Addressing the Challenges of Applying Machine Learning for Predicting Mental Disorders and Their Prognosis Using Two Case Studies

by

Seyedehreyhaneh Ghoreishiamiri

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science

Department of Computing Science
University of Alberta

Abstract

One of the principal applications of machine learning in psychiatry is to build automated tools that can help clinicians predict the diagnosis and prognosis of mental disorders using available data from patients' profiles. Here, in two different studies, we investigate ways to use machine learning to produce models that can predict mental disorders and their prognosis, using different neuroimaging modalities, genotype data, and clinical information.

The first study addresses the challenge of producing a classifier that a human clinician can interpret and potentially use in clinical practice. In this study, we were seeking a simple and accurate classifier that can correctly distinguish Alzheimer's disease (AD) patients from healthy controls (HC). We wanted to learn this classifier from the data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, using just a fairly small set of input features, including grey matter volumes of 33 regions of interest derived from brain structural MRI, as well as the APOE genotype. Running our overall learner, involving standard feature selection processes and three simple base-learners on these features, produced a 7-feature elastic net model that achieved accuracy of 89.28% on the test set. Next, we ran the same overall learner using two more-complex base-learners over the same initial dataset. The accuracy of the best model here (SVM-RBF

over 23 features) was 90.47%, which was not statistically different from the performance of our much simpler linear model, over just 7 features. We, therefore, introduce this simple 7-feature model as our accurate and simple classification model.

Our second study explored the utility of machine learning methods in predicting the response of a group of schizophrenia patients (n= 51 to 90, depending on the response criterion) to a specific treatment, given their functional magnetic resonance imaging data, structural magnetic resonance imaging data, diffusion tensor imaging data, and clinical information. In this study, we explored various clinical measures for defining treatment response, various feature types for imaging and non-imaging data, various machine learning tasks and learning algorithms, but (probably due to the small sample size), we were not able to obtain any significant results.

Preface

The third chapter of this dissertation, Alzheimer's diagnosis prediction, is taken from our collaborative research manuscript, "A Simple Classification Framework for Predicting Alzheimer's Disease from Region-based Grey Matter Volume and APOE Genotype Status" by Reyhaneh Ghoreishiamiri, Graham Little, Matthew R.G. Brown, and Russell Greiner.

Data preprocessing and feature extraction in Chapter 2 was performed by Graham Little. Data preprocessing and feature extraction in Chapter 3 is my original work, under the supervision of Dr. Sunil Kalmady. The script that was used for brain parcellation in Section 3.3.2 and Section 3.3.3 was developed by Dr. Sunil Kalmadi.

The survival regression tools, mentioned in Section 3.6, were developed by Humza Heidar and Dr. Russell Greiner.

Acknowledgements

First, I would like to offer my special thanks to my supervisor, Professor Russell Greiner, without whose help and support this work would have been impossible. I also would like to express my great appreciation to my co-supervisor, Professor Matthew Brown, for all of his advice and feedbacks during my master's.

I am particularly grateful for all the assistance given by Dr. Sunil Kalmadi and Graham Little in my projects.

I also would like to thank Roberto Vega for his great advice and recommendations in my projects and studies.

Thanks, all of my friends in computational psychiatry group and Edmonton for supporting me during this time.

I would like to thank my mother, father and sister for their kindness and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

With the gradual collection of large psychiatric datasets, psychiatric research has also entered the era of Big Data. These datasets often include thousands of heterogeneous variables, including clinical, neuroimaging, genomic measures and possibly other modalities [6]. Analyzing such datasets is challenging, especially when dealing with a high number of measurements and a high number of individuals, and may be further complicated by highly correlated variables [6]. Supervised machine learning or prediction modelling is a branch of statistical research that focuses on first, effectively learning the relationship between a large group of input variables and a target variable, and then, predicting the target variable for previously unseen instances. Therefore, prediction modelling allows for individualized prediction of early diagnosis and treatment outcome (as target variables) in psychiatry research [7].

This dissertation is composed of two psychiatric prediction modelling studies: Alzheimer's diagnosis prediction and schizophrenia prognosis prediction. The two following chapters, Chapter 2 and Chapter 3, are dedicated to each of these studies. The first study, Alzheimer's diagnosis prediction, is taken from our manuscript, "A Simple Classification Framework for Predicting Alzheimer's Disease from Region-based Grey

Matter Volume and APOE Genotype Status". This study aims to build a simple classifier that can accurately distinguish Alzheimer's patients from healthy controls using their structural magnetic resonance imaging and a single genotype, APOE. In this study, we were successful building a 7-feature elastic net model – 6 regional grey matter volumes and one genotype, APOE – with 89.28% accuracy, 84.00% specificity, and 93.54% sensitivity and show that the accuracy of this model is not statistically significantly different from those of more-complex classifiers. The second study aims to classify responder schizophrenia patients from non-responders using their functional magnetic resonance imaging data, structural magnetic resonance imaging data, diffusion tensor imaging data, and clinical information. In this study, despite exploring many different machine learning tasks and methodologies, we were not able to achieve any significant results. Each of the Chapters 2 and 3 will include their own introduction sections (Sections 2.1 and 3.1), which discusses the background and motivations of each of these studies in more detail.

The rest of this chapter provides relevant background material in order to clarify the remaining chapters of this thesis.

1.1 Types of Magnetic Resonance Imaging

A magnetic resonance (MR) scanner uses superconducting electromagnets to create a static high-strength magnetic field, usually ranging from 1.5 to 3 Tesla for hospital scanners [5] (see Figure 1.1). The radio-frequency coils inside the scanner generate magnetic pulses; when a pulse is turned on, it modifies the alignment of hydrogen protons (mostly in water molecules) within the magnetic field and when it is turned off, the protons relax to their original state, which releases energy that forms the raw MR signal [5]. Another magnetic source, known as gradients, is responsible for creating the spatial resolution; as the strength of each gradient varies linearly along

each dimension, the three orthogonal gradients create a signal in three spatial dimensions [5].

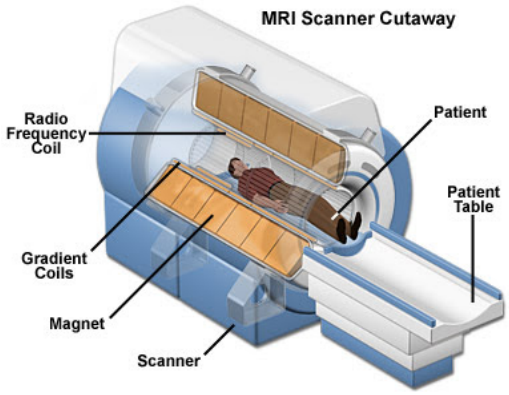

Figure 1.1: Illustration of a MRI Scanner (taken from [1])

Despite using similar equipment settings, different categories of MR imaging techniques – structural MR, functional MR and diffusion tensor imaging – serve different purposes. Structural MR aims for capturing anatomical abnormalities by measuring the density of water molecules [5]; functional MR aims at capturing the functional activity of different areas of brain by measuring the changes in the blood oxygen-level dependent (BOLD) signal [5]; and diffusion tensor imaging aims at capturing the anatomical abnormalities of the brain (specifically, white matter integrity) by measuring the shape of diffusion of water molecules [3]. In this section, we will briefly introduce each of these MR imaging techniques.

1.1.1 Structural Magnetic Resonance Imaging

The goal of structural MRI (sMRI) is usually to measure the density of water molecules (hydrogen protons, specifically) at a given location, which reveals much information about the normal or abnormal underlying tissue – *e.g.*, bone, gray matter, white matter or tumor [5]. sMR scans are high-resolution (spatial resolution) images that not only provide elabo-

rate details about the underlying anatomy of the brain, but can be used as reference images for other imaging modalities (like fMRI and DTI) for co-registeration and normalization purposes [5]. We will discuss the measures that we have derived form sMRI data in Section 3.3.2.

1.1.2 Diffusion Tensor Imaging

Diffusion tensor imaging is a means of exploring the tissue structure of the brain by measuring the shape of diffusion of water molecules, which reveals the complex structure of the fiber tracts in the brain and white matter integrity [3]. As diffusion of water molecules in the tissue can vary with direction (anisotropic diffusion, as opposed to isotropic diffusion, which is the same in all orientations; see Figure 1.2), modeling this diffusion reveals much information about the orientation of the underlying tissue [3].

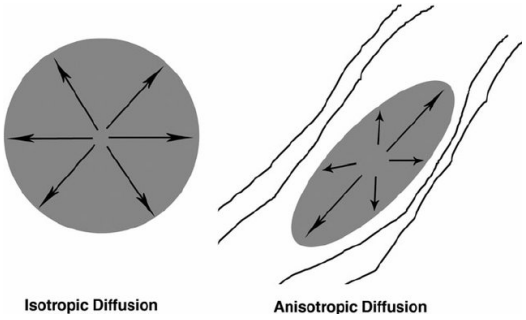

Figure 1.2: Isotropic vs anisotropic diffusion (taken from [2])

Diffusion tensor (DT, a.k.a. diffusion model) models the diffusion of water molecules using a Gaussian distribution. At each voxel, DT is modeled as a 3×3 positive definite and symmetric matrix with three eigenvectors, where the major one shows the fastest diffusion direction which often corresponds to the fiber tract axis of the underlying tissue [3]; see

Figure 1.3. The diffusion tensor matrix for anisotropic diffusion has form:

$$\begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{1.1}$$

where the three diagonal elements (D_{xx}, D_{yy}, D_{zz}) represent diffusion coefficients along each of the principal (x, y and z) axes and the six off-diagonal terms (D_{xy}, D_{yz}, ...) represent the correlation of motions between each pair of directions [8]. The diffusion tensor matrix for isotropic diffusion, on the other hand, has form:

$$\begin{bmatrix} D & 0 & 0 \\ 0 & D & 0 \\ 0 & 0 & D \end{bmatrix} \tag{1.2}$$

where the diagonal elements are all equal and the off-diagonal elements are all zero. The three orthogonal eigenvectors at each spatial position in the brain form a local coordinate system. Thus, the three orthogonal eigenvectors, alongside the three positive eigenvalues, showing the direction of diffusity in each of the eigenvectors, are described as an ellipsoid that shows the isosurface of diffusion probability [3]; see Figure 1.4.

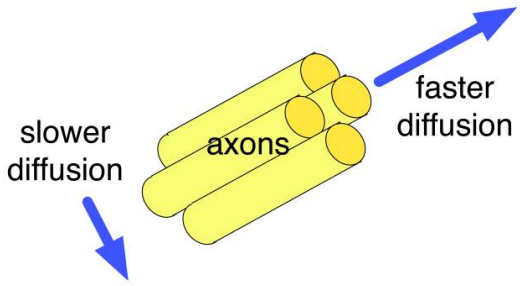

Figure 1.3: Illustration of anisotropic diffusion of water molecules in the tissue, measured via DTI (taken from [3])

To measure diffusion using magnetic resonance imaging, external magnetic field gradients are applied to create an image that reflects diffusion

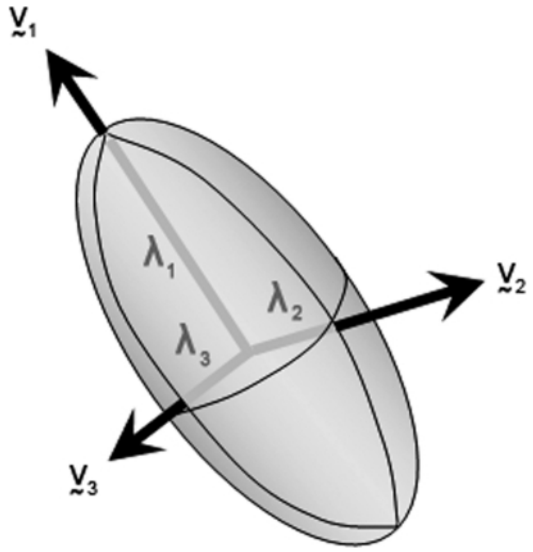

Figure 1.4: Diffusion ellipsoid (diffusion tensor model) with 3 eigenvectors: v_1, v_2 and v_3, and 3 eigen-values: λ_1, λ_2 and λ_3 as an estimation for the diffusion of water molecules (taken from [4])

in each orientation. Then, this process is repeated in multiple directions to estimate a three-dimensional diffusion model or tensor [3]. As a result, a diffusion tensor raw image is a 4D image with the first 3 dimensions showing the spatial position in the brain and the fourth dimension showing the diffusion-sensitizing directions that have been applied during acquisition. We will discuss the scalar measures that are derived form DTI data in Section 3.3.4.

1.1.3 Funtional Magnetic Resonance Imaging

Functional magnetic resonance imaging is a means of capturing the neural activity by measuring the real-time changes in the blood oxygen-level dependent (BOLD) signal [5], which reveals much information about the neural and cognitive architecture of the brain. When the neural activity in a region of brain increases, oxygenated hemoglobin races to the region due

to the increased metabolic need, which causes a rise in the BOLD signal [5]. The BOLD signal is a measure of the ratio of oxygenated to deoxygenated hemoglobin, a molecule in the blood that is responsible for carrying oxygen.

In an fMRI scanning session, the spatial resolution in the functional scan is sacrificed for temporal resolution, so the functional image is typically accompanied by a high-resolution structural scan [5]. The scanning session parameters are among the factors that drastically affect the fMRI signal and should not be disregarded, particularly when using collected data from multiple sites. These parameters include scanner magnetic field (which is a key parameter for all MR modalities), repetition time (TR), which is the time between each of the whole brain scans and the voxel size, which determines the spatial resolution of the functional brain scan [5].

Another issue is that the BOLD signal that we extract from the fMRI data is not the neural activity itself, but a function of it [5]:

$$B_i(t) \quad = \quad f[\, N_i(t)\,] \tag{1.3}$$

where $B_i(t)$ is the BOLD response at time t and location i, and $N_i(t)$ is the neural activation at time t and location i. Neural activation is mostly considered as a latent variable that is inferred from observed BOLD response. Thus, to derive the relationship between the observable variable and the latent variable, assuming the superposition principle holds for BOLD response, we can model the BOLD response to any neural activation function as [5]:

$$B(t) \quad = \quad \int_0^t N(\tau)\, h(t-\tau)\, d\tau \tag{1.4}$$

where $h(t)$ is the hemodynamic (impulse) response function or hrf, which is typically a gamma function or difference of two gammas in most of the studies [5]. Time origin (time 0) is the time when the an event happens or the neural activity is induced.

