Predicting Treatment Outcomes for Schizophrenia Patients Using Machine Learning

May 2022

Kate Čevora, Diego Fajardo Rojas, Weiwei Qian, Xinyu Zhang [1]

Supervised by Dr. Emma Robinson, Dr. Lucy Vanes and Dr. Logan Williams

Abstract

Schizophrenia is a severe and heterogeneous brain disorder, with a significant proportion of patients demonstrating resistance to treatment with anti-psychotic medication. Understanding structural and functional features of the brain that differ between treatment responsive and treatment resistant patients can improve treatment outcomes for resistant patients and inform the future development of anti-psychotics. In this project we extract measures of cortical thickness, sulcal depth, curvature and surface area over Desikan-Killiany anatomical regions from T1w MRI brain images of 54 schizophrenia patients and 43 healthy controls. These features are then used to train machine learning models (Random Forest and Fully-connected Neural Network) to classify the subjects. We address two classification tasks: classifying patients from healthy controls [HCs] and classifying treatment resistant [TRS] and non-treatment resistant [NTR] patients from each other. We find that our best-performing model can distinguish patients and HCs with 73% accuracy on an unseen test dataset, but is unable to distinguish TRS and NTR patients above random performance. We then use feature importance rankings from the Random Forest model to extract and visualise the anatomical regions which were most important when classifying patients and HCs.

[1] Authors listed in alphabetical order.

Contents

1 Introduction

Schizophrenia is a severe brain disorder which is highly heterogeneous in nature. All current licensed anti-psychotics used to treat schizophrenia symptoms (such as clozapine and risperidone) act on dopamine D2 receptors [1], but between 14% and 60% [2, 3, 4] of patients are resistant to such treatment, despite adequate D2 receptor occupancy [5, 6]. Treatment-resistant patients are often highly symptomatic and may require extensive periods of hospital care [7]. Understanding the factors which contribute to treatment resistance is key to tailoring the next generation of anti-psychotic medication to provide better outcomes for these patients.

There is a body of research suggesting differences in both brain structure and function of patients with schizophrenia [8, 9, 10, 11]. Further, more recent research suggests differences in brain structure and function also exist between treatment resistant [TRS] and non treatment resistant [NTR] schizophrenia patients [12, 13, 14] . If markers derived from structural and functional magnetic resonance imaging [MRI] could be used to predict which patients were likely to have a worse response to anti-psychotic medications within the first few weeks of treatment, interventional strategies targeted at specific subgroups could be implemented early [15], improving treatment outcomes.

In this paper we employ Machine Learning methods to classify TRS and NTR patients from features derived from structural MRI data using several different models. We also attempt to classify patients from healthy controls [HCs] to provide a baseline performance of our model. We deconfound age and sex from the data using two different methods and compare predictive performance to non-deconfounded data. Finally, we visualise the most important predictive features to aid model explainability.

2 Background

In this section we present a selection of current research related to the topic of classifying treatment resistant [TRS] and non-treatment resistant [NTR] schizophrenia patients from neuroimaging data using Machine Learning [ML]. We discuss evidence for the possibility of predicting TRS and NTR from both structural and functional neuroimaging data, present existing research on ML models for prediction of NTR and TRS schizophrenia, and discuss methods for mitigating the influence of confounding variables on neuroimaging data. We also review several methods for extracting and visualising important predicive features from ML models to aid model explainability.

2.1 Neuroimaging Features of TRS Schizophrenia

There is a body of research that provides evidence for structural and functional differences in the brains of schizophrenia patients compared to healthy controls, and also between NTR and TRS patients. A Machine Learning model has the

potential to learn multiple of these distinguishing features from MRI data, in order to differentiate these groups.

A number of studies that have evaluated brain structure at the time of illness onset as well as following treatment have reported an association between smaller gray matter volumes and poorer clinical outcomes [15]. Kasparek et al. [12] found smaller volumes of prefrontal areas to be predictive of poorer outcomes after one year for patients with first-episode psychosis. In a 16-year longitudinal study, Jääskeläinen et al. [13] found that higher density of left frontal grey matter was associated with not being on a disability pension, alongside a number of other correlations between structural features in the brain and course of illness. Barry et al. [16] studied differences in cortical volume, thickness and surface area for healthy controls, TRS and NTR patients. They found TRS individuals showed significantly greater reductions in cortical volume and thickness compared to NTR patients in the right frontal and precentral regions, right parietal and occipital cortex, left temporal cortex and bilateral cingulate cortex.

