more vulnerable to noise compared to BRF. However, we introduce an alternative RF algorithm without tuning the class weights or the cutoff parameter, which is called repeated random subsampling (RRS) random forest algorithm [23]. Random sub-sampling [22] is known as Monte Carlo cross validation, Random subsampling validation is based on randomly splitting the data into subsets. In our particular case, the 169 inactive data points has been randomly partitioned into three groups, such that the ratio of active data points and inactive data points from each subset was fixed to be 1:1.14. Eventually, every inactive sample in the training data was selected once, while every active sample was selected three times. After training and test the model on all the sub-samples, we employed a majority vote for the three sub-samples sets classification results, by which it means average votes were rounded to the nearest integers. We evaluated the classification performance using 50 different random 3-fold splits of the

data, followed by a held-out cross validation algorithm. We computed the average majority votes for accuracy, sensitivity and specificity scores.

Using the RRS idea we described, we applied RRS-Naïve Bayes classifier to modify the imbalanced dataset impacts on the Naïve Bayes classification analysis. We evaluated the classification performance using 50 different random 3-fold splits of the data, followed by a held-out cross validation algorithm. We computed the average majority predicted outcomes for accuracy, sensitivity and specificity scores given by those posterior probabilities we obtained.

Table 3.3 shows an experiment on the data set *volume.dat*, allowing each model to produce 50 classifier solutions. The accuracy rates, sensitivity rates, and specificity rates were averaged over 50 trials. Accordingly, two similar experiments were performed on the data sets *slicearea.dat* and *subslicearea.dat*. The experimental results are listed in Table 3.4 and Table 3.5.

Table 3.3 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *volume.dat* data set (experiment 1). Overall accuracy rates, sensitivity rates and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| **Weighted SVM** | **69.51%** | **±1.92%** | **53.37%** | **±3.70%** | **78.05%** | **±3.01%** |
| RRS-RF | 53.17% | ±4.97% | 55.36% | ±10.66% | 50.04% | ±7.78% |
| RRS-NB | 50.95% | ±6.48% | 51.36% | ±10.60% | 50.74% | ±11.08% |

Table 3.4 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *slicearea.dat* data set (experiment 2). Overall accuracy rates, sensitivity rates and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| **Weighted SVM** | **78.56%** | **±2.02%** | **63.54%** | **±3.12%** | **84.28%** | **±3.18%** |
| RRS-RF | 62.18% | ±5.41% | 55.74% | ±10.07% | 64.65% | ±8.34% |
| RRS-NB | 57.65% | ±4.94% | 57.43% | ±9.94% | 57.75% | ±7.55% |

Table 3.5 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *subslicearea.dat* data set (experiment 3). Overall accuracy rates, sensitivity rates and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| **Weighted SVM** | **80.12%** | **±2.89%** | **67.29%** | **±3.01%** | **85.05%** | **±4.68%** |
| RRS-RF | 58.87% | ±5.64% | 55.85% | ±7.49% | 59.36% | ±10.16% |
| RRS-NB | 58.12% | ±4.52% | 60.99% | ±10.22% | 57.68% | ±9.50% |

We have observed from the Table 3.3, Table 3.4 and Table 3.5 that the classification performance of three classifiers. Generally, weighted SVM classifiers over-performed the other two classifiers. And overall, the performance of RRS Naïve

Bayes classifiers was not as good as that of RRS random forest classifiers. Regardless of the advantages of Naïve Bayes classifiers and random forest classifiers, they have respectively instinctive disadvantages: 1) Random forest classifier may be too complex to for such a classification problem where the parameters are estimated from a limited number of training samples. 2) Even though 10-fold cross validation on the classification models with a finite number of training samples will expectedly bring about high variances results, random forest classifiers always produce even higher variances than weighted SVM classifiers in three experiments. 3) Naïve Bayes classifier assumption is always not true, especially for these three data sets. 4) RRS estimation algorithm is unable to perform as good as the weighted classifiers algorithm in terms of dealing with data imbalance problems, although it is much more computationally efficient and easy to implement.