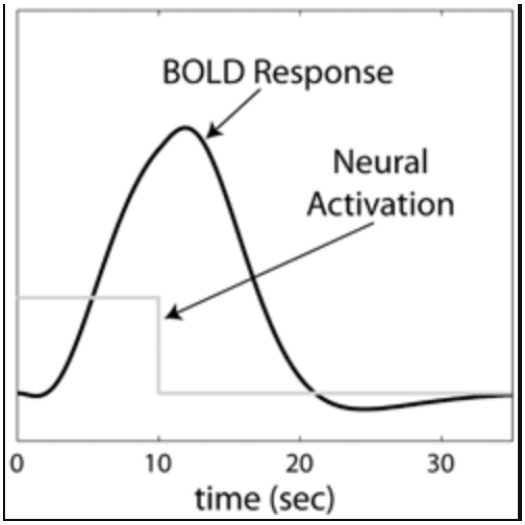

Figure 1.5: The relationship between BOLD signal, $B_i(t)$, and neural activity, $N_i(t)$, at location i (taken from [5]).

In a task-based fMRI study, the patients perform a particular task according to the experiment design to analyze the changes in their neural activity, while in resting-state, a.k.a. task-free fMRI, the patient is only asked to rest and not perform any tasks, as the goal here is to analyze the interaction between their brain regions. We used resting-state fMRI data in this study. We will discuss the measures that we have derived form fMRI data in Section 3.3.3.

1.2 Treatment Response Measures

There are currently several different categories of symptomatic and functional severity measures for schizophrenia. To have a well-defined criterion for treatment response (how well a patient responds to a certain treatment), it is crucial to first, select the severity measure and second, determine the treatment response, based on the selected measure. In this section, we will present a brief description of each of these severity mea-

sures.

1.2.1 Clinical Global Impression (CGI)

CGI provides an overall measure that encapsulates all the available information, including patient's history, symptoms, behavior, cognitive state and functioning, measured by an expert clinician (completely based on the clinician's subjective assessment) [9].

The CGI includes two companion one-item measures: First, CGI severity (**CGI-S**) describes, on a scale of 1 (normal) to 7 (among the most extremely ill patients), how mentally ill the patient is at the time of visist, based on the clinician's total clinical experience with the particular population which the patient belongs to [9]. Second, CGI improvement (**CGI-I**) compares the severity of illness between one week before the initiation of medication and the visit after taking the prescribed medication, on a scale of 1 (very much improved) to 7 (very much worse) [9]. Section 3.2.1 will discuss our treatment response definition based on CGI-S and CGI-I criteria.

1.2.2 SAPS and SANS

SANS (Scale for the Assessment of Negative Symptoms) and SAPS (Scale for the Assessment of Positive Symptoms) are each applied frequently in clinical settings as a reliable and consistent measure of severity of schizophrenia [10]. SAPS asks the clinician to report values for positive symptoms using 34 items, each on a 6-point (0-5) scale. Items fall under 4 categories: hallucinations, delusions, bizarre behavior, and positive formal thought disorder (FTD). SANS, on the other hand, asks the clinician to report values for negative symptoms using 25 items, each on a 6-point (0-5) scale. Items fall under five categories: affective blunting, alogia, avolition, anhedonia, and attention. Note that this scaling is different from the positive

and negative symptom scale (PANSS) [10]. Section 3.2.2 will discuss our treatment response definition based on SAPS and SANS criteria.

Chapter 2

Alzheimer's Diagnosis Prediction

2.1 Introduction

Alzheimer's disease (AD) is a highly prevalent neurodegenerative disease affecting an estimated 5.5 million Americans [11]. In Alzheimer's disease, neurons in all areas of the brain are eventually damaged or destroyed, including those that enable a person to perform basic daily functions such as walking and swallowing [11]. People in the final stages of AD are bedbound and demand uninterrupted care and the disease is ultimately fatal [11]. Patients with AD experience progressive cognitive impairment associated with patterns of structural brain atrophy more severe than the volumetric loss typical of healthy aging populations, but some of these structural changes may not be may not be visible to a clinician's eye by looking at the patient's MRI scan until the late stages of the disease. The high prevalence of AD combined with downstream progressive impairment has motivated investigations into advanced diagnosis strategies capable of early detection of the disease. Moreover, recent investigations that apply machine learning techniques to structural brain imaging have shown promise in accurately discriminating AD patients from controls. Multiple studies have used voxel-based morphometry (VBM) to distin-

guish AD patients from controls. In four consecutive studies, VBM features combined with different feature selection methods showed high prediction accuracies of 89% to 96% [12, 13, 14, 15]. VBM combined with texture analysis features has also been successful in classifying AD patients achieving 92.86% accuracy [16]. Another study, using multimodal features – including voxel-wise structural MR and fluorodeoxyglucose (FDG)-positron emission tomography (PET) imaging features, cerebrospinal fluid (CSF) biomarkers, cognitive scores and APOE genotype data – to predict conversion of Mild Cognitive Impairment (MCI) patients to Alzheimer's disease, achieved an accuracy of 92.4% [17]. Additionally, applying a 3D convolutional neural network (CNN) on 3D T1-weighted structural MR images has shown 99.2% accuracy, 99.5% specificity, and 98.5% sensitivity [18] in classifying AD patients vs controls (to our knowledge, this is the current best accuracy result with this dataset on classifying AD patients vs controls). Note that in all of these studies, the diagnostic labels are determined by expert clinicians. Despite achieving impressive accuracies of 92% to 99% in classifying AD patients, all of these methods use a large number of features (ranging from 100 to 2000) and complicated diagnostic models that are difficult for a human to interpret [19] and thus, are not applicable in clinical environments where troubleshooting diagnostic tools is of critical importance; see Section 2.5.1.

Regional brain features, based on cortical/subcortical segmentation, involve many fewer variables than voxel-based features; this is more appropriate for simple classification models. One study – which combined segmentation-based features of cortical thickness, cortical area, cortical curvature, grey matter density, subcortical volumes and hippocampal shape – achieved 0.98 AUC[1] [20]. A multimodal study achieved 93% accuracy by combining region-based features from structural MRI, FDG-PET and CSF proteins (189 features total) [21]. Such models, which use the features for

[1] AUC = Area Under the Curve of the Receiver Operating Curve (ROC)

all brain regions (one feature for each region, at least), involve a total of hundreds of brain features; this means they are necessarily very complex.

A priori selection of the brain regions typically impacted by AD has been a successful strategy for reducing the complexity of classification models. Using 6 features – namely the left and right hippocampus volume, amygdala volume and entorhinal cortical thicknesses – a support vector machine (SVM) classifier with the radial basis function (RBF) kernel, scored 0.89 AUC [22], suggesting that only few brain features are needed to discriminate AD from healthy controls. Additionally, grey matter volumes and diffusion-based MRI parameters over predetermined brain regions have shown utility in classifying MCI patients from healthy controls, achieving 89.7% accuracy [23]. Taken together, these studies suggest that simple classification models, on a limited number of brain features, are sufficient to discriminate AD patients from controls. However, more work is needed to assess whether simple predictive models involving a limited number of features can achieve classification results comparable to those of more complex diagnostic models.

This chapter explores the challenge of learning a simple classifier that can accurately distinguish patients with Alzheimer's disease (AD) from healthy controls. Section 2.2 describes the dataset we used, that describes each of the 752 patients using a small set of 33 brain volumes along with the APOE genotype status. Section 2.3 describes how we use that database to produce an accurate classifier. This process involves the pre-processing and feature extraction step, base-learners, feature selection methods, evaluation method, and overall learners. Section 2.4 presents our training, test, and feature selection results. Using these 34 features, we first compare the value of applying our overall learner, involving standard feature selection processes, with 3 base-learners selected for their simplicity – decision tree, elastic net and linear SVM – versus 2 relatively complicated base-learners – SVM with RBF kernel and extreme gradient boosting learner. Section 2.5

then compares the result of our simple classifier, produced by our overall learner with the simple base-learners, to other studies that also classify Alzheimer's patients versus healthy controls.

2.2 Participants / Imaging Data

This analysis used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, with the primary goal of testing whether a combination of serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be used to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [24, 25]. For up-to-date information, see www.adni-info.org.

Out of initially 793 subjects from various ADNI projects who had baseline scans, 35 subjects were excluded due to segmentation problems and 6 due to missing APOE genotype data. Our analysis considers the remaining (n=752) individuals, described using their baseline MR imaging and genetic sequencing (including the APOE genotype status, indicating the number of the APOE4 alleles at the APOE gene locus) – including 337 with a diagnosis of AD (age 75.26 ± 7.81, 44.8% ♀) and 415 controls (age 74.79 ± 5.72, 49.6% ♀). Mini-mental state examination (MMSE) [26] scores as well as clinical dementia rating sum of boxes (CDRSB) [27] scores were collected for all subjects. MMSE scores ranged from 18-28 for AD patients and from 24-30 for controls, and CDRSB scores ranged from 1-10 for AD patients and from 0-1 for controls.

The ADNI data was acquired from 60 different sites across US and Canada and the diagnostic status of the subjects was labeled by expert clinicians. To demonstrate our methodology, we merged data from 48 of their sites to become the training set, ADNI_TRAIN (with n=584 subjects)

	Training Set (ADNI_TRAIN)		Test Set (ADNI_HO)	
	AD	HC	AD	HC
Number	149 ♂ / 113 ♀	160 ♂ /162 ♀	37 ♂ / 38 ♀	49 ♂ /44 ♀
Age	75.04 ± 7.87	75.03 ± 5.66	74.54 ± 7.40	73.92 ± 5.90
MMSE	23.16 ± 2.08	29.07 ± 1.13	23.42 ± 2.07	29.06 ± 1.09
CDRSB	4.45 ± 1.64	0.03 ± 0.13	4.25 ± 1.71	0.03 ± 0.11
Number of APOE4	0.86 ± 0.70	0.31 ± 0.53	0.84 ± 0.75	0.25 ± 0.43

Table 2.1: Demographic information for the participants included in the training and test datasets (divided based on the data acquisition sites). Numbers for age, MMSE, CDRSB, and the number of APOE4 alleles are each shown as mean $\pm$ STD.

and merged the remaining 12 sites into the held-out test set, ADNI_HO (with n=168 subjects). Note these two sets are disjoint. Table 2.1 provides demographic information, showing the sex, age, diagnostic distributions and other information, for the training and test sets.

2.3 Materials and Methods

2.3.1 Image Acquisition and Segmentation

As part of the ADNI data collection, standardized structural MR imaging data was acquired for all participants using a sagittal magnetization-prepared rapid gradient echo sequence, with $1 \times 1\text{mm}^2$ in-plane resolution, 1.2mm slice thickness, and a field of view of $192 \times 192\text{mm}^2$ [28]. The scanner field strength varied from 1.5 to 3.0T, depending on the site [28]. Here, we focused on the participants' baseline imaging data. We used the Freesurfer (version 5.3) segmentation pipeline to extract regional cortical and subcortical volumetric measurements from each subject's MRI scan [29]. This process generated 68 regional cortical volumes (34 in each hemisphere), as well as 43 subcortical volumes for each subject. To compensate for the possible existing variance in the brain sizes of individuals,

we used the *normalized percent volumes* of each region, which is 100 times the ratio of the volume of that region, divided by the total intracranial volume, for each individual.

From the segmentation output, we selected the 33 normalized regional brain volumes that have shown robust group-level differences in previous imaging studies of AD: 10 subcortical regions (left and right thalamus [30], putamen [30], amygdala [31], hippocampus [32] and lateral ventricles [32]), 10 medial temporal regions [31] (right and left parahippocampal gyrus, entorhinal cortex, inferior temporal gyrus, middle temporal gyrus and superior temporal gyrus regions), 8 parietal regions [33] (left and right posterior cingulate gyrus, isthmus of cingulate gyrus, inferior parietal lobule and precuneus), 3 callosal regions [34] (posterior, central and anterior corpus callosum), and bilateral cerebellar cortex [35]. In addition to these brain imaging results, Corder *et al.* [36] showed that the number of the APOE4 alleles at the APOE gene locus is widely associated with late onset Alzheimer's disease. Our dataset therefore described each patient using this one genotype, as well as the normalized grey matter volumes of these 33 regions.

We z-scored each feature to have a mean value of 0 and a standard deviation of 1. Our training dataset describes each subject

$$x \quad = \quad [x_1, ..., x_{34}]$$

based on normalized brain volumes from 33 brain regions (x_i for $i \in \{1, 2, ..., 33\}$), and the APOE genotype status x_{34}.

2.3.2 Base Learning Algorithms

We input this data to various base learning algorithms, all implemented in the Python software packages scikit-learn[1] and XGBoost[2]. *Binary decision trees* (DT) are one of the most visually simple classifiers; they are also similar to clinical algorithms used for sequential diagnosis in medicine. We limited the depth of our decision trees to 10 and the number of leaf nodes to 20 for further simplicity and also to reduce the chance of overfitting. We also consider two linear models, each learning a weight vector $W = [w_0, w_1, ..., w_{34}] \in \Re^{34}$, using the function

$$y_W(x) \quad = \quad \sum_{i=1}^{34} w_i \times x_i \quad + \quad w_0 \tag{2.1}$$

Here, the model predicts the subject x has AD $\iff$ $y_W(x)$ is larger than 0. One linear model, *logistic regression with elastic net penalty* (EN), was our second simple classifier, combining L1 (Lasso) and L2 (Ridge) regularizers with a ratio (*L1_ratio*) that weighs the two penalties and an α parameter that weights the penalty term [37]. *Support Vector Machines* (SVM) is one of the most commonly used classifiers in Alzheimer's prediction studies [12, 13, 14, 15, 21]. In this study, we consider SVM classifiers with two different kernels: the linear and radial basis function (RBF) kernel [38]. The *linear SVM* (ℓSVM) is an example of a simple classifier while *SVM with RBF kernel* (rSVM) is used as an example of a more complex non-linear method. *Extreme Gradient Boosting* (XGB) [39] is a gradient boosted decision method, which is also used as as an example of a more complex non-linear base-learner in this study.

"Overall learner" OL is the system that invokes the pre-processing steps, etc., before running the base-learners. It also does the grid-searches to find

[1] See the scikit website, http://scikit-learn.org/

[2] See the XGBoost website, https://xgboost.readthedocs.io

the best learning algorithm and feature selection parameter settings, and also selects the best of these learners, based on the performance over the training data (see Figure 2.1).

2.3.3 Learning Process

We ran our OL twice, over different sets of base-learners: once for the simple base-learners – ℓSVM, EN, DT – and once for the more-complex base-learners – rSVM and XGB; for further clarity, we name the first one S.OL and the second one C.OL. Each of these runs produced a classifier, one simple and one complex.

Recall we used a training set ADNI_TRAIN, of 584 subjects, and a disjoint test set ADNI_HO, of 168 subjects. In both cases, OL partitioned ADNI_TRAIN into 5 folds and and used the same folds for all 5 base-learners. Each base-learner had to determine the best values for a set of hyper-parameters (described below), and use these when learning its classifier. The learning algorithm identified the best settings for these hyper-parameters based on average 5-fold cross-validation (CV) accuracies using grid search.