In terms of differences in brain function, there is evidence that reward feedback processing mechanisms could be impaired for patients with schizophrenia [8, 9, 17]. Vanes et al. [14] compared treatment resistant [TRS] and non-treatment resistant [NTR] patients by examining neural correlates of reward prediction error [RPE] signals (associated with reinforcement learning [18]) using fMRI data acquired while patients completed a probabilistic reinforcement learning task. Results suggested that while the behavioural output of TRS and NTR patients during the reinforcement learning task was similar, the underlying neural mechanisms may differ, supporting the hypothesis that TRS may represent a different disease from NTR schizophrenia.

2.2 Classifying Schizophrenia Patients with ML

The high complexity and dimensionality of neuroimaging data relevant to TRS and NTR schizophrenia classification lends itself to analysis via Machine Learning [ML]. As discussed in Section 2.1, there are many features which appear to be somewhat correlated with treatment outcome for schizophrenia patients, and ML offers the opportunity to combine many of these features together to build a more powerful predictor. Here we cover the methods and outcomes of a number of ML models applied to the task of classifying schizophrenia patients from neuroimaging data.

Most research to date has focused on the problem of classifying schizophrenia patients from healthy controls. Greenstein et al. [19] used 74 anatomic brain MRI sub-regions derived from structural MRI data to classify 98 childhood onset schizophrenia patients and 99 controls using Random Forests and achieved predictive accuracy of 73.7%. Iwabuchi et al. [20] used gray matter and white matter MRI images as inputs to a Support Vector classifier and achieved a highest accuracy of 77% (with a 7T scanner) in distinguishing 19 patients and 20 control subjects. Lu et al. [21] combined Support Vector Machines with recursive feature elimination to discriminate 41 schizophrenia patients from 40

controls using structural MRI data from a 3T scanner, and achieved 88.4% classification accuracy.

Ambrosen et al. [22] developed a framework for using multi-modal imaging data to classify healthy-controls [HCs] and patients, and also non-treatment resistant [NTR] and treatment resistant [TRS] patients using Machine Learning methods. The multi-modal data included regional measures of cortical thickness, surface area and mean curvature derived from T1-weighted MRI images, data measuring cognitive performance on standardized tests, and electrophysiology data. They combined this data with additional simulated data to train an ensemble of basic Machine Learing models including linear regression algorithms, Support Vector Machines and Random Forests. Ambrosen et al. were able to classify HCs from TRSs and NTRs with a maximum class-balanced accuracy of 64% using an ensemble of trees, but were unable to classify TRSs and NTRs with significance.

2.3 Dealing with Confounds in Medical Imaging Data

One issue faced by all predictive models is that of confounding variables in the training data. For example, if we consider a model built to predict whether a subject is diagnosed with schizophrenia or not based on structural brain features, there are a number of factors that are known to correlate with schizophrenia, such as gender [23] and substance abuse [24]. Both of these variables are known to affect gray matter structure [25], and can therefore be considered confounds of the target variable. A model that appears to be able to differentiate between schizophrenia patients and controls from structural features may actually be driven by features relating to gender or substance abuse (the confounding variables). This becomes an issue when attempting to explain what features drive the model's performance, as one could wrongly conclude that features relating for example to gender, are also features of interest in understanding the schizophrenic brain.

There are number of proposed methods for removing the influence of confounds from imaging data. One simple way to control for confounds is to make sure that the confounds are balanced between experimental groups [26]. In the case of gender being the confounding variable, this would mean including an even number of men and women in all experimental groups. In practise however, it is not always possible to counterbalance confounds. If the confounding variable is brain size, this can only be determined after data collection. Counterbalancing data that has already been collected effectively means throwing data away, and this can have a large impact on predictive model performance when data sets are typically small due to a high cost of data collection, as is the case for neuroimaging data.