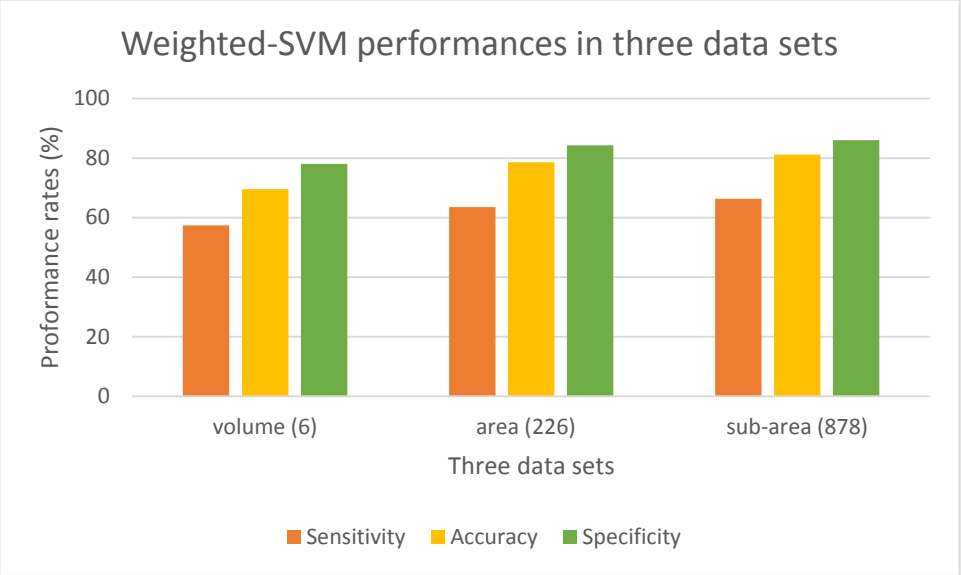
### 3.1.3 Feature Dimensionality Influence Comparisons

In this section we compare the performance of classification analysis on different data sets. As we discussed in section 2.1.3, we extracted features from the MRI brain images and created three distinct data sets, which end up with three experiments using weighted SVM classifiers, RRS random forest classifiers and RRS Naïve Bayes classifiers. Although the data sets are different, they are actually correlated in terms of feature dimensionality. It is due to the fact that we generalized *volume.dat, slicearea.dat,* and *subslicearea.dat* based on gradually increasing the feature resolutions. Moreover,

the ideas of three experiments are absolutely identical. People always say data beats algorithms, which is especially applicable to the cases of mining a finite number of data. In particular, the samples of our HIV DNP study are certainly limited and imbalanced. It is necessary to discover more useful information from the perspective of feature exaction. We therefore make comparison of the influence of feature dimensionality in our particular HIV DNP classification study, which explicitly illustrates the correlation between data and classification performances. For example, if we simply select the performance rates of weighted-SVM classifiers in these three experiments, we are certainly able to observe the trends of classification improvements we make as the feature dimensionality increases.

Figure 3.1 shows the performance of weighted SVM classifiers in the three experiments with different feature dimensionalities of the data sets. It is clear to observe that sensitivity rates, overall accuracy rates and specificity rates considerably increases in the area space data set (226 features) with respect to that of the volume space data set (6 features). Moreover, the overall performance slightly improves using the sub-regional area space data set (878 features) with respect to area space data set. We therefore conclude that it is potentially worthy to employ even higher dimensional features data set in the study to explore the trends of classification performance in feature resolutions series.

Figure 3.1 Comparison of weighted-SVM classifiers performances in three data sets with different feature resolutions. (·) represents the number of features. Overall speaking, sub-regional area space features set performs the best than the other two lower dimensional feature sets.




### 3.1.4 Detrending Effect Comparison

#### 3.1.4.1 Group Mean Detrending

With the idea of group mean detrending described in section 2.1.3.1, three data sets with detrending values were used in the supplementary experiments: *volume_norm.dat, slicearea_norm.dat, subslicearea_norm.dat*. In section 3.1.2, we compared the performance of weighted-SVM classifiers, RRS random forest classifiers, and RRS Naïve Bayes classifiers on the feature set version without detrending. In this section, we replicated three classifiers algorithms on three group mean detrending data

sets and compare the performance with the results shown in section 3.1.2. Table 3.6, Table 3.7 and Table 3.8 below show the classification results.