OL then used the 5-fold cross-validation accuracies of these base-learners (accompanied by various feature selection methods, described in Section 2.3.4) with these selected hyper-parameters, to identify the best-learners. It then ran that best base-learner on the entire training set to produce classifiers; we then tested the learned classifiers (simple and complex) on the test set, ADNI_HO. We evaluated our final models using specificity and sensitivity, as well as accuracy. For all three measures, for each model (simple and complex), we reported the performance of that model on the test set, as well as the mean and standard deviation, over the 5 folds of the training set.

For SVM methods, the C hyper-parameter was chosen from {1E-5, 1E-4, ..., 1E3, 1E4}, and γ was chosen from {1E-6, 1E-5, ..., 1E1, 1E2}. For

elastic net, the α hyper-parameter was chosen from {1E-4, 1E-3, 1E-2, 1E-1, 2E-1, ..., 9E-1} values and *L1_ratio* from {0, 5E-2, 1E-1, ..., 9E-1, 9.5E-1, 1} values [1]. Note that setting the *L1_ratio* to 1 means the learner only applies L1 regularization (aka Lasso classification) and setting it to 0 only applies L2 regularization (aka Ridge classification). For the decision tree base-learner, we set the maximum depth of the tree to 10 for further simplicity of our tree model and then used internal cross-validation to find the best values of three hyper-parameters: *minimum samples split* is the minimum percentage of training set instances required to split an internal node, chosen from a range of values between 0.005 to 0.480 (of total number of samples); *minimum samples leaf* is the minimum percentage of instances required to be at a leaf node, chosen from a range of values between 0.005 to 0.480 (of total number of subjects); and *maximum number of leaf nodes* controls the width of the tree at its leaf level, chosen from range of 2 to 20. For the extreme gradient boosting learner, the *number of tree estimators* was chosen from {50, 100, 150, 200}, the *maximum depth of the trees* from {2, 4, 6, 8}; the *learning rate* from {0.0001, 0.001, 0.01}; the *minimum child weight* (which is the minimum sum of instance weight that is needed in a child) from {1, 3, 5}; the *subsample* (which is the subsample ratio of the training instance) from {0.6, 0.7, 0.8, 0.9}; and *column sample by tree* is the subsample ratio of the columns when constructing each of the trees from {0.6, 0.7, 0.8, 0.9}[2].

To statistically compare the accuracy of our classifiers (on ADNI_HO) against each other and see if their classification rates are significantly different, we used the mid-p-value McNemar test [40] and reported the null hypothesis test result at $\beta = 0.05$ significance level, as well as the p-values. Any p-value smaller than β suggests rejection of the null hypothesis.

[1] See Elastic Net's API, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.ht

[2] See XGBoost's API, https://xgboost.readthedocs.io/en/latest/python/python_api.html

2.3.4 Feature Selection

First, note that the decision tree learner (similarly, extreme gradient boosting learner, which is a tree-based learner) has its own inherent way of choosing the best subset of features. At each internal node, this learner splits the available training instances based on the feature that best separates the class labels in terms of reducing the Gini impurity criterion. This process stops when the current node is sufficiently pure; this means the resulting decision tree will typically only use a small subset of the features.

The OL system also considered several approaches to learn yet simpler models, which involved fewer features. Here, it explored two *filtering* feature selection methods, each as a pre-processing step to reduce the number of features that are given to the base-learner (and hence the learned classifier): a simple univariate feature selector (UFS), and minimum redundancy maximum relevance (mRMR) [41]. Univariate feature selection method selects the top k features based on ANOVA [1] F-values [42]. This method first computes the F-value for each individual feature, and then selects those with top k^* values, using k^* found by grid-searches. Such univariate feature selection methods, however, do not consider the correlation between the features [42]. The mRMR method addresses this by sequentially seeking a set of features that maximizes the mutual information between each feature and the target classification variable while minimizing the mutual information between the currently-selected features.

For the linear models (SVM-linear and elastic net), S.OL also considered the "recursive feature elimination (RFE)" algorithm [43]: a *wrapper* feature selection method that sequentially removes the least important features, based on the value of learned linear weights – *i.e.*, initially the feature indexed by $i^* = \text{argmin}_{i \in \{1,..,34\}} |w_i|$. (There are fewer feature weights to consider in successive iterations.)

[1] Analysis of Variance

All of our feature selection methods – *i.e.*, UFS, mRMR and RFE – take as input, both the initial dataset and a number k^*, which is the number of features to use. To determine k^*, OL first computes the average cross-validation accuracy $\widehat{\text{acc}}(k)$ for each number of features $k \in \{1, 2, \ldots, 34\}$; then sets $a^* = \max_k\{\widehat{\text{acc}}(k)\}$ to be the most accurate, and $k^* = \arg\max_k\{\widehat{\text{acc}}(k)\}$ to be the associated value. To find this k^*, OL of course ran the feature selection methods "in-fold" – *i.e.*, determining the best size-k subset of features for each fold during training. Note that setting $k = 34$ in each of the of feature selection methods is equivalent to applying no feature selection.

Figure 2.1 summarizes our method, showing the combinations of base-learners and feature selection methods, within each version of OL.

2.4 Results

This section first describes our cross-validation results on the training set ADNI_TRAIN (composed of subjects' data from 48 acquisition sites), *i.e.*, the cross-validation accuracies of the best classifiers – one from S.OL and one from C.OL (Section 2.4.1). This analysis identified the best learners; we then ran just these two resulting classifiers on the independent held-out test set, ADNI_HO; those results appear in Section 2.4.2. Section 2.4.3 describes the features selected by the simple classifier, EN_7.

2.4.1 Cross-validation Accuracy on the Training Set, ADNI_TRAIN

Table 2.2 and Figure 2.2 show the mean and standard deviation of the cross-validation performance of our best simple and complex learners, EN_7 and rSVM_{23}. Note that the mean cross-validation accuracy, specificity, and sensitivity of the two models are close to each other – *i.e.*, within the boundaries of each other's error bars.

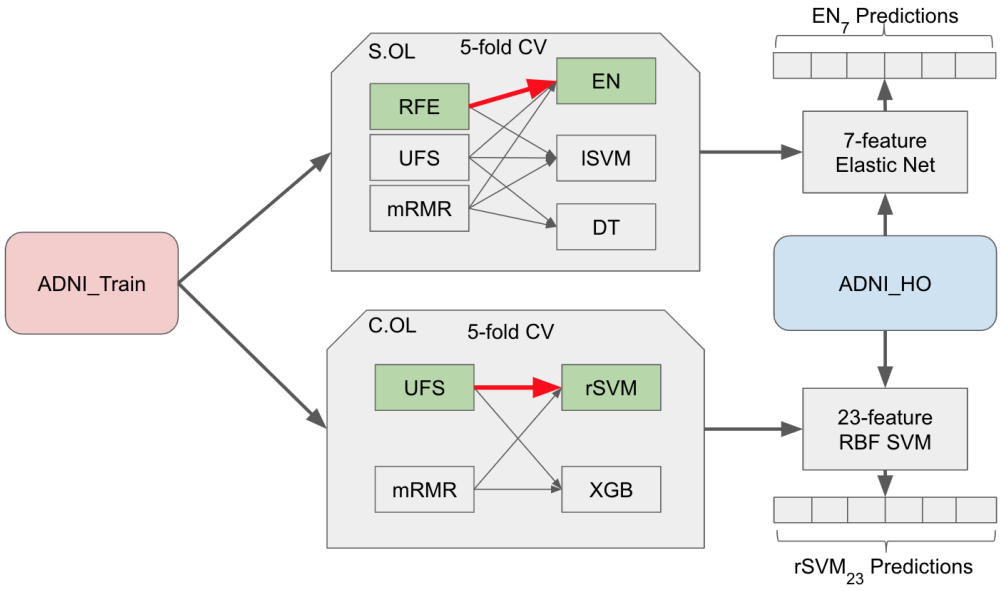

Figure 2.1: Our overall framework: Left-to-right is the training component, to produce both a simple classifier (here EN$_7$, on top, using S.OL) and a complicated classifier (here rSVM$_{23}$, on bottom, using C.OL). Each of these OLs considered a set of feature selection methods, from { RFE, UFS, mRMR }, and a given set of possible base-learners – S.OL considered EN, ℓSVM and DT, while C.OL: rSVM and XGB. (Note that DT did not use RFE). The RED arrow, in each, is the combination of feature selection method and base-learner with the best 5-fold cross-validation accuracy. We then evaluated each of these classifiers, by running each on the ADNI_HO (vertical, on right).

2.4.2 Results on the Held-Out Test Set, ADNI_HO

As described in Section 2.3.3, we twice ran our overall learner OL over the training set ADNI_TRAIN (composed of 48 sites, with a total of n=584 patients) to produce two classifiers – here, the elastic net model with 7-features, EN$_7$, and the RBF SVM model with 23-features, rSVM$_{23}$. Then, to evaluate and compare the effectiveness of these classifiers, we ran those classifiers on the held-out test set, ADNI_HO (composed of 12 sites, with a total of n=168 patients). Table 2.3 shows the test accuracies of these

	EN_7	$rSVM_{23}$
Accuracy	87.50 ± 1.76	87.67 ± 3.16
Specificity	83.56 ± 2.98	82.40 ± 5.44
Sensitivity	90.70 ± 3.19	91.93 ± 3.92

Table 2.2: Mean and standard deviation (STD) of the 5-fold cross-validation (CV) performance of EN_7 and $rSVM_{23}$ models

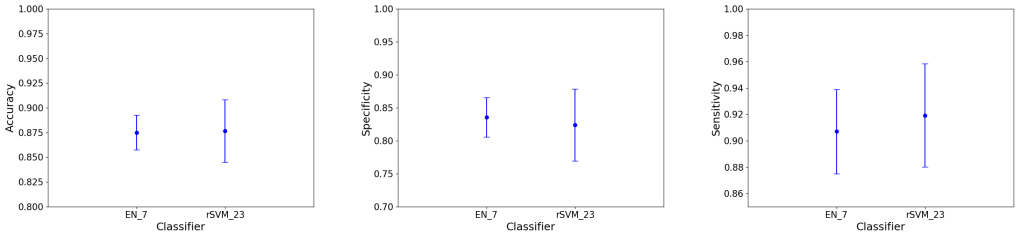

Figure 2.2: Mean and standard deviation (STD) of the 5-fold cross-validation (CV) performance of EN_7 and $rSVM_{23}$ models, on ADNI_TRAIN. The red dots show the hold-out performance of these models, on ADNI_HO.

two produced classifiers, along with the result of their statistical comparison, based on the McNemar test. Since the p-value of the McNemar test was above 0.05 (0.4531), no statistical difference was found between the accuracy of the simple and complex models on the held-out dataset (ADNI_HO).

2.4.3 Feature Importance, Based on EN_7

EN_7 selected APOE and 6 brain regions; Figure 2.3 shows the locations of the 6 regions, and Table 2.4 shows their associated weights, which corresponds to their "importance". Note that interestingly, the automatic feature selection results show 3 regions in both hemispheres (3 bilateral regions).

	Model		p-value
	EN_7	$rSVM_{23}$	
Accuracy	89.28	90.47	0.4531
Specificity	84.00	86.66	
Sensitivity	93.54	93.54	

Table 2.3: Test (hold-out) results using EN_7 and $rSVM_{23}$ models and the p-value of the statistical comparison of their accuracy based on McNemar test

Name	Location	Weight
Left hippocampus	top left	0.4221
Right hippocampus	bottom left	0.2896
Left inferiortemporal volume	top middle	0.3036
Right inferiortemporal volume	bottom middle	0.2535
Left entorhinal volume	top right	0.3802
Right entorhinal volume	bottom right	0.2181
APOE		-0.3621

Table 2.4: EN_7's weights for the features. (The 1st column refers to the name of the features and the 2nd column refers to the location in Figure 3.)

2.5 Discussion and Analysis

2.5.1 Performance of Simple Alzheimer's Disease Classification (EN_7)

In this study, we applied various machine learning algorithms to APOE genotype status, and regional grey matter volumes from 33 brain regions (that previous clinical studies have shown to be influenced by progression of AD) to learn a model that can predict Alzheimer's disease. We considered five base learners (including three simple models within S.OL). We also considered the effect of feature selection.

As noted in Section 2.1, there are many previous studies on prediction

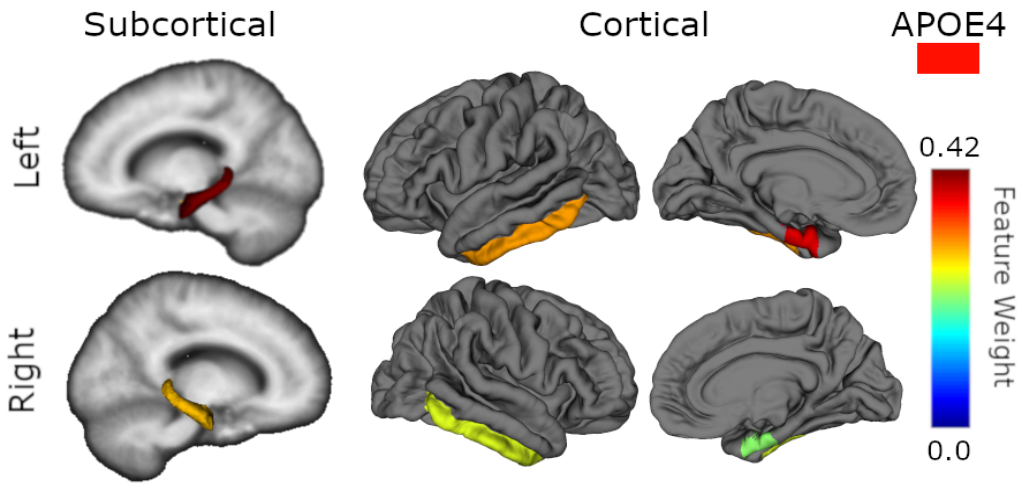

Figure 2.3: Locations of the features used by the EN$_7$ models. Color is based on the absolute value of EN$_7$'s weight of the feature.

of AD using structural MRI. Most of these previous studies concentrated on either achieving high prediction accuracy or mere simplicity (by using the grey matter volumes for a very small number of recognized brain regions), but this study is an attempt to create a balance between prediction accuracy and simplicity of the prediction framework. An earlier study demonstrated that a 3D convolutional neural network could achieve high accuracy (99%) using a large number of voxel-based features [18]. There were other region-based studies using structural MRI data, either combining a variety of measures with regional grey matter volumes, including cortical thickness, surface area and cortical curvature [20], or combining regional data from different imaging modalities [21], that can achieve high classification performances of 0.98% AUC and 93% accuracy, respectively. The problem with these approaches is that a system that involves too many features might not be used in a clinical environment. This means that even though they have high accuracies, clinicians may be uncomfortable using them because clinicians might find it difficult to understand

processes that involve a large number of features.