Confound regression is a popular method for removing the influence of confounding variables from data, and has been used recently in the context of neuroimaging data [27, 28, 29, 30, 31]. Typically, a linear regression model is fitted to each feature in the dataset (e.g. each image voxel) with the confound (e.g. age, sex) as a predictor [25]. The variance explained by the confound is

then subtracted from the data to remove its influence. Snoek et al. [25] introduce a modified version of confound regression, called *cross-validated confound regression [CVCR]*. Instead of fitting the confounds to the features of the entire dataset (train and test set combined), they are fitted to a number of *folds* of the training data and the learned parameters of the regressor are then used to deconfound both the training and test data. When deconfounding brain size from structural MRI data for the prediction of gender, they found that CVRC retained predictive accuracy significantly better than standard whole-dataset confound regression.

2.4 Model Explainability

For some Machine Learning methods, the nature of the process by which the model makes predictions is relatively transparent, making it easy to understand the relative importance of input features to the model prediction. For example, feature importances can be easily extracted from a Random Forest classifier by measuring the average decrease in *impurity* (or mixing of categories) when a dataset is optimally split on each feature. The features that result in the greatest decrease in impurity are the largest contributors to model predictive accuracy.

In contrast, neural networks are considered to be black-box models due to the complex and often non-linear mapping between input features and output predictions. A number of methods exist to extract relative feature importances from deep neural network models.

The Shapley value approach compares the output of a model for one input datapoint and a comparison group of datapoints, and attributes how much difference is accounted for by each feature. For example, if our datapoint of interest is a female with a schizophrenia diagnosis that our model predicts to be treatment resistant, and our comparison group is a number of females with schizophrenia diagnosis that our model predicts to be non-treatment resistant, we wish to find which model input features contribute most to this difference in prediction. This is achieved by effectively perturbing the values of input features and measuring the effect on model output, for example by setting a given feature to the average value across the training dataset. Ibrahim et al. [32] use this approach to generate feature importances from a model trained to predict acute myocardial infarction from ECG values.

Saliency mapping is a set of methods for visualising which parts of an image are most important to a convolutional neural network's (CNN) decision making process. For a classification network, this can be done by computing the gradient of the score for a class with respect to the image pixels, in effect determining which pixels need to be changed the smallest amount to change the classification [33]. Current methods include GRAD-CAM [34] ans GRAD-CAM++ [35]. Saliency map methods have been applied to a number of models in the medical image analysis domain in an attempt to localize areas important to the decision-making process [36, 37, 38, 39]. However, there is evidence that these methods should be used with caution when attempting to localize features in high-risk

domains [40].

3 Methods

In this section we present our methods for deriving structural and functional features from MRI and fMRI data, and for building models to predict treatment outcomes of patients based on these features. We also include information on the dataset used for this project.

3.1 Dataset

54 schizophrenia patients with first episode psychosis (FEP) (according to ICD-10 criteria) and 43 healthy controls [HCs] with matched sex, age and socioeconomic background were recruited to participate in this project. Patients were classified into treatment resistant schizophrenia [TRS] (n=18) and treatment response [NTR] (n=35) groups. TRS group was defined as persistent psychotic symptoms with at least 4 scores on at least 2 positive symptoms of the Positive and Negative Syndrome Scale (PANSS) over 5 years [41] while NTR group got 3 or less scores on all items in PANSS [42] with stable symptoms for at least 6 months [43]. All patients in the TRS group had at least twice drug trials for 4-6 weeks without clinical improvement. The dataset comprises one T1w MRI image, one T2w MRI image and several task fMRI sequences for each candidate.

These data were provided by Dr. Vanes, who acquired it as part of her previous work on the subject . It can be verified [14, 44] that all participants provided written and informed consent to take part on the studies, and that ethical approval was provided by the London Camberwell St Giles Research and Ethics Committee.

3.2 Surface Features

The analysis on cortical surface maps is based on the evidence that they are influenced by genetic factors [45]. The sulcal depth is an important marker for brain morphological studies. It is obtained from the distance between the cortical surface and its cerebral hull [46]. Curvature provides information on how a point on the cortical surface is embedded in space [47]. The human cortex contains a highly folded sheet of neurons, where thickness varies from 1 to 4.5mm. Abnormal changing in cortical thickness may lead to neurodegenerative and psychiatric disorders [48]. Cortical thickness and surface area have been reported as relevant features in patients with schizophrenia. A previous study [49] compared cortical thickness and surface area between 4474 schizophrenia cohorts and 5098 healthy cohorts. The results report that individuals with schizophrenia show cortical thickness and surface abnormalities in frontal and temporal lobe regions compared with controls.