Table 3.6 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *volume_norm.dat* data set (experiment 1). Overall accuracy, sensitivity, and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| Weighted SVM | **72.06%** | **±1.92%** | **52.86%** | **±3.34%** | **79.22%** | **±3.06%** |
| RRS-RF | 53.08% | ±3.78% | 53.94% | ±8.32% | 52.64% | ±8.57% |
| RRS-NB | 56.15% | ±6.28% | 57.73% | ±9.88% | 55.35% | ±10.76% |

Table 3.7 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *slicearea_norm.dat* data set (experiment 2). Overall accuracy, sensitivity, and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| Weighted SVM | **76.98%** | **±2.25%** | **63.78%** | **±3.06%** | **79.68%** | **±3.10%** |
| RRS-RF | 55.12% | ±4.92% | 52.36% | ±8.99% | 56.56% | ±8.28% |
| RRS-NB | 51.72% | ±8.17% | 57.81% | ±9.75% | 52.37% | ±12.42% |

Table 3.8 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *subslicearea_norm.dat* data set (experiment 3). Overall accuracy, sensitivity, and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| **Weighted SVM** | **72.03%** | **±2%** | **64.09%** | **±2.87%** | **75.37%** | **±3.36%** |
| RRS-RF | 57.16% | 3.93% | 56.82% | ±10.41% | 58.81% | 6.75% |
| RRS-NB | 55.38% | 7.52% | 56.77% | ±11.49% | 54.74% | 10.49% |

We then pairwise compare the experimental results in Table 3.6, Table 3.7 and Table 3.8 with that in Table 3.3, Table 3.4 and Table 3.5. We observe that weighted-SVM classifiers still bring in the best performances in each supplementary experiments. But we are also able to see that the overall performances have no improvement if group-mean detrending algorithm applied to the data. In other words, we overestimated the effect of group mean detrending in our MRI brain image data. However, it does not imply trends elimination process is useless or unnecessary in our medical imaging analysis. It just indicates that grouping-mean idea is not reliable in our experiments due to the finite number of data samples we have. Very few data points in each group will cause the high variances in feature subsets.

### 3.1.4.2 Detrending by Additional Covariates

Group mean detrending algorithm brought about several disadvantages, however, we implemented another detrending approach using additional covariates. we have three data sets with demographic covariates were used in the supplementary experiments: *volume_plus.dat, slicearea_plus.dat, subslicearea_plus.dat*. In section 3.1.2, we compared the performance of weighted-SVM classifiers, RRS random forest classifiers, and RRS Naïve Bayes classifiers on the feature set version without detrending. In this section, we replicated three classifiers algorithms on three additional covariates detrending data sets and compare the performance with the results shown in section 3.1.2. Table 3.9, Table 3.10 and Table 3.11 below show the classification results.

Table 3.9 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *volume_plus.dat* data set (experiment 1). Overall accuracy, sensitivity, and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| **Weighted SVM** | **70.56%** | **±1.46%** | **53.56%** | **±3.87%** | **78.91%** | **±2.76%** |
| RRS-RF | 56.67% | ±5.03% | 58.81% | ±9.79% | 51.76% | ±8.87% |
| RRS-NB | 54.95% | ±7.42% | 55.63% | ±9.40% | 52.87% | ±10.36% |

Table 3.10 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *slicearea_plus.dat* data set (experiment 2). Overall accuracy, sensitivity, and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| Weighted SVM | **80.12%** | **±1.96%** | **63.90%** | **±2.93%** | **86.26%** | **±3.01%** |
| RRS-RF | 60.48% | ±5.11% | 60.80% | ±9.29% | 60.32% | ±7.47% |
| RRS-NB | 56.58% | ±4.89% | 56.67% | ±9.16% | 56.54% | ±8.43% |