In another study, Jongkreangkrai *et al.* [22] learned an RBF-kernel SVM over bilateral hippocampus and amygdala grey matter volumes and entorhinal cortical thickness features. This achieved an AUC of 0.89, which is especially impressive as it used only 6 features. However, the resulting "SVM with RBF kernel" classifier involves a complex combination of the features, which prevents users from reasoning about the influence of each feature. By contrast, it is easy to reason about linear classifiers (Equation 2.1) as the sign of the coefficient w_i tells whether increasing that feature's value x_i (*mutatis mutandis*) increases the risk of AD, or decreases it; see Table 2.4.

2.5.2 Explaining EN$_7$'s Feature Selection Results

Table 2.4 shows the weights for the 7 features that appeared in EN$_7$. Jongkreangkrai *et al.* [22]'s 6-feature RBF SVM model, mentioned in Section 2.5.1, also used the cortical thickness for 4 of these brain regions: left and right hippocampus and entorhinal cortex. However, there is no easy way to read off the influence of a variable, nor even the directionality, in non-linear models, like RBF SVM or decision trees, in general. This is possible in linear models, such as EN: here finding a feature whose associated weight is positive, means the chance of AD increases as that variable's value increases, *mutatis mutandis*.

Previous studies on dynamics of grey matter loss in Alzheimer's disease suggest that bilateral hippocampus regions are areas of the brain that are most strongly affected by AD, which makes them appear as the most discriminating features in the classification task [44]. We also saw that the feature weights of bilateral regions are not similar to each other. This is consistent with the findings of clinical studies that claim grey matter loss in AD is asymmetric [45]. Studies claim that entorhinal cortex, which is the gateway to hippocampus, is one of the first areas that AD begins to af-

fect, which suggests that grey matter volume for this area may help identify patients at early stages of AD [46]. All of these 6 marked areas were located in the temporal lobe, which is consistent with the previous literature on diagnosis of AD [47]. Genetic studies show that the number of the APOE4 alleles at the APOE gene locus is strongly associated with late onset Alzheimer's disease [36], explaining why APOE genotype appears among discriminating features in our AD diagnosis prediction framework.

2.6 Conclusion

In this study, we attempted to build a classification framework for learning a *simple* model that can *accurately* distinguish patients with Alzheimer's disease from healthy controls. The performance results, on the 168 subjects in our test set, show that a learned simple linear classifier using only a small set of features – grey matter volume for 6 brain regions and a single genotype datum – can accurately distinguish Alzheimer's patients from controls. We found that the APOE genotype status had one of the highest feature importance in our EN_7 linear classifier (Table 2.4).

Although we started from 34 features that were already identified as relevant to AD, we provide a learned linear classifier using just 7 of these features that is statistically as accurate as its more complex counterparts. The best accuracy on this task and dataset in the literature (99%) [18] was achieved using a much more complicated, non-interpretable model (convolutional deep neural network). Our simple method, achieving 89% accuracy, approaches clinical relevance, which justifies future research into simple systems whose decision process would be accessible to clinicians and could help improve clinical diagnosis.

Chapter 3

Schizophrenia Prognosis Prediction

3.1 Introduction

Schizophrenia is a chronic psychotic disorder that often begins at young adult ages and persists for a lifetime, affecting 1% of the general population [48]. Due to the heterogeneous nature of this disease [49] and the intertwined symptoms of mental disorders [50], computational measures – based on patient's clinical history, neuroimaging scans, genetic profile or an ensemble of all – has become prevalent in the world of clinical science. Machine learning tools, with their ability to learn models that connect combinations of many variables to a label (e.g., disease state), can help clinicians diagnose this complicated disease in an automated manner, either without using any prior knowledge or with the help of domain-based knowledge of psychiatrists.

In addition to diagnosis – distinguishing schizophrenia patients from healthy controls – another challenge is prognosis – *i.e.*, , accurately predicting how a patient responds to a particular treatment, given his/her clinical or/and imaging profile. This question is especially important to

address because few of the current antipsychotic treatments consider the heterogeneous nature of schizophrenia and thus, few incorporate the delicate symptomatic distinctions between the patients, which leads to many hit-and-miss treatment policies [51].

Given the adverse side-effects of most of the antipsychotic drugs, accurate prediction of the impact of consuming a particular drug on a particular patient would be significantly valuable.

Recently, there has been a growing research on the statistical association between neuroimaging/clinical markers and treatment response. In a systematic review, Dazzan *et al.* [51] list 11 clinical studies, based on structural MRI data, that attribute particular markers in brain anatomy to treatment outcome (functional and symptomatic) of first-episode schizophrenia patients in the first 12 months of their treatment period. These anatomical markers include alterations in medial temporal and prefrontal cortical regions of the brain [51]. In a study on 76 schizophrenia subjects, Doucet *et al.* [52] studied the relationship between resting-state functional network (default mode network) connectivity (Z-transformed Pearson's correlation coefficients)/brain structure and 24-weeks clinical outcome of antipsychotic treatment. Despite not noticing any significant anatomical predictors, they found that functional network connectivity or higher internal cohesiveness of default mode network (a fMRI-based a.k.a functional predictor) is strongly linked to improvement in positive and anxiety/depression symptoms [52]. In another study, Sarpal *et al.* introduce baseline striatal connectivity index, calculated based on functional connectivity of 91 regions that are functionally connected to striatum, as a predictor of treatment outcome [53]. Siegel *et al.* studied the relationship between clinical variables – demographics and cognitive and symptomatic measures – and long-term (2–8 years) functional outcome in two groups of first-episode and previously treated schizophrenia patients. This study found that initial functioning level, positive/negative symptoms, gender,

education level and duration of illness are strongly associated with functional outcome [54]. In a study on 44 first-episode psychosis patients, Luck *et al.* found that abnormal fronto-temporal connectivity, measured by white matter fractional anisotropy in three main tracts that connect frontal and temporal regions of the brain is linked to poor short-term clinical outcome [55]. Altogether, these studies have observed associations between different baseline clinical/brain characteristics and treatment outcome in several different schizophrenic populations.

While there are many clinical univariate testing studies, seeking variables that are, by themselves, related to the treatment outcome, there are relatively few research studies that apply machine learning methodologies to build discriminative models to predict the treatment outcome for previously unseen schizophrenia patients [56]. This may be due to the unavailability of large public datasets with follow-up diagnostic measures. Cao *et al.* conducted one of the very first studies trying to predict treatment outcome in schizophrenia patients. In a study on a group of 43 drug-naive first-episode schizophrenia patients, using functional connectivity – correlation and mutual information – between bilateral superior temporal cortex and all the other cortical regions (68 regions in total), they were able to reach an accuracy of 82.5% by applying a linear support vector regression model to discriminate responders, defined as the subjects showing more than 30% reduction in their total PANSS score, from non-responders [57].

Our study explores ways to learn a prediction model of treatment response, defined by different measures of clinical and functional outcome, using patients' data collected at baseline: clinical data, functional magnetic imaging data, structural magnetic imaging data and diffusion tensor imaging data. Our general stacking procedure [58] combined best models from each data modality (if better than chance, independently). In addition to the responder vs non-responder classification experiments, we also explored several other questions: First, predicting follow-up posi-

tive/negative symptoms (9 symptoms in total) using a multitask artificial neural network regressor. One of the most clinically important questions regarding the prognosis of schizophrenia is to predict how each category of positive and negative symptoms change with consumption of the prescribed medication; we address this critical question using a multitask neural network. Second, predicting the time when the patients leave their medication using survival regression models. The motivation behind this task is that: with having an accurate estimation of the amount of time the patients adhere to their treatment, the clinicians can choose the type of medication and its dosage more insightfully. Third, classifying patients based on the type of their CGI severity progression in order to avoid putting the resistant or relapse patients under the risk of dealing with unnecessary side effects. Unfortunately, we could not achieve any significant results in any of these tasks, possibly due to the small number of patients.

3.2 Schizophrenia Dataset

We received this schizophrenia dataset from a hospital in India, which includes clinical, fMRI, sMRI and DTI data for schizophrenic patients, each with some of the follow-up scores: some patients only have CGI scores, some only SAPS/SANS, and some both. Based on the availability of follow-up severity measures, we created two versions of the dataset (see Figure 3.1). Here, we will describe each of the versions.

3.2.1 CGI Dataset

This dataset includes a total of 90 patients, whose CGI scores are available at a follow-up time point, ranging from 32 to 1099 days from the baseline visit. See Table 3.1 and Table 3.2 for more details on this dataset. We define two types of response for this dataset, each corresponding to a separate criterion; the "response", based on CGI-S criterion, is defined as having a

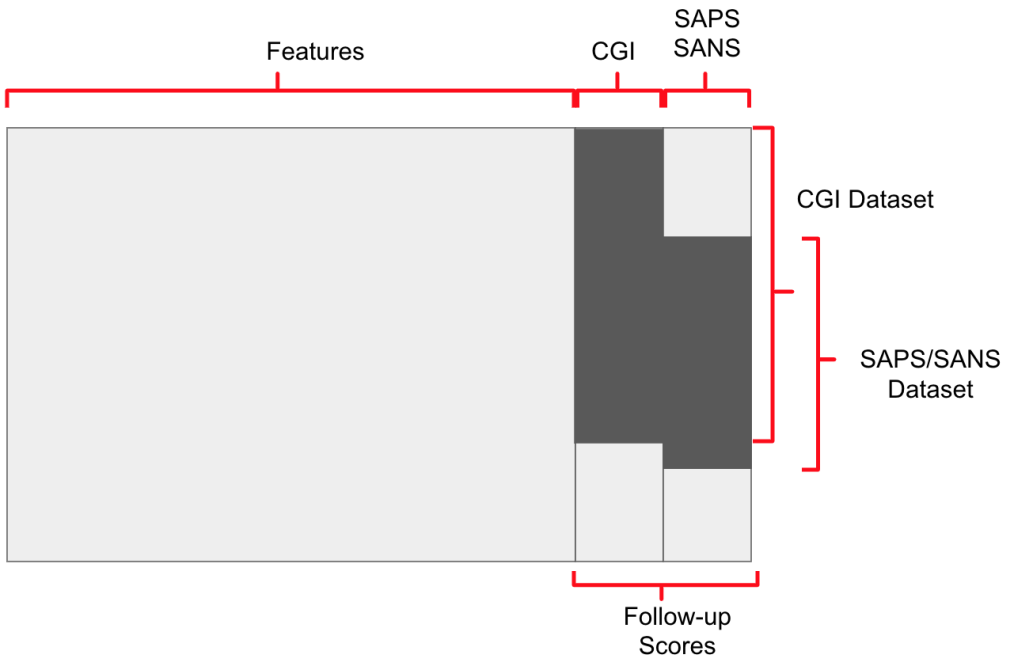

Figure 3.1: The two versions of the dataset featuring patients with available follow-up CGI scores and follow-up SAPS/SANS scores

follow-up CGI-S score that is less under 3 and based on CGI-I criterion, the "response" is defined as having a CGI-I score that is under 3.

3.2.2 SAPS/SANS Dataset

This dataset includes a total of 57 patients whose SAPS/SANS scores are available at a follow-up time point, ranging from 85 to 270 days. We define two types of response for this dataset, each corresponding to a separate criterion; the "response", based on SAPS criterion, is defined as full remission of positive symptoms (100% reduction in the total SAPS score - *i.e.*, having a follow-up total SAPS score of 0) and based on SANS criterion, it is defined as having a less-than-median (median across the patient population) follow-up total SANS score. See Tables 3.3 and 3.4 for more details

	R	NR	R	NR
	Clinical Data		fMRI Data	
Number	27♂ / 21♀	25♂ / 17♀	26♂ / 18♀	25♂ / 17♀
Age	30.69 ± 7.22	31.15 ± 6.37	30.40 ± 7.24	31.15 ± 6.37
Baseline CGI-S	4.60 ± 0.91	5.02 ± 0.89	4.59 ± 0.92	5.02 ± 0.89
	sMRI Data		DTI Data	
Number	27♂ / 20♀	25♂ / 17♀	27♂ / 18♀	24♂ / 16♀
Age	30.45 ± 7.10	31.15 ± 6.37	30.38 ± 6.79	31.01 ± 6.49
Baseline CGI-S	4.6170 ± 0.92	5.02 ± 0.89	4.64 ± 0.93	5.02 ± 0.89

Table 3.1: Demographic information for the responders (R) and non-responders (NR), based on follow-up CGI-S criterion. Numbers for age and baseline CGI-S are shown as mean $\pm$ STD.

	R	NR	R	NR
	Clinical Data		fMRI Data	
Number	45♂ / 29♀	7♂ / 9♀	44♂ / 26♀	7♂ / 9♀
Age	30.47 ± 6.86	32.21 ± 6.13	30.42 ± 6.65	31.96 ± 6.55
Baseline CGI-S	4.74 ± 0.92	5.06 ± 0.92	4.74 ± 0.92	5.06 ± 0.92
	sMRI Data		DTI Data	
Number	45♂ / 28♀	7♂ / 9♀	45♂ / 26♀	6♂ / 8♀
Age	30.47 ± 6.86	32.21 ± 6.13	30.42 ± 6.65	31.96 ± 6.55
Baseline CGI-S	4.75 ± 0.92	5.06 ± 0.92	4.77 ± 0.92	5.07 ± 0.91

Table 3.2: Demographic information for the responders (R) and non-responders (NR), based on follow-up CGI-I criterion. Numbers for age and baseline CGI-S are shown as mean $\pm$ STD.

on this dataset.

3.3 Data Preprocessing and Feature Extraction

3.3.1 Clinical Data Cleaning and Imputation

We began with 351 features, including: socio-demographic information, physical measurements, baseline functional and clinical measures, psychopathology, lifetime symptoms, family history and medication log (see

	R	NR	R	NR
	Clinical Data		fMRI Data	
Number	15♂ / 11♀	19♂ / 12♀	13♂ / 8♀	19♂ / 11♀
Age	30.85 ± 6.72	32.33 ± 7.00	31.13 ± 7.51	31.93 ± 6.44
BL SAPS	22.19 ± 7.91	31.25 ± 13.16	25.95 ± 11.31	27.83 ± 12.54
BL SANS	24.65 ± 21.67	35.80 ± 27.45	32.0 ± 30.07	31.36 ± 23.07
	sMRI Data		DTI Data	
Number	11♂ / 12♀	22♂ / 8♀	12♂ / 9♀	20♂ / 10♀
Age	30.18 ± 6.89	32.69 ± 6.47	31.89 ± 7.08	30.97 ± 6.36
BL SAPS	23.69 ± 10.00	29.23 ± 12.68	29.04 ± 12.57	25.1 ± 11.12
BL SANS	24.30 ± 22.60	35.13 ± 27.97	34.04 ± 30.25	28.86 ± 23.39

Table 3.3: Demographic information for the responders (R) and non-responders (NR), based on follow-up SAPS criterion. Numbers for age, baseline total SAPS score, and baseline total SANS score are shown as mean $\pm$ STD.

Figure 3.2).