Although many studies focus on the volume of the brain, it can not reveal neurodevelopmental abnormalities in the cortex, such as cortical sulcal depth,

thickness, curvature, or surface area. Compared with a 3D volume, it is much easier to visualise cortical features on a surface. The variability in cortical organization within healthy controls, and between controls and patients with diseases can be better captured with a cortical surface. Mapping BOLD signals to the surface could also provide more accurate localisation of cortical signals and functional activities with less blurring of BOLD signals. In our study, we intend to analyse the cortical development of schizophrenia patients to find abnormal patterns in their cortex.

3.3 Image Processing

3.3.1 Functional Features

Freesurfer, Ciftify and Connectome Workbench were used to pre-process the functional MRI (fMRI) images. The Human Connectome Project (HCP) is an ambitious 5-year effort to characterise brain connectivity and function [50]. Their analytic approach is surface-based and uses CIFTI "grayordinate" file format for the cerebral cortex, which is more sensitive than volume-based approaches. It has made significant contributions in data acquisition and analysis in neuroimaging. Ciftify is a flexible framework to manage non HCP legacy data based on the concept of the preprocessing pipelines of HCP. It can analyze all functional datasets with anatomical data in CIFTI format. Easier surface-based analysis can therefore be produced with optimal surface-level results across different settings of acquisition [51].

The pre-processing steps were computationally intensive and took upwards of 15 hours per subject to run. The pipeline used for extracting features from the fMRI data is as follows:

1. Freesurfer `recon-all` command used to extract cortical surfaces from T1w images for all subjects. T2w images were included in the cortical reconstruction process to improve contrast.

2. Ciftify `ciftify_recon_all` command used to convert cortical surfaces to HCP style folder structure.

3. Cortical surfaces were re-sampled to create low-resolution 32k versions with separate left and right hemispheres.

4. Timeseries fMRI volumes for each task were registered to the T1w image.

5. Registered timeseries fMRI volumes were mapped to the 32k cortical surfaces using the Connectome Workbench `wb_command -volume-to-surface-mapping` command. Output was separate left and right hemispheres.

6. Left and right registered fMRI surface hemispheres were re-combined using the Connectome Workbench `wb_command -cifti-create-dense-timeseries` command.

It was originally planned to complete a group Independent Component Analysis (ICA) of the fMRI data mapped to the cortical surfaces using FSL's MELODIC Task ICA package. The extracted signals would form a complementary set of features to the structural features. Unfortunately during quality control of the last stage of processing it was discovered that the registration of the fMRI volumes to the low-resolution surfaces was too poor to continue with processing and this set of features was discarded from the project.

3.3.2 Structural Features

Image processing to derive structural features was achieved using Multimodal surface matching (MSM) to register the sulcal depth maps obtained in Step 2 of Section 3.3.1 to a template image.

MSM is a surface-based cortical registration method. The method maps the surfaces to a sphere. It drives alignment using various descriptors of univariate, multivariate or multimodal features. In more detail, metrics such as curvature, sulcal depth and myelin are univariate, resting state networks or task fMRI are multivariate, and the combinations of folding, myelin and fMRI are multimodal.

MSM matches input spherical surfaces and reference spherical surfaces. A low-resolution regular control point grid (Figure 1.A) was warped to perform image registration. Each control point is deformed independently in every iteration of rotations. Labels are defined by a set of regularly spaced points that are surrounded by the control point (Figure 1.B). The control point is then deformed to its optimal label point by using the rotations about the centre of the sphere (Figure 1.C) [52].

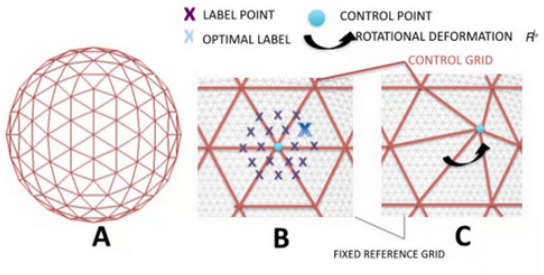

Figure 1: Theory of how MSM performs registration using a grid of control points.