Table 3.11 Comparison of Weighted SVM, RRS random forest and RRS Naïve Bayes classifiers on the *subslicearea_plus.dat* data set (experiment 3). Overall accuracy, sensitivity, and specificity rates were listed, which are the average of 50 experimental trials.

| **Classifiers** | **Accuracy (mean)** | **Accuracy (stdev)** | **Sensitivity (mean)** | **Sensitivity (stdev)** | **Specificity (mean)** | **Specificity (stdev)** |
|---|---|---|---|---|---|---|
| **Weighted SVM** | **81.42%** | **±1.24%** | **67.54%** | **±2.57%** | **86.16%** | **±3.16%** |
| RRS-RF | 63.08% | ±4.62% | 66.36% | ±8.86% | 61.39% | ±7.82% |
| RRS-NB | 58.46% | ±5.76% | 59.09% | ±11.63% | 58.14% | ±9.16% |

We then pairwise compare the experimental results in Table 3.9, Table 3.10 and Table 3.11 with that in Table 3.3, Table 3.4 and Table 3.5. We observe that weighted-SVM classifiers still bring in the best performances in each supplementary experiments.

But we are also able to see that the overall performances achieve a little improvement if additional covariates detrending algorithm applied to the data. In other words, the effect of additional covariates detrending algorithm over-performs the group mean detrending algorithm on our MRI brain image data. After all, trends elimination process is necessarily reliable in our medical imaging analysis.

# Chapter 4

# CONCLUSION

## 4.1 Summary

This thesis presents a novel study relative to existing HIV DNP studies in terms of the optimization of DNP classification performance based on the MRI brain images, using low dimensional feature subsets and computationally efficient models.

In chapter 1, the state of the arts of pattern recognition applications in bioinformatics were discussed. Key statistical regression and classification concepts for supervised learning problems were reviewed. In chapter 2, we introduced the data sets for regression and classification models. Under the assumption of the association between HIV DNP and the brain volume changes, we can test the hypothesis using multivariate regression analysis. Then for the purpose of diagnosing the presence of HIV DNP based on the brain images, a classification analysis was applied. We presented some supervised learning approaches such as linear regression model, SVR model, LDA, Gaussian kernel SVM, Bayes decision classifiers, and decision tree classifiers.

Chapter 3 presented the experimental results based on comparisons of the regression analysis and classification performances of several models. We first compared the nonlinear regression model with the linear multivariate regression model. We

concluded that our nonlinear regression analysis performances validated the results we obtained from the linear regression model. Second, we compared the performances of three classifiers we implemented on the non-detrended data sets. We concluded that the weighted-SVM classifier is the most reliable classifier in our three classification experiments. Then we "vertically" compared the performances of the non-detrended data sets with distinct feature resolutions. We concluded that the classification performances gradually improved as the feature resolutions increased. As a result, the data set with sub-regional area space features performed the best, and the overall accuracy rate reached 80%. Finally, in order to understand the helpfulness of trends elimination for feature vectors, we pairwise compared the performances of non-detrended data sets with those of detrended data sets. We concluded that the data sets with group mean detrending performed not as good as the non-detrended data sets in all three classification experiments. It is due to the fact that our group mean detrending algorithm controlled three covariates and thereby generalized 16 groups, however, high variance in the feature vectors was caused by high dimensional group averaging on our limited data samples. Alternatively, when we simply added those covariates into the data sets rather than taking group averaging, we observed and concluded that the classification results of the data sets with additional covariates over-performed those of the non-detrended data sets. In other words, trends elimination approaches improve the overall performance in our classification analysis.

## 4.2 Suggestions for Future Studies

Although we maximized the overall accuracy, as well as sensitivity and specificity rates in our classification analysis, absolute high classification performances haven't been achieved in the existing area space. In the future, we may want to further improve the HIV DNP classification performances by extending our study into voxel space. Based on the MRI brain images, new feature extraction algorithms are supposed to be developed to extract voxel intensity features from gray scale images or binary features from binary images. Furthermore, we are able to develop a HIV DNP regional detection algorithm based on classification analysis in voxel space. On the other hand, an intermediate solution might also be investigable to implement. That is, we could try even higher feature resolution data in area space, for instance, 16 quadrant area space data.