We started cleaning the clinical dataset by removing the features with more than 30% missing values – this removed 62 features. After applying one-hot encoding to convert categorical nominal variables to numerical values, we eliminated the constant features. To put all of the medication measures into the same scale, we used the Chlorpromazine dose equivalents for each patient's medication entry. We condensed the 59 psychopathology entries (SAPS/SANS) into 9 categories (total hallucination score, total delusion score etc., as described in Section 1.2.2). Our final cleaned clinical dataset contained a total of 242 features, all viewed as real-valued – as even 1-hot encoded features are $\{0,1\}$.

After eliminating the columns with more than 30% missing values, we used Multivariate Imputation by Chained Equations (MICE) [59] to impute the remaining missing values in the clinical feature set, which we assume are missing at random (MAR) [60].

In the MICE method, at first, all of the missing values across the whole dataset are replaced by column-wise mean values; then, missing values

	R	NR	R	NR
	Clinical Data		fMRI Data	
Number	21♂ / 9♀	13♂ / 14♀	18♂ / 11♀	14♂ / 8♀
Age	30.26 ± 7.52	33.21 ± 5.78	32.60 ± 6.55	30.28 ± 7.13
BL SAPS	26.53 ± 12.64	27.77 ± 11.21	28.93 ± 12.05	24.59 ± 11.67
BL SANS	36.06 ± 27.12	24.77 ± 22.33	32.82 ± 23.53	30.04 ± 29.22
	sMRI Data		DTI Data	
Number	20♂ / 9♀	13♂ / 11♀	19♂ / 10♀	13♂ / 9♀
Age	30.98 ± 7.74	32.35 ± 5.27	33.26 ± 6.48	28.83 ± 6.03
BL SAPS	27.48 ± 13.09	26.04 ± 10.30	28.41 ± 11.93	24.5 ± 11.46
BL SANS	36.58 ± 28.16	23.0 ± 21.68	33.72 ± 25.58	27.40 ± 27.33

Table 3.4: Demographic information for the responders (R) and non-responders (NR), based on follow-up SANS criterion. Numbers for age, baseline total SAPS score, and Baseline total SANS score are shown as mean $\pm$ STD.

are sequentially set back to null, imputed and replaced by regression models – linear regression for continuous variables and logistic regression for discrete variables – using other observed variables [59]. For all of the classification methods, we used MICE as our imputation method, except for Extreme Gradient Boosting (XGBoost) learning algorithm, which provides its own way of choosing a default path for missing values and therefore, does not require prior imputation of missing values.

3.3.2 sMRI Preprocessing and Feature Extraction

We used SPM's CAT12 toolbox for preprocessing and feature extraction of structural MRI data in this study [61]. In CAT12's pipeline, data was first normalized to the MNI[1] standard space, and then segmented (hard segmentation) into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), using SPM's Tissue Probability Map (TPM) [61]. After this step, we applied spatial smoothing with a Gaussian kernel of $8mm$ width

[1]Montreal Neurological Institute, https://www.mcgill.ca/neuro/

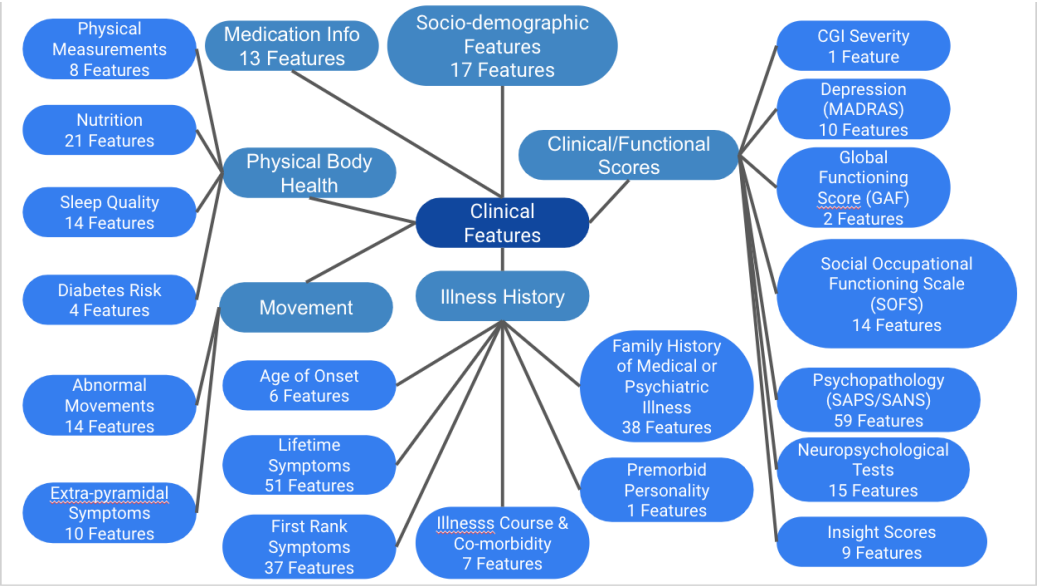

Figure 3.2: Clinical features diagram (lines represent inclusion relationship)

at half of the maximum value (full-width-half-maximum, or FWHM) to the modulated normalized grey matter images to reduce the noise and improve the signal-to-noise ratio (SNR).

We then parcellated the resulting smoothed modulated normalized grey matter images using 9 different brain atlases – AAL [62], MSDL [63], Harvard-Oxford cortical and subcortical atlases distributed with FSL (see FSL's website), Destrieux [64] and BASC (Bootstrap Analysis of Stable Clusters) multiscale atlases [65] – and computed the regional grey matter volume for each of the parcellations, which resulted in the 9 different sMRI feature sets that we used in our classification framework. The number of features in each of the created feature sets depends on the number of brain regions in the corresponding brain atlas. To compensate for the possible existing variance in the brain sizes of individuals, we regressed out age, gender and total intracranial volume (TIV) as covariates. In this residual approach, we used a linear regression between the regional volumes and

these 3 covariates – *i.e.*, age, gender and total intracranial volume – to predict the normalized volumes [66].

3.3.3 fMRI Preprocessing and Feature Extraction

Preprocessing is a critical step in fMRI data analysis to remove uninteresting variations in data, caused by various sources of noise, and preparing the data for feature extraction and statistical analysis [67]. We used the DPARSF toolbox (Advanced version) [68] for preprocessing our resting-state fMRI dataset. DPARSF is a SPM-based Matlab toolbox for pipeline processing of resting-state fMRI data, providing a wide variety of measures, derived from the signal. For parcellation of preprocessed brain images and extracting region-based measures, using a Nilearn[1] script written by Dr. Sunil Kalmadi, we created 13 different brain parcellations, based on 13 brain atlases – AAL [62], MSDL [63], Harvard-Oxford cortical and subcortical atlases distributed with FSL (see FSL's website), Destrieux [64], BASC (Bootstrap Analysis of Stable Clusters) multiscale atlases [65], Smith [69], Craddock [70], Power [71] and Dosenbach [72]. In this section, we briefly explain the steps of our fMRI preprocessing pipeline. Figure 3.3 presents the pipeline.

Merging Since the functional magnetic resonance imaging data is acquired in form of 3D images, each belonging to a particular time point, these 3D images were merged to form a 4-dimensional fMRI image for each patient, as the beginning step.

Removing First 10 Time Points Each of our 4D functional images originally consisted of 153 3D volumes, but we removed the first 10 volumes to account for the patients' adjustment to the scanner noise [68], leaving 143 volumes for each subject.

[1] A machine learning library for neuro-imaging in Python, https://nilearn.github.io/

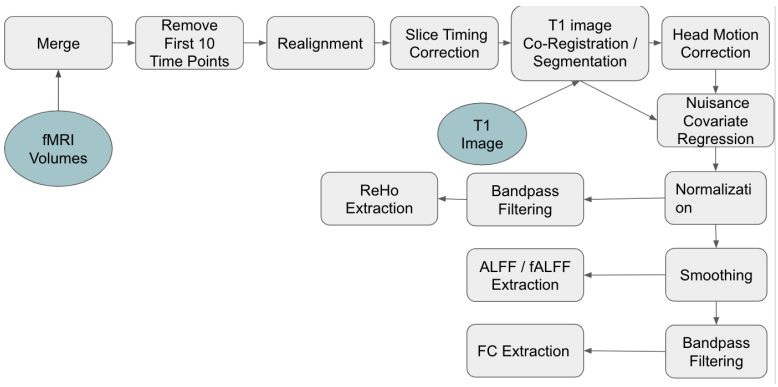

Figure 3.3: fMRI preprocessing / feature extraction pipeline

Slice Timing Correction In fMRI acquisition, each of the whole brain volumes are acquired one slice at a time, within the repetition time (TR = 2s) [68]. Because these slices are acquired at different time points (within each TR), we corrected the timing when merging them into a whole brain volume using the slice ordering information, a parameter in scanning time [68].

T1 Image Co-Registeration and Segmentation In this step, first the T1 image was co-registered to the mean realigned functional image and was then segmented into different tissues (grey matter, white matter and CSF) using unified T1 segmentation [73].

Realignment and Head Motion Correction When head moves during the scanning time, the location of the brain volumes varies from one time point to another. The goal of head motion correction is to reposition all of the volumes in the time series to the same location [67]. In this step, all of the volumes were realigned and co-registered to a reference volume (average volume) using Friston 24-parameter model [74] and then, the effects of head motion (24 parameters) were regressed out from the signal. Also, during this step, the individuals with excessive head movements were identified and removed from the study.

Nuisance Covariate Regression To remove the impact of non-neuronal and psysiological artifacts from the fMRI signal, mean global signal, CSF, and white matter signal were linearly regressed out, as covariates. In this residual approach, a linear regression between the BOLD signal and these 3 covariates – *i.e.*, mean global signal, CSF, and white matter signal – is used to predict the cleaned BOLD signal and remove possible artifacts. In recent studies, mean global signal or whole-brain signal is associated with the effect of respiration [68].

Normalization Because the size, shape and brain anatomy of individuals is different and this factor might influence the result of statistical analyses, all of the scans are spatially normalized to a standard brain template [68]. Using DARTEL tool [75], each functional brain scan was nonlinearly co-registered to MNI standard space.

Smoothing Spatial smoothing with a Gaussian kernel of $4mm$ width at half of the maximum value (full-width-half-maximum, or FWHM) was applied to the functional images to reduce the noise (random Gaussian noise) and improve signal to noise ratio (SNR).

Bandpass Filtering Most studies associate low-frequency fluctuations of fMRI signal between 0.01Hz-0.08Hz to neuronal activity that originates from grey matter tissue of the brain, while associating the higher frequencies to white matter and non-neuronal activities [68]. Therefore, a bandpass filter (0.01Hz-0.08Hz) was applied to the data to reduce the effect of non-neuronal artifacts.

Extracting Regional Homogeneity Regional homogeneity (ReHo) is a measure of regional spontaneous neuronal activity that is measured for each voxel by calculating Kendall's coefficient of concordance (KCC) of time series of the voxel with those of its 27 nearest neighbors (sharing the

same faces, edges, or corners) [76]. Once the KCC for each voxel is calculated, it is divided by the global mean regional homogeneity value within the whole-brain mask to reduce global variability effects [68]. Regional Homogeneity of a region is the average of the ReHo values of its voxels.

Extracting ALFF and fALFF ALFF, which is the power spectral density or mean square root of the power spectrum of low-frequency fluctuations (LFF, 0.01Hz-0.08Hz), is another measure of regional spontaneous neuronal activity; however, since some studies have shown that ALFF is prone to large fluctuations of high frequency physiological noise, Zou *et al.* [77] introduced a more robust measure, fALFF, as the ratio of total amplitude within the low-frequency range to the total amplitude of the entire detectable frequency range. Since the predictive value of these two measures in between-group studies is still unknown [68], we extracted both and used them in the classification framework.

Extracting Functional Connectivity Measures Functional connectivity is a measure that is commonly used in analysis of resting-state functional MRI, taking into account the inter-regional correlations in the BOLD signal [68]. Using the image after realignment, head motion correction, nuisance removal, normalization, smoothing and bandpass-filtering steps, we computed the inter-regional **correlation**, **partial correlation**, **precision** and **covariance** measures, using Nilearn library.

Extracting Functional Graph Characteristic After extracting functional correlation matrix, we converted it to a binary matrix, representing a bi-directional graph, known as brain's functional graph, by thresholding it by a cut-off value of 0.7 (chosen by trying values from $\{0.7, 0.75, 0.8\}$) on the absolute value of entries. We included local (nodal) characteristics of the graph in our feature set, including node degree, which is the number of neighbors of each node (region), and local efficiency, which is the average

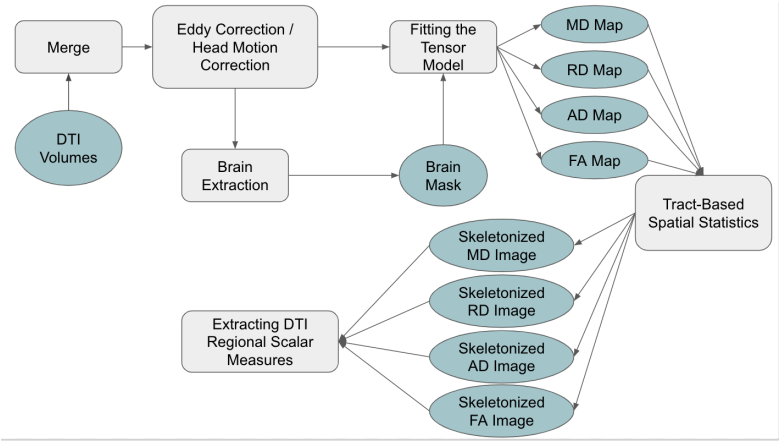

Figure 3.4: DTI preprocessing / feature extraction pipeline

inverse shortest path length in the sub-network that is created by a node's (region) immediate neighbors [78].

3.3.4 DTI Preprocessing and Feature Extraction

For preprocessing and feature extraction of DTI data, we followed Engima Group's DTI processing protocols and templates. All the preprocessing steps were performed in FSL through Nipype's pipelines and interfaces. In this section, we briefly explain the steps of our DTI preprocessing pipeline. The pipeline is shown in Figure 3.4.

Merging As the diffusion tensor imaging data is acquired in form of 3D images (65 images, in this dataset), each belonging to a particular direction, these 3D images were merged to form a 4-dimentional DTI image, as the beginning step. The first 3 dimensions show the spatial position in the brain and the fourth dimension shows each of the scalar diffusion-sensitizing directions that have been applied during acquisition.

Eddy Current and Head Motion Correction The two main artifacts in DTI acquisition are eddy current distortions and head motion [79]. In DTI, the gradients are much longer than in other magnetic resonance imaging acquisitions, which means there might be changes in the local magnetic field that result in induction of circular electric currents in the different conducting surfaces of the MRI scanner creating image distortions [79]. Eddy currents vary with the diffusion gradient direction and, therefore, there might be misregistration between successive volumes [79]. We used FSL's eddy correction method, which removes both eddy current distortions and head motion artifacts – by rigid-body co-registration to a reference volume (average volume).