Image processing was achieved using MSM to register the sulcal depth maps obtained from Step 2 in Section 3.3.1 to a template. The registration was run in left-to-right and right-to-left directions in each case with all configuration levels. The output contains three files and one of the output files is a file that was undergoing quality control (QC). For instance, the cortical surface template is shown in Figure 2a. Figure 2b is the QC result of one of the subjects.

Specifically, the white dots highlight all the folds present in the template sphere and manually compare whether the subject folds match the template folds.

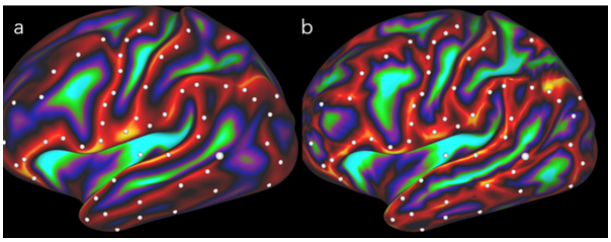

Figure 2: Template (left); Quality control of the output by using MSM of the left hemisphere of one subject (right).

A further resampling step was then performed. The `-metric-resample` function in the Connectome Workbench was applied to resample metrics to different meshes.

Asymmetry maps of each metric (thickness, sulcal depth and curvature) were obtained by subtracting the right hemisphere metric from the left hemisphere metric, then using Equation 1 to normalize it.

$$\frac{L - R}{0.5(L + R)} \tag{1}$$

Then all subjects per metric were merged and all midthickness surfaces were merged together to generate a design matrix for PALM statistical analysis. Furthermore, all metrics were merged per subject for ROI averaging analysis.

Once a merged file per subject was produced, a Python script, based on a previous script written by Dr. Robinson [2], was implemented to average each of the metrics within 125 ROIs per hemisphere (see Figure 3). In turn, these measurements were normalised and registered in a .csv file to be used as features by the Machine Learning models.

Figure 3: Template showing the 125 ROIs used for feature averaging per hemisphere.

[2] https://github.com/ecr05/Random_forests_for_cortical_imaging_data/blob/master/extract_roi_data.py

3.4 Data Deconfounding

Two different techniques to deconfound neuroimaging data were investigated, namely including confounds in the data and confound regression. In the present work, the use of the former is intrinsically related to the model in which it is implemented and it is explained in the next section.

Confound regression was briefly explained in Section 2.3. The mathematical formulation of this process is straightforward. If X is a feature matrix and C is a matrix with an intercept in the first column and confounds in the remaining ones, the Ordinary Least Squares regression has the solution

$$\beta_C = \left(C^T C\right)^{-1} C^T X.$$

Then the variance can be subtracted from the data matrix to yield a new deconfounded data matrix X_{corr}:

$$X_{corr} = X - \beta_C C.$$

Instead of performing this operation on the whole dataset, a cross-validated version of this algorithm is implemented, as advocated by [25]. Cross-validated confound regression was performed using the module of the open-source skbold Python package [53].

3.5 Models

As mentioned in Section 1, two classification problems were addressed. The first task was distinguishing between schizophrenia patients and healthy controls [HCs], and the second one was classifying TRS and NTR patients. To achieve this, two fundamentally different approaches were implemented. The chosen classification algorithms were Random Forests and Fully-connected Neural Networks. Based on the feature importances given by the Random Forest models, feature selection was also explored for the input of the Fully-connected Neural Networks.

All the experiments were performed in Python (version 3.6). The Random Forests models, as well as the feature selection derived from these, were implemented via scikit-learn (version 1.0.2). PyTorch (version 1.10.0) was used for the implementation of the Neural Networks. An 80/20 split of the data was performed to obtain training and test sets. The same sets in each problem were used for every experiment.

The first evaluated model was a Random Forest which was fed the original, normalised data. A second Random Forest was fed the data obtained by applying confound regression to the original data. To find the best parameters of these models, a grid-search was performed using the GridSearchCV method of scikit-learn, with 5 default folds used for cross-validation. To address class imbalance, during these searches the data was stratified according to their labels. The parameters over which the searches were conducted, alongside their possible values, are shown in Table 1.