# REFERENCE


[1] J. Keltner, C. Fennema-Notestine, F. Vaida, D. Wang, et al, "HIV-Associated Distal Neuropathic Pain is Associated with Smaller Total Cerebral Cortical Gray Matter". *The Journal of Neuro Virology*, 2013.

[2] C. M. Bishop, Pattern Recognition and Machine Learning, *Springer-Verlag New York Inc.*, 2006.

[3] L. Wiebe, T.J. Phillips, J.M. Li, J.A. Allen, K. Shetty, "Pain in HIV: an evolving epidemic". *Journal of Pain*, 2011.

[4] R.J. Ellis, D. Rosario, D.B. Clifford, J.C. McArthur, D. Simpson, T. Alexander, et al, "Continued high prevalence and adverse clinical impact of human immunodeficiency virus-associated sensory neuropathy in the era of combination antiretroviral therapy: the CHARTER Study". *Archives of Neurology*, 2010.

[5] R.S. Swenson. (2006). *Review of Clinical and Functional Neuroscience* [Online]. Available: http://www.dartmouth.edu/~rswenson/NeuroSci/

[6] S.M. Smith, M. Jenkinson, M.W. Woolrich, C.F. Beckmann, T.E.J. Behrens, H. Johansen-Berg, P.R. Bannister, M. De Luca, I. Drobnjak, D.E. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J.M. Brady, and P.M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL". *NeuroImage*, 23(S1):208-19, 2004

[7] Analysis Group, FMRIB, Oxford, UK. (2012). FMRIB Software Library v5.0 [Online]. Available: http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL

[8] A. May, "Structural Brain Imaging: A Window into Chronic Pain". *The Neuroscientist*, 17:209, April 2011.

[9] P.Y. Geha, M.N, Baliki, R.N. Harden, W.R. Bauer, T.B. Parrish, A.V. Apkarian. "The brain in chronic CRPS pain: abnormal gray-white matter interactions in emotional and autonomic regions". *Neuron*, 60: 570-581, 2008.

[10] M.H. Kutner, C.J. Nachtsheim, J. Neter, *Applied Linear Regression Models (4$^{th}$ edition)*, McGraw- Hill/Irwin Press, 2004, pp. 232.

[11] C.M. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer, 2006, pp. 293-294.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection". *J. Mach. Learn. Res*., vol. 3, pp. 1157-1182, 2003.

[13] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* **2** (12): 1137–1143, 1995.

[14] C.C. Chang and C.J. Lin: LIBSVM: "a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology* 2(3): 27:1-27:27, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[15] H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, and V.N. Vapnik, (1997) "Support Vector Regression Machines", in *Advances in Neural Information Processing Systems 9, NIPS 1996*, 155–161, MIT Press.

[16] A.J. Smola and B. Schölkopf, "A tutorial on support vector regression*." *Statistics and computing* 14(3): 199-222, 2004

[17] Y. Aksu, D.J. Miller, G. Kesidis and Q.X. Yang, "Margin-Maximizing Feature Elimination Methods for Linear and Nonlinear Kernel SVMs". *IEEE Trans. Neural Networks* 21(5):701-717, May 2010.

[18] K. Fukunaga, *Statistical Pattern Recognition*, 2nd ed., San Diego, Academic Press Inc., 1990.

[19] L. Breiman, "Random Forests". *Machine Learning,* **45** (1): 5–32, 2001

[20] S.J. Winham, R.R. Freimuth, J.M. Biernacka, "A weighted random forest approach to improve predictive performance". *Statistical Analysis and Data Mining*, **6** (6): 496-505, 2013.

[21] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data", University of California, Berkeley; 2004.

[22] R.R. Picard and R.D. Cook, "Cross-validation of regression models". J. Am. Stat. Assoc., 79: 575-583, 1984.

[23] M. Khalilia, S. Chakraborty, and M. Popesu, "Predicting disease risks from highly imbalanced data using random forest". *BMC Medical Informatics and Decision Making*, 2011.