Brain Extraction Brain extraction or brain/non-brain segmentation is considered as the preparation step for registration and segmentation. We used FSL's BET tool that is reportedly capable of removing non-brain tissues with various contrasts and geometries (skull, marrow and etc) and is robust to local intensity changes [80]. We set the brain extraction threshold manually by trying different values and picking the value that worked the best (0.4) by visual checking. This step creates a binary mask for brain/non-brain separation.

Fitting the Tensor Model In this step, using FSL's dtifit, a diffusion tensor model was fit at each voxel, as described in Section 1.1.2. The resulting model includes the 3D maps of 1st, 2nd and 3rd eigenvectors and eigenvalues, Mean Diffusity (MD) and Fractional Anisotropy (FA) with voxelwise entries. Axial Diffusivity (AD) is the first eigenvalue (*i.e.*, largest eigenvalue) and Radial Diffusivity (RD) is the average of second and third eigenvalues ($\frac{\lambda_2+\lambda_3}{2}$) [3]. MD, a.k.a the apparent diffusion coefficient (ADC), is the average of tensor's eigenvalues and relates to the total amount of diffusion in a voxel [3]. FA measures are ratios of the eigenvalues that are used to quantify the shape of the diffusion and is basically a normalized

variance of the eigenvalues [3].

$$FA = \frac{\sqrt{(\lambda_1 - MD)^2 + (\lambda_2 - MD)^2 + (\lambda_3 - MD)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{3.1}$$

where λ_1,λ_2, and λ_3 are the 3 eigenvalues [3]. FA is commonly used as a measure of white matter integrity.

Tract-Based Spatial Statistics Using FSL's TBSS method [81] and following the Enigma protocol, we registered and skeletonized all of the 3D FA images (maps) to Enigma's DTI atlas. In this step, the images were first slightly eroded and zero padded to remove the possible outliers from the tensor model fitting; then, they were non-linearly registered to the Enigma DTI template and linearly registered to the MNI space. All of the standardized FA images were masked using Engima's mask template and consequently, skeletonized by projecting the ENIGMA skeleton onto them. These steps were repeated for dtifit's MD, AD and RD maps, creating individual-specific skeletonized AD, MD, FA and RD images.

Extracting DTI Regional Scalar Measures Using Engima's protocol for ROI analysis, for each of the FA, MD, AD and RD measures, mean ROI values were extracted based on the JHU atlas (ICBM-DTI-81 white-matter labels atlas) parcellation, alongside an average value across the entire skeleton. In the second feature extraction step, information from the first output was used to average related (e.g., average of L and R external capsule) regions to get an average value weighted by volumes of the regions – AD_avg, FA_avg, MD_avg and RD_avg.

3.4 Machine Learning Methods

3.4.1 Unsupervised Dimensionality Reduction for Functional Connectivity Features

Among the set of fMRI features that we used in our schizophrenia prognosis prediction experiments, functional connectivity features – *i.e.*, , correlation, partial correlation, covariance and precision – are high-dimensional feature vectors of size $\frac{n\times(n-1)}{2}$, where n is the number of regions in a given brain atlas that is used for parcellation. Considering the size of our datasets, finding ways to reduce the dimensionality of these large feature vectors before using them in the classification framework is absolutely necessary, in order to overcome the curse of dimensionality. For these features, we tried two unsupervised methods to reduce the dimensionality: First, a linear compression method, Prinicipal Component Analysis (PCA) and second, a non-linear transformation method, denoising autoencoder.

In the first approach, we simply reduced the dimensionality of the dataset to $m-1$, where m is the number of instances, using PCA based on Singular Value Decomposition (SVD) [82]. In the second approach, we used a denoising autoencoder structure to learn a compact representation of the data that produces the least reconstruction error a.k.a. L2 loss (see Section 3.6). For the framework, described in Sections 3.4.2 to 3.4.5, all of the fMRI-driven functional connectivity features are compressed using PCA, after standard scaling to have a mean value of 0 and a standard deviation of 1.

3.4.2 Base Learners

In this framework, for each of the data modalities – fMRI, sMRI, DTI and clinical – we tried 6 different classes of base learners, including tree-based methods: *extreme gradient boosting* (XGB) and *random forest* (RF), linear

methods: *linear support vector machine* (ℓSVM) and *logistic regression with lasso penalty* (lasso), and kernelized support vector machines: *SVM with RFB kernel* (rSVM) and *SVM with polynomial kernel* (pSVM). We briefly described support vector machines and extreme gradient boosting in Section 2.3.2. Lasso is similar to elastic net, described in Section 2.3.2, but it only uses L1 penalty (*L1_ratio* = 1); we chose the lasso penalty due to the high dimensionality of our feature space. Similar to the extreme gradient boosting learner, random forest [83] is an ensemble tree-based method, but instead of boosting, it uses bagging, which grows each tree from a random selection (without replacement) of instances in the training set and then, combines their results.

3.4.3 Feature Selection

Each of these learners were accompanied by a feature selection method to improve their computational speed and also, to avoid overfitting. For linear methods, ℓSVM and lasso, we performed feature selection by thresholding the absolute value of the linear weights and pruning the features whose absolute weight is less than the selected threshold. Due to its non-recursive approach, thresholding the linear weights is a faster alternative for the RFE method, described in Section 2.3.4, when we have a high-dimensional feature space. The thresholds were chosen from {median, 2*median, 3*median, 4*median, 5*median} of the absolute value of the weights.

For SVM with RBF or polynomial kernels, we performed univariate feature selection (UFS), based on the mutual information [84] between each of the features and the labels. We used the mutual information measure for univariate feature selection, instead of the standard F-value or chi2 statistics, in order to be able to cover positive, negative, discrete and continuous features of our heterogeneous dataset. The number of features for univariate feature selection was chosen from $\{10, 20, 30, 40, 50\}$.

The tree-based learners, XGBoost and random forest, have their own inherent way of choosing the best subset of features. At each internal node of each tree, the tree splits the available training instances using the feature that best separates the class labels in terms of reducing the Gini impurity criterion. This process stops when the current node is sufficiently pure; this means the resulting tree will typically only use a small subset of the features. These ensemble tree-based learners then combine the results of these trees using bagging (random forest) or boosting (XGBoost).

3.4.4 Evaluation

We ran our "overall learner" OL[1] 24 times, over different sets of base-learners – ℓSVM, EN, rSVM, pSVM, RF, and XGB – and different data modalities – clinical, sMRI, DTI, and fMRI; each of these runs produced a classifier.

In each iteration of our 10×3 external cross-validation (10-fold cross-validation, repeated 3 times), we divided the dataset into a train (90% of instances) and a test set (10% of instances). OL partitioned the train set into 5 internal folds and used the same internal folds for all of the runs. OL then used the 5-fold cross-validation accuracies of these base-learners (accompanied by various feature selection methods, described in Section 3.4.3) to identify the best feature hyperparameters (described below), feature selection hyperparameters (described in Section 3.4.3), and base-learner hyperparameters (described below). It then ran that best model on the entire training set to produce classifiers; we then tested the learned classifiers on the test set. As for the choice of evaluation metric, for predicting response based on CGI-S, SAPS and SANS scores, because the dataset was almost balanced (55%-60% baseline accuracy), we used accuracy as our

[1] OL is the system that invokes the pre-processing steps, etc., before running the base-learners. It also does the randomized-searches [85] seeking the best learning algorithm, feature, and feature selection parameter settings, based on the performance over the training data; see the text.

performance measure, while for predicting CGI-I score, we used balanced accuracy:

$$Balanced\ \ Accuracy \quad = \quad \frac{Specificity + Sensitivity}{2} \tag{3.2}$$

since the dataset was highly imbalanced toward the positive class. In this case, we also used a weighted cost function for all of the base-learners to compensate for the imbalance in our dataset by penalizing false positives more than false negatives. In this approach, we weighted the false positive cost by the ratio of the number of positive instances to the number of negative instances.

For SVM methods, the C hyper-parameter was chosen from {1E-5, 1E-4, ..., 1E3, 1E4}, γ was chosen from {1E-6, 1E-5, ..., 1E1, 1E2}, and degree of the polynomial kernel from {1, 2, 3, 4, 5}. For lasso, the α hyper-parameter was chosen from {1E-4, 1E-3, 1E-2, 1E-1, 2E-1, ..., 9E-1} values and *L1_ratio* was set to 1 [1], which means the learner only applies L1 regularization (aka Lasso classification). For the random forest base learner, the *number of tree estimators* was chosen from {50, 100, 150, 200, 300, 500}; the *maximum depth of the trees* from {2, 4, 6, 8}; *minimum samples split* is the minimum percentage of training set instances required to split an internal node, chosen from a range of values between 0.005 to 0.480 (of total number of samples); *minimum samples leaf* is the minimum percentage of instances required to be at a leaf node, chosen from a range of values between 0.005 to 0.480 (of the total number of samples); and *maximum number of leaf nodes* controls the width of the trees at their leaf level, chosen from range of 2 to 20. For the extreme gradient boosting learner, the *number of tree estimators* was chosen from {50, 100, 150, 200}, the *maximum depth of the trees* from {2, 4, 6, 8}; the *learning rate* from {0.0001, 0.001, 0.01}; the *minimum child weight* (which is the minimum sum of instance weight that is needed in a child) from {1,

[1] See Elastic Net's API, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.ht

3, 5}; the *subsample* (which is the subsample ratio of the training instance) from {0.6, 0.7, 0.8, 0.9}; and *column sample by tree* is the subsample ratio of the columns when constructing each of the trees, chosen from {0.6, 0.7, 0.8, 0.9}[1].

In this classification framework, our overall learner tuned the feature hyperparameters alongside the base learner and feature selection hyperparameters, based on average 5-fold inner cross-validation performance using randomized search [85] in the hyperparameter space. Feature hyperparameters include choice of brain parcellation atlas for sMRI and fMRI data and choice of feature type for fMRI and DTI data. Brain parcellation atlas for sMRI was chosen from {AAL, MSDL, Harvard-Oxford cortical atlas, Harvard-Oxford subcortical atlas, Destrieux, BASC multiscale atlases (4 atlases)}; brain parcellation atlas for fMRI was chosen from {AAL, MSDL, Harvard-Oxford cortical atlas, Harvard-Oxford subcortical atlas, Destrieux, BASC multiscale atlases (4 atlases), Smith, Craddock, Power, Dosenbach}; feature type for fMRI was chosen from {correlation, partial correlation, covariance, precision, ALFF, fALFF, ReHo, node degree, local efficiency}; and feature type for DTI was chosen from {FA, MD, AD, RD, FA_avg, MD_avg, AD_avg, RD_avg}. The workflow of our overall framework is shown in Figure 3.5.

3.4.5 Combining Modalities

To create a multimodel model, we used a stacking classifier to combine the best models from each modality, based on average internal cross-validation performance. We avoided simple concatenation of the feature sets for two reasons. First, our small sample size already imposes a significant challenge on the classification task, which would even get worse after concatenating the feature sets. Second, our feature sets include features with different scales and types – continuous, categorical nominal, and categori-

[1]See XGBoost's API, https://xgboost.readthedocs.io/en/latest/python/python_api.html

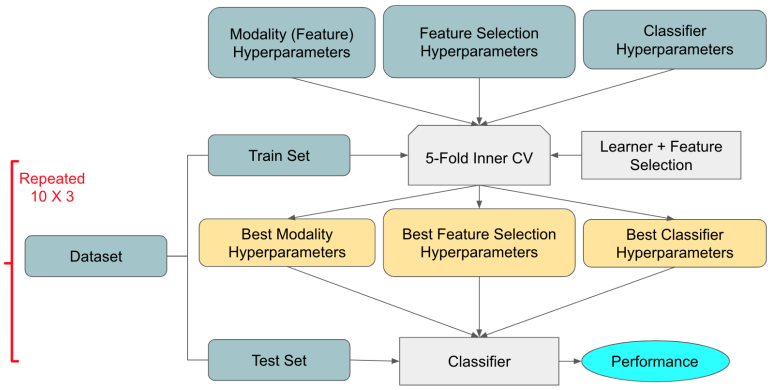

Figure 3.5: Overall framework for schizophrenia prognosis prediction

cal ordinal – and if the model that we are using is not a tree-based learning algorithm (XGBoost and random forest), it might be adversely affected by these scale/type discrepancies. So, after finding the best model for each modality, using 5-fold inner cross-validation and randomized search, we combined their predictions into a new train and test set and trained the meta-classifier (XGB) on the new train set and evaluated it on the new test set (for each iteration of 10×3 outer cross-validation). The advantage of using a stacking algorithm over averaging the decisions made by each of the models or manual voting (weighting the decisions) is that the stacking algorithm automatically tunes the decision weights, based on the predictive power of each model (in our case, each modality). Our stacking framework is shown in Figure 3.6.

3.5 Prognosis Prediction Results and Discussion

Using the classification framework, described in Sections 3.4.2 to 3.4.4, we tried to classify patients into {responder, non-responder}, based on each of the treatment response criterion, described in Section 1.2. The classification results for each prediction task are shown in **Tables 3.5** to **3.8**. As it

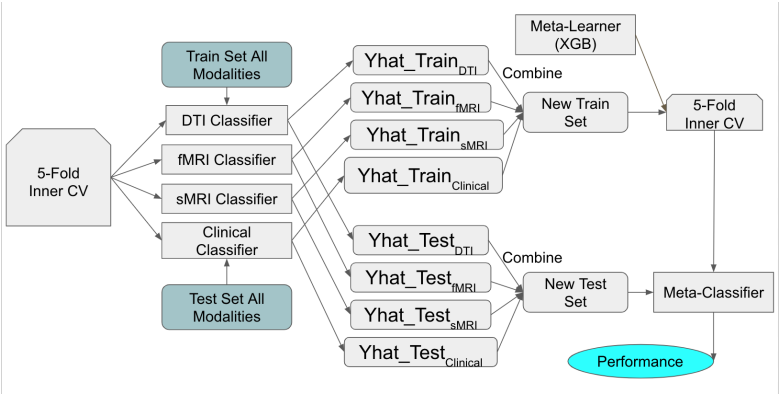

Figure 3.6: Stacking method for combining modalities

is observable in these tables, none of the results were significantly above the baseline performance. The huge errorbars could be mainly attributed to the small size of our dataset. Another factor that might be affecting the results is our missing information on the adherence level of individuals to the prescribed medication. While we know that the medications were prescribed for the patients by an expert clinician, the patients were not hospitalized to take the medications under an authorized supervision. In addition, in our dataset, there were significant differences between the medication type and follow-up duration of patients that might be influencing the outcome variable, particularly when using neuroimaging features per se for the classification task.