Parameter	Values
max_depth	3, 5, 10, 20, 50
max_features	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64, 68, 72, 76, 80, 84, 88, 92, 96, 100, 104, 108, 112, 116, 120, 124, 128, 132, 136, 140, 144, 148, 152, 156, 160, 164, 168, 172, 176, 180, 184, 188, 192, 196, 200
n_estimators	10, 25, 50, 100, 150, 200

Table 1: Parameters searched to obtain the best performing Random Forests.

Let RF_{orig} be the Random Forest that was fed the original data, defined by the parameters that resulted in the best performance. Analogously, let RF_{decon} be the best performing Random Forest that was fed the data produced by confound regression. The most important features obtained by each of these models were selected using the SelectFromModel method from scikit-learn, with the maximum number of features parameter set to 100, to keep the number of selected features similar to the size of the datasets. Table 2 shows the number of features selected from each of the models for each classification problem.

	$\boldsymbol{RF_{orig}}$	$\boldsymbol{RF_{decon}}$
HCs / Patients	52	100
TRS / NTR	100	42

Table 2: Number of features selected by SelectFromModel.

All the remaining models were Fully-connected Neural Networks with two hidden layers. The number of neurons in the first hidden layer was a tuneable hyperparameter. First, a Neural Network with the data previously deconfounded by confound regression was optimised and evaluated. This process was carried out twice, once with only the features selected by RF_{decon} and once with all the available features. In both networks the second hidden layer had 16 neurons.

Two further networks were trained to evaluate the performance of including confounds in the data fed to the network. To implement this deconfounding technique, the approach by Fawaz et al. in [54] was followed, thus the confounds were added in the second hidden layer of the networks. The data provided to these networks was the original, normalised data, once with only the features selected by RF_{orig} and once with all the available features. In these networks the second hidden layer had 18 neurons, including two neurons for the confounds.

The optimisation of every Neural Network was carried out with the Adam optimiser and a learning rate of 0.0001. Dropout with probability 0.4 was applied for regularisation after each layer. The number of training epochs, as well as the number of neurons in the first hidden layer of each network were tuned manually. Let NN_{decon} and $NN_{decon\_reduced}$ be the best performing networks

that were fed the data deconfounded by regression, with full and selected features respectively. Similarly, NN_{orig} and $NN_{orig\_reduced}$ are the best performing networks that were fed the original data and the confounds in the second hidden layer. Tables 3 and 4 show the selected hyperparameters for these networks in each classification problem.

HCs/Patients Problem		
Network	**Training Epochs**	**Neurons in the First Hidden Layer**
NN_{orig}	301	128
$NN_{orig\_reduced}$	601	64
NN_{decon}	221	128
$NN_{decon\_reduced}$	301	128

Table 3: Hyperparameters of the Neural Networks for the HCs/Patients problem.

TRS/NTR Problem		
Network	**Training Epochs**	**Neurons in the First Hidden Layer**
NN_{orig}	251	64
$NN_{orig\_reduced}$	601	64
NN_{decon}	151	64
$NN_{decon\_reduced}$	301	32

Table 4: Hyperparameters of the Neural Networks for the TRS/NTR problem.

Finally, feature importances were derived from the Random Forests models, which in turn were used for further explainability. This is presented in Section 4.

3.6 PALM Analysis

In order to determine whether there were any statistically significant differences between the distributions of structural features for healthy controls and patients, a Permutation Analysis of Linear Model (PALM) was performed. These results provide a comparison for the features importances derived from the Machine Learning models which classify patients from controls.

3.6.1 Background

PALM is an FSL tool used for inference with permutation methods, providing exact control of false positives via Matlab or Octave. Plenty of features that

could not be used in other software are available in PALM [55]. PALM is being used increasingly in neuroimaging analysis.

PALM analysis is based on the student t-test, which is the most common statistical test used to determine the significant difference between the mean of two groups, assuming the sample is normally distributed and with a null hypothesis that the difference is zero.

There are three types of t-tests:

1. The one sample t-test compares the average of one group against the set mean.

2. The independent two sample t-test is used to compare the mean between two groups.

3. The paired sample t-test is used to compare one sample over time.

For this PALM analysis, an independent two sample t-test was used to compare features between controls and patients. This is calculated as follows:

$$\frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{m_1} - \frac{s_2^2}{m_2}}} \tag{2}$$

Where m_i is the mean of group i, and s_i is the standard deviation of group i.