	Clinical Baseline 50%	fMRI Baseline 50%	DTI Baseline 50%	sMRI Baseline 50%	Multimodal Baseline 50%
	Accuracy Baseline 82.22%	Accuracy Baseline 81.39%	Accuracy Baseline 83.52%	Accuracy Baseline 82.02%	
XGBoost	60.92 ± 21.69	42.38 ± 15.42	43.45 ± 17.16	54.16 ± 19.01	
Random Forest	45.62 ± 16.78	44.88 ± 14.24	39.61 ± 12.43	54.37 ± 16.02	
Linear SVM	43.21 ± 18.46	43.69 ± 21.75	43.45 ± 17.16	36.45 ± 19.37	
RBF SVM	49.52 ± 16.95	48.57 ± 12.07	62.11 ± 16.68	51.48 ± 19.04	
Poly SVM	49.73 ± 14.14	53.80 ± 23.52	45.83 ± 17.40	40.74 ± 17.43	
Lasso	45.44 ± 12.19	45.95 ± 23.26	50 ± 0.0	50.65 ± 15.21	
Stacking Classifier					48.25 ± 15.28

Table 3.5: Treatment response prediction results, based on CGI-I score (CCG-I $\leq$ 2 or not). The performance measure for this task is balanced accuracy and the results show average 10-fold $\times$ 3 (repeated 3 times) balanced accuracy.

3.6 Other Experiments

We also investigated other clinically relevant questions with regards to the dataset, in addition to prognosis prediction experiments, described in Section 3.5. Here, we briefly explain these experiments.

Using denoising autoencoder for fMRI dimensionality reduction For this approach, we picked regional functional correlation features, based on

	Clinical Baseline 53.33%	fMRI Baseline 51.16%	DTI Baseline 52.95%	sMRI Baseline 51.69%	Multimodal
XGBoost	56.45 ± 13.74	40.13 ± 12.59	44.49 ± 12.08	39.90 ± 12.79	
Random Forest	50.99 ± 12.23	52.06 ± 15.88	49.25 ± 15.19	42.96 ± 14.67	
Linear SVM	50.75 ± 8.77	54.82 ± 16.58	43.65 ± 13.00	46.47 ± 14.84	
RBF SVM	47.50 ± 12.62	43.14 ± 19.38	42.54 ± 13.47	41.27 ± 17.25	
Poly SVM	48.67 ± 16.25	46.34 ± 13.56	49.90 ± 10.12	48.85 ± 13.83	
Lasso	47.67 ± 9.40	56.26 ± 15.64	49.81 ± 3.92	46.89 ± 10.52	
Stacking Classifier					43.33 ± 15.28

Table 3.6: Treatment response prediction results, based on CGI-S score (CCG-S $\leq$ 2 or not). The performance measure for this task is accuracy and the results show average 10-fold $\times$ 3 (repeated 3 times) accuracy.

aal parcellation. The encoder architecture includes $\frac{116\times115}{2}$ = 6670 nodes in the input layer, 1000 and 100 nodes in the first and second hidden layers, respectively, and $\{10, 20, 30, 40, 50\}$ nodes in the code or representation layer. The decoder architecture includes 100 and 1000 nodes in the first and second hidden layers, respectively and 6670 nodes in the output layer. We used soft sign ($f(x) = \frac{x}{1+|x|}$) activation function, which produces values in $(-1, 1)$ interval, the same range as functional correlation values and does not saturate easily (we also tried relu as the activation function for hidden units, but this increased the L2 reconstruction loss). The architecture of a denoising autoencoder is shown in Figure 3.7. Unfortunately, we couldn't achieve any positive results using these dimensionality-reduced fMRI datasets.

	Clinical Baseline 54.39%	fMRI Baseline 56.87%	DTI Baseline 54.91%	sMRI Baseline 54.72%	Multimodal (Clinical-DTI) Baseline 54.91%
XGBoost	65.88 ± 26.37	46.72 ± 17.62	67.72 ± 21.66	45.44 ± 22.86	
Random Forest	61.93 ± 25.58	49.77 ± 18.78	55.49 ± 20.78	39.83 ± 22.65	
Linear SVM	47.34 ± 18.66	47.55 ± 20.71	42.66 ± 16.97	51.50 ± 19.22	
RBF SVM	58.22 ± 21.22	50.55 ± 14.88	47.55 ± 15.74	58.00 ± 21.24	
Poly SVM	47.05 ± 12.01	53.77 ± 15.22	47.05 ± 12.01	46.77 ± 23.90	
Lasso	49.66 ± 7.06	45.16 ± 21.21	49.66 ± 7.06	54.11 ± 15.12	
Stacking Classifier					55.05 ± 23.39

Table 3.7: Treatment response prediction results, based on follow-up SAPS score (CCG-I $\leq$ 2 or not). The performance measure for this task is accuracy and the results show average 10-fold $\times$ 3 (repeated 3 times) accuracy.

Follow-Up 9-Score SAPS/SANS Prediction Using Multitask Regression
One of the most clinically import questions regarding treatment outcome prediction in psychosis is to predict how each of the 9 subcategories of SAPS/SANS – hallucinations, delusions, bizarre behavior, positive formal thought disorder, affective blunting, alogia, avolition, anhedonia, and attention – are influenced by the prescribed treatment. For this experiment, we used a multitask artificial neural network (ANN) architecture with an additional hidden layer, shared among all the tasks and trained simultaneously for all the tasks, to be able to use domain-specific information learned from one task to improve the results of all other related tasks [86]. Figure 3.8 shows the architecture of this ANN. In this experiment, we used a subset of our original dataset for which the follow-up SAPS/SANS scores are available (n = 51 to 57, depending on the data modality). But,

	Clinical Baseline 52.63%	fMRI Baseline 50.98%	DTI Baseline 50.98%	sMRI Baseline 50.94%	Multimodal
XGBoost	57.11 ± 19.25	65.83 ± 19.24	48.27 ± 20.83	46.38 ± 22.83	
Random Forest	49.44 ± 13.93	66.50 ± 19.28	43.38 ± 17.10	49.77 ± 18.77	
Linear SVM	51.66 ± 17.69	52.77 ± 17.34	34.55 ± 16.75	31.38 ± 21.14	
RBF SVM	54.33 ± 16.55	**68.27 ± 19.37**	44.44 ± 12.95	49.11 ± 18.00	
Poly SVM	45.55 ± 15.13	53.44 ± 24.30	38.88 ± 18.54	42.77 ± 13.46	
Lasso	56.11 ± 18.50	60.77 ± 18.13	49.66 ± 3.14	48.61 ± 14.27	
Stacking Classifier					54.77 ± 24.83

Table 3.8: Treatment response prediction results, based on follow-up SANS score (SANS $\leq$ 2 or not). The performance measure for this task is accuracy and the results show average 10-fold $\times$ 3 (repeated 3 times) accuracy.

because the multitask assumption biases the learner toward hypotheses that explain more than one task [86], we also tried an alternative architecture, eliminating the shared representation layer; but we were not able to reach a significant performance – measured by computing a separate R^2 error for each of the tasks – despite considering several different network architectures, number of hidden layers, number of hidden units, loss functions and activation functions.

Classifying patients, based on type of CGI Severity Progression In this experiment, we first tried to label the patients, based on the pattern of change in their CGI severity scores using available follow-up data at 1-3 time points (See Figure 3.9). We labeled the patients with a consistent decrease in their CGI severity as responders (n = 61), the patients with either

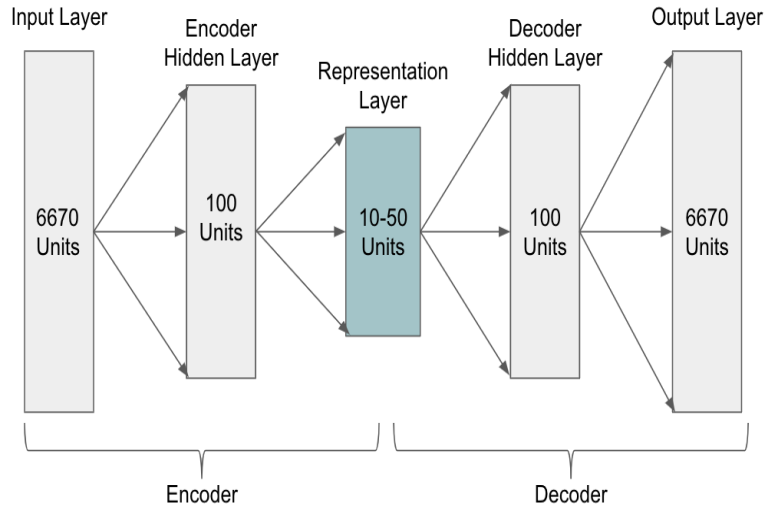

Figure 3.7: Denoising autoencoder architecture

no change or increase in their CGI severity as resistant (n = 10) and the patients switching between increase/decrease/no change (n = 19) as relapse patients. In Figure 3.9, subject S0004 represents a responder; subject S0003 represents a resistant and subject S0034 represents a relapse patient. We then tried to classify responder versus resistant versus relapse patients using a combination of MRI and clinical features in the same classification framework as described in Section 3.4, but we were not able to reach any

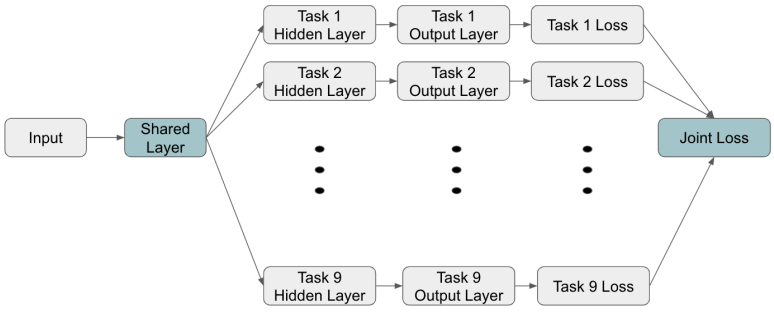

Figure 3.8: Multitask neural network architecture

significant results.

Survival regression Our CGI dataset, recorded 19/90 patients who claim to have stopped their medication at some point in their follow-up visits. First, we tried to classify these patients, who have a tendency of leaving medication, from adherent patients, but we were not able to obtain a significant performance on this task. Thus, we took a survival regression approach to predict the time a patient leaves medication. For the uncensored patients, we consider the first time point when they claim to have left medication as the the event (in this case, leaving medication) time and for the censored patients, we consider the last time point that they were recorded to be taking medication as the event time.

Applying various survival regression algorithms including multi-task logistic regression [87] (See PSSP website, http://pssp.srv.ualberta.ca/), cox proportional hazards regression, cox regression with elastic net penalty and random survival forest (using a R library, developed by Humza Heidar), we were not able to reach a significant concordance index or D-calibration measure on this task. One issue that might be affecting the results is that in addition to being right-censored, our dataset is also left-censored because we only have the patients' records at certain follow-up time points and the exact time that they left their medication is not available. Since most of these tools only deal with right-censored data, the interval-censored nature of our data might probably be one of the factors affecting the results of this survival regression task.

Figure 3.9: Prognosis curves, based on CGI severity. Each of the curves show the progression of CGI severity from the baseline time point to the existing follow-up time points (based on availability)

Chapter 4

Conclusions

Medical imaging data, particularly magnetic resonace imaging, has shown high potential for building automated tools for predicting diagnosis and prognosis status in mental disorders. In our first study, we addressed the challenge of building a simple classification framework that is explainable to the clinical society, without losing accuracy. Simple models are easy to follow and explain and thus, more likely to be accepted as a reliable diagnostic tool by clinicians. Still, coming up with a measure to quantify the balance between accuracy and simplicity is an ongoing challenge, mostly as we lack an objective definition for simplicity or interpretability. Thus, reaching a consensus on ways to quantify these model attributes would be of great value. In addition, for the Alzheimer's prediction task, there's still room for building more accurate, yet still simple diagnostic tools.

Our second study attempts to build a multimodal classification framework that can combine the information from different data modalities to predict the response of schizophrenia patients to treatment. Although we investigated various tasks and methods in this study, we were not able to reach any significant results. This may be partly attributed to the small number of patients, while dealing with high-dimensional features. While the small size of hospital datasets does make it challenging to produce

predictive models of the treatment outcome, we anticipate that gathering large logitudinal datasets with detailed information on the patients' medication log and follow-up severity measures would significantly contribute to the growth of research in this area. Another issue that might be affecting our results in the second study is the subjective nature of our response criteria, especially CGI scores. Because of this issue, there may be discrepancies between the scores recorded by different clinicians, and even when using the scores from a single clinician, we cannot be sure about the exact mapping between these rankings and numerical values. Therefore, one of the main steps for future work would be to analyze these subjective measures and find a more numerically meaningful mapping for them.

Bibliography

[1] J. Vizcarra, Diffusion tensor imaging (dti) with an mri gives insight on how concussions differ from person to person (2012).
URL www.quantumday.com/2012/06/diffusion-tensor-imaging-dti-with-mri.html

[2] K. H. Karlsgodt, D. Sun, A. M. Jimenez, E. S. Lutkenhoff, R. Willhite, T. G. Van Erp, T. D. Cannon, Developmental disruptions in neural connectivity in the pathophysiology of schizophrenia, Development and psychopathology 20 (4) (2008) 1297–1327.

[3] L. J. O'Donnell, C.-F. Westin, An introduction to diffusion tensor image analysis, Neurosurgery Clinics 22 (2) (2011) 185–196.

[4] G. Dougherty, Image analysis in medical imaging: recent advances in selected examples, Biomedical imaging and intervention journal 6 (3) (2010) e32.

[5] F. G. Ashby, An introduction to fmri, in: An introduction to model-based cognitive neuroscience, Springer, 2015, pp. 91–112.

[6] R. Iniesta, D. Stahl, P. McGuffin, Machine learning, statistical learning and the future of biological research in psychiatry, Psychological medicine 46 (12) (2016) 2455–2465.

[7] D. Bzdok, A. Meyer-Lindenberg, Machine learning for precision psychiatry, arXiv preprint arXiv:1705.10553.

[8] J. S. Shimony, R. C. McKinstry, E. Akbudak, J. A. Aronovitz, A. Z. Snyder, N. F. Lori, T. S. Cull, T. E. Conturo, Quantitative diffusion-tensor anisotropy brain mr imaging: normative human data and anatomic analysis, Radiology 212 (3) (1999) 770–784.

[9] J. Busner, S. D. Targum, The clinical global impressions scale: applying a research tool in clinical practice, Psychiatry (Edgmont) 4 (7) (2007) 28.

[10] S. Kumari, M. Malik, C. Florival, P. Manalai, S. Sonje, An assessment of five (panss, saps, sans, nsa-16, cgi-sch) commonly used symptoms rating scales in schizophrenia and comparison to newer scales (cains, bnss), Journal of addiction research & therapy 8 (3).

[11] Alzheimer's Association, 2017 alzheimer's disease facts and figures, Alzheimer's & Dementia 13 (4) (2017) 325–373. doi:10.1016/j.jalz.2017.02.001.

[12] I. Beheshti, H. Demirel, A. D. N. Initiative, et al., Probability distribution function-based classification of structural mri for the detection of alzheimer's disease, Computers in biology and medicine 64 (2015) 208–216.

[13] I. Beheshti, H. Demirel, F. Farokhian, C. Yang, H. Matsuda, A. D. N. Initiative, et al., Structural mri-based detection of alzheimer's disease using feature ranking and classification error, Computer methods and programs in biomedicine 137 (2016) 177–193.

[14] I. Beheshti, H. Demirel, A. D. N. Initiative, et al., Feature-ranking-based alzheimer's disease classification from structural mri, Magnetic resonance imaging 34 (3) (2016) 252–263.

[15] I. Beheshti, H. Demirel, H. Matsuda, A. D. N. Initiative, et al., Classification of alzheimer's disease and prediction of mild cognitive

impairment-to-alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm, Computers in Biology and Medicine 83 (2017) 109–119.

[16] Y. Ding, C. Zhang, T. Lan, Z. Qin, X. Zhang, W. Wang, Classification of alzheimer's disease based on the combination of morphometric feature and texture feature, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE, 2015, pp. 409–412.

[17] C. Hinrichs, V. Singh, G. Xu, S. C. Johnson, A. D. N. Initiative, et al., Predictive markers for ad in a multi-modality framework: an analysis of mci progression in the adni population, Neuroimage 55 (2) (2011) 574–589.

[18] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, A. D. N. Initiative, et al., Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks, NeuroImage: Clinical (2018) 101645.

[19] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.

[20] F. de Vos, T. M. Schouten, A. Hafkemeijer, E. G. Dopper, J. C. van Swieten, M. de Rooij, J. van der Grond, S. A. Rombouts, Combining multiple anatomical mri measures improves alzheimer's disease classification, Human brain mapping 37 (5) (2016) 1920–1929.

[21] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative, et al., Multimodal classification of alzheimer's disease and mild cognitive impairment, Neuroimage 55 (3) (2011) 856–867.

[22] C. Jongkreangkrai, Y. Vichianin, C. Tocharoenchai, H. Arimura, A. D. N. Initiative, et al., Computer-aided classification of alzheimer's disease based on support vector machine with combination of cerebral image features in mri, in: Journal of Physics: Conference Series, Vol. 694, IOP Publishing, 2016, p. 012036.

[23] Y. Zhang, N. Schuff, M. Camacho, L. L. Chao, T. P. Fletcher, K. Yaffe, S. C. Woolley, C. Madison, H. J. Rosen, B. L. Miller, et al., Mri markers for mild cognitive impairment: comparisons between white matter integrity and gray matter volume measurements, PloS one 8 (6) (2013) e66367.

[24] M. W. Weiner, P. S. Aisen, C. R. Jack, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett, et al., The alzheimer's disease neuroimaging initiative: progress report and future plans, Alzheimer's & Dementia 6 (3) (2010) 202–211.

[25] D. M. Jones-Davis, N. Buckholtz, The impact of adni: What role do public-private partnerships have in pushing the boundaries of clinical and basic science research on alzheimer's disease?, Alzheimer's & dementia: the journal of the Alzheimer's Association 11 (7) (2015) 860.

[26] M. F. Folstein, S. E. Folstein, P. R. McHugh, Mini-mental state: A practical method for grading the cognitive state of patients for the clinician., Journal of Psychiatric Research.

[27] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, R. Martin, A new clinical scale for the staging of dementia., The British journal of psychiatry 140 (6) (1982) 566–572.

[28] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward, et al.,

The alzheimer's disease neuroimaging initiative (adni): Mri methods, Journal of magnetic resonance imaging 27 (4) (2008) 685–691.

[29] B. Fischl, Freesurfer, Neuroimage 62 (2) (2012) 774–781.

[30] L. De Jong, K. Van der Hiele, I. Veer, J. Houwing, R. Westendorp, E. Bollen, P. De Bruin, H. Middelkoop, M. Van Buchem, J. Van Der Grond, Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an mri study, Brain 131 (12) (2008) 3277–3285.

[31] T. St J, J. C. Pruessner, F. Faltraco, C. Born, M. Rocha-Unold, A. Evans, H.-J. Möller, H. Hampel, Comprehensive dissection of the medial temporal lobe in ad: measurement of hippocampus, amygdala, entorhinal, perirhinal and parahippocampal cortices using mri, Journal of neurology 253 (6) (2006) 794–800.

[32] P. M. Thompson, K. M. Hayashi, G. I. De Zubicaray, A. L. Janke, S. E. Rose, J. Semple, M. S. Hong, D. H. Herman, D. Gravano, D. M. Doddrell, et al., Mapping hippocampal and ventricular change in alzheimer disease, Neuroimage 22 (4) (2004) 1754–1766.

[33] H. I. Jacobs, M. P. Van Boxtel, J. Jolles, F. R. Verhey, H. B. Uylings, Parietal cortex matters in alzheimer's disease: an overview of structural, functional and metabolic findings, Neuroscience & Biobehavioral Reviews 36 (1) (2012) 297–309.

[34] M. Di Paola, F. Di Iulio, A. Cherubini, C. Blundo, A. Casini, G. Sancesario, D. Passafiume, C. Caltagirone, G. Spalletta, When, where, and how the corpus callosum changes in mci and ad a multimodal mri study, Neurology 74 (14) (2010) 1136–1142.

[35] E. Canu, G. B. Frisoni, F. Agosta, M. Pievani, M. Bonetti, M. Filippi, Early and late onset alzheimer's disease patients have distinct pat-

terns of white matter damage, Neurobiology of aging 33 (6) (2012) 1023–1033.

[36] E. Corder, A. Saunders, W. Strittmatter, D. Schmechel, P. Gaskell, G. a. Small, A. Roses, J. Haines, M. A. Pericak-Vance, Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families, Science 261 (5123) (1993) 921–923.

[37] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[38] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[39] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[40] T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural computation 10 (7) (1998) 1895–1923.

[41] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on pattern analysis and machine intelligence 27 (8) (2005) 1226–1238.

[42] Y.-W. Chen, C.-J. Lin, Combining svms with various feature selection strategies, in: Feature extraction, Springer, 2006, pp. 315–324.

[43] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (1-3) (2002) 389–422.

[44] C. R. Jack, R. C. Petersen, P. C. O'brien, E. G. Tangalos, Mr-based hippocampal volumetry in the diagnosis of alzheimer's disease, Neurology 42 (1) (1992) 183–183.

[45] S. Derflinger, C. Sorg, C. Gaser, N. Myers, M. Arsic, A. Kurz, C. Zimmer, A. Wohlschläger, M. Mühlau, Grey-matter atrophy in alzheimer's disease is asymmetric but not lateralized, Journal of Alzheimer's Disease 25 (2) (2011) 347–357.

[46] G. W. Van Hoesen, B. T. Hyman, A. R. Damasio, Entorhinal cortex pathology in alzheimer's disease, Hippocampus 1 (1) (1991) 1–8.

[47] T. Erkinjuntti, D. H. Lee, F. Gao, R. Steenhuis, M. Eliasziw, R. Fry, H. Merskey, V. C. Hachinski, Temporal lobe atrophy on magnetic resonance imaging in the diagnosis of early alzheimer's disease, Archives of Neurology 50 (3) (1993) 305–310.

[48] C. A. Tamminga, D. R. Medoff, The biology of schizophrenia, Dialogues in clinical neuroscience 2 (4) (2000) 339.

[49] L. Antonucci, G. Pergola, D. Dwyer, S. Torretta, M. T. Attrotto, R. Romano, B. Gelao, R. Masellis, A. Rampino, G. Caforio, et al., O9. 4. predicting schizophrenia: Identification of multimodal markers of disease through a machine learning approach, Schizophrenia Bulletin 44 (suppl_1) (2018) S100–S101.

[50] M. Shim, H.-J. Hwang, D.-W. Kim, S.-H. Lee, C.-H. Im, Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level eeg features, Schizophrenia research 176 (2-3) (2016) 314–319.

[51] P. Dazzan, C. Arango, W. Fleischacker, S. Galderisi, B. Glenthøj, S. Leucht, A. Meyer-Lindenberg, R. Kahn, D. Rujescu, I. Sommer, et al., Magnetic resonance imaging and the prediction of outcome in

first-episode schizophrenia: a review of current evidence and directions for future research, Schizophrenia bulletin 41 (3) (2015) 574–583.

[52] G. E. Doucet, D. A. Moser, M. J. Luber, E. Leibu, S. Frangou, Baseline brain structural and functional predictors of clinical outcome in the early course of schizophrenia, Molecular Psychiatry (2018) 1.

[53] D. K. Sarpal, M. Argyelan, D. G. Robinson, P. R. Szeszko, K. H. Karlsgodt, M. John, N. Weissman, J. A. Gallego, J. M. Kane, T. Lencz, et al., Baseline striatal functional connectivity as a predictor of response to antipsychotic drug treatment, American Journal of Psychiatry 173 (1) (2015) 69–77.

[54] S. J. Siegel, F. Irani, C. M. Brensinger, C. G. Kohler, W. B. Bilker, J. D. Ragland, S. J. Kanes, R. C. Gur, R. E. Gur, Prognostic variables at intake and long-term level of function in schizophrenia, American Journal of Psychiatry 163 (3) (2006) 433–441.

[55] D. Luck, L. Buchy, Y. Czechowska, M. Bodnar, G. B. Pike, J. S. Campbell, A. Achim, A. Malla, R. Joober, M. Lepage, Fronto-temporal disconnectivity and clinical short-term outcome in first episode psychosis: a dti-tractography study, Journal of psychiatric research 45 (3) (2011) 369–377.

[56] M. Gheiratmand, I. Rish, G. A. Cecchi, M. R. Brown, R. Greiner, P. I. Polosecki, P. Bashivan, A. J. Greenshaw, R. Ramasubbu, S. M. Dursun, Learning stable and predictive network-based patterns of schizophrenia and its clinical symptoms, NPJ schizophrenia 3 (1) (2017) 22.

[57] B. Cao, R. Y. Cho, D. Chen, M. Xiu, L. Wang, J. C. Soares, X. Y. Zhang, Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity, Molecular Psychiatry (2018) 1.

[58] C. C. Aggarwal, Data classification: algorithms and applications, CRC Press, 2014, pp. 498–500.

[59] M. J. Azur, E. A. Stuart, C. Frangakis, P. J. Leaf, Multiple imputation by chained equations: what is it and how does it work?, International journal of methods in psychiatric research 20 (1) (2011) 40–49.

[60] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[61] C. Gaser, F. Kurth, Manual computational anatomy toolbox-cat12, Structural Brain Mapping Group at the Departments of Psychiatry and Neurology, University of Jena.

[62] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain, Neuroimage 15 (1) (2002) 273–289.

[63] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, B. Thirion, Multi-subject dictionary learning to segment an atlas of brain spontaneous activity, in: Biennial International Conference on Information Processing in Medical Imaging, Springer, 2011, pp. 562–573.

[64] C. Destrieux, B. Fischl, A. Dale, E. Halgren, Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature, Neuroimage 53 (1) (2010) 1–15.

[65] P. Bellec, P. Rosa-Neto, O. C. Lyttelton, H. Benali, A. C. Evans, Multi-level bootstrap analysis of stable clusters in resting-state fmri, Neuroimage 51 (3) (2010) 1126–1139.

[66] O. Voevodskaya, A. Simmons, R. Nordenskjöld, J. Kullberg, H. Ahlström, L. Lind, L.-O. Wahlund, E.-M. Larsson, E. Westman,

A. D. N. Initiative, et al., The effects of intracranial volume adjustment approaches on multiple regional mri volumes in healthy aging and alzheimer's disease, Frontiers in aging neuroscience 6 (2014) 264.

[67] S. Huettel, A. Song, G. McCarthy, Signal, noise, and preprocessing of fmri data, Functional Magnetic Resonance Imaging 2.

[68] C. Yan, Y. Zang, Dparsf: a matlab toolbox for" pipeline" data analysis of resting-state fmri, Frontiers in systems neuroscience 4 (2010) 13.

[69] S. M. Smith, P. T. Fox, K. L. Miller, D. C. Glahn, P. M. Fox, C. E. Mackay, N. Filippini, K. E. Watkins, R. Toro, A. R. Laird, et al., Correspondence of the brain's functional architecture during activation and rest, Proceedings of the National Academy of Sciences 106 (31) (2009) 13040–13045.

[70] R. C. Craddock, G. A. James, P. E. Holtzheimer III, X. P. Hu, H. S. Mayberg, A whole brain fmri atlas generated via spatially constrained spectral clustering, Human brain mapping 33 (8) (2012) 1914–1928.

[71] J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al., Functional network organization of the human brain, Neuron 72 (4) (2011) 665–678.

[72] N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, et al., Prediction of individual brain maturity using fmri, Science 329 (5997) (2010) 1358–1361.

[73] J. Ashburner, K. J. Friston, Unified segmentation, Neuroimage 26 (3) (2005) 839–851.

[74] K. J. Friston, S. Williams, R. Howard, R. S. Frackowiak, R. Turner, Movement-related effects in fmri time-series, Magnetic resonance in medicine 35 (3) (1996) 346–355.

[75] J. Ashburner, A fast diffeomorphic image registration algorithm, Neuroimage 38 (1) (2007) 95–113.

[76] Y. Zang, T. Jiang, Y. Lu, Y. He, L. Tian, Regional homogeneity approach to fmri data analysis, Neuroimage 22 (1) (2004) 394–400.

[77] Q.-H. Zou, C.-Z. Zhu, Y. Yang, X.-N. Zuo, X.-Y. Long, Q.-J. Cao, Y.-F. Wang, Y.-F. Zang, An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff, Journal of neuroscience methods 172 (1) (2008) 137–141.

[78] J. Kim, J. R. Wozniak, B. A. Mueller, X. Shen, W. Pan, Comparison of statistical tests for group differences in brain functional networks, NeuroImage 101 (2014) 681–694.

[79] J. Soares, P. Marques, V. Alves, N. Sousa, A hitchhiker's guide to diffusion tensor imaging, Frontiers in neuroscience 7 (2013) 31.

[80] S. M. Smith, Fast robust automated brain extraction, Human brain mapping 17 (3) (2002) 143–155.

[81] S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews, et al., Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data, Neuroimage 31 (4) (2006) 1487–1505.

[82] M. E. Tipping, C. M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (3) (1999) 611–622.

[83] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[84] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Physical review E 69 (6) (2004) 066138.

[85] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (Feb) (2012) 281–305.

[86] R. Caruana, Multitask learning, Machine learning 28 (1) (1997) 41–75.

[87] C.-N. Yu, R. Greiner, H.-C. Lin, V. Baracos, Learning patient-specific cancer survival distributions as a sequence of dependent regressors, in: Advances in Neural Information Processing Systems, 2011, pp. 1845–1853.