In statistics, two types of mis-classification may occur. Type I error (false positive) occurs when the null hypothesis is rejected but it is in fact true, while type II error (false negative) arises when the null hypothesis is not rejected, but in reality is false. On average, when 5% of all experiments have one or more false positive voxels, it is described as family-wise error rate (FWER).

True positives occur when the null hypothesis is rejected and is false in reality, and true negatives occur when the null hypothesis is not rejected and is true in reality. Sensitivity is the ratio between true positives, and true positives plus true false negatives.

Cluster-based thresholding is the most popular approach to increase sensitivity by using spatial neighborhood information. With spatial smoothing applied before, the raw statistical image is set with threshold and p-values for each voxel are calculated after identifying the clusters of contiguous supra-threshold voxels. However, the pre-smoothing extent is arbitrary and unstable caused by initial hard thresholding. A new method called Threshold-Free cluster enhancement (TFCE) is then introduced to minimize those problems. It is similar to optimal cluster-based thresholding but stable and non-arbitrary. It goes through all the voxels and each voxel returns a cluster enhanced value. The output at each voxel is measured of local cluster-like support. In the TFCE enhanced images, graduation of values shows the significance [56].

3.6.2 Method

Statistical analysis was performed using FSL PALM [57] with TFCE. Merged files combining controls and patients for each feature from MSM image registration were used as inputs: left hemisphere curvature, right hemisphere curvature, left hemisphere sulcus, right hemisphere sulcus, left hemisphere thickness, right hemisphere thickness, asymmetry curvature, asymmetry sulcus, and asymmetry thickness. TFCE was performed on the group average mid-thickness anatomical surface. Left and right hemisphere mid-thickness surface files were corresponding with input files and left hemisphere mid-thickness surface file was used for asymmetry files. Independent two sample t-tests were used to investigate differences in features between controls and patients (C1: control > patient and C2: patient > control). FWER correlation was applied across modalities and contrasts.

4 Results

In the following section results are presented for the classification accuracy of the trained ML models on two tasks: classifying healthy controls [HCs] from patients, and classifying treatment resistant [TRS] and treatment non-resistant [NTR] schizophrenia patients. The feature importances derived from these models, alongside the outcomes of the PALM statistical testing are also explained. Before proceeding to that, the success of data deconfounding is evaluated.

4.1 Data Deconfounding

Figure 4 shows the distribution of the age of subjects in the patient and control groups. The difference in these distributions poses a potential problem, as there are many changes in structural brain features associated with the ageing process [58, 59, 60, 61, 62]. The difference in age distribution between the patient and control groups means that a Machine Learning model trained to predict which group a patient belongs to may be able to do so based on structural features associated with ageing, rather than with schizophrenia.

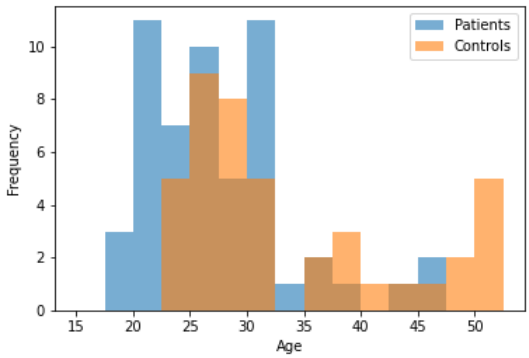

Figure 4: Distribution of age in the patients and control groups.

To determine the potential impact of age and sex as confounding variables on the model, we trained a classifier to predict these variables from the rest of the data. We then used confound regression as described in Section 3.4 to compensate for the effect of these variables on the structural features dataset, and trained the same model again to predict these features from the deconfounded structural data.

A Fully-connected Neural Network with two hidden layers was trained to predict sex, achieving an accuracy of 0.8 on test set when using the original data and an accuracy of 0.6 with the deconfounded data. Table 5 shows the scores achieved by the sex predicting network on the test set. Figure 5 displays the training curves with both datasets, providing further evidence of the success of confound regression. Various network architectures were tested to try to get similar insights in the case of age, to no avail. In both the cases of original and deconfounded data, a model that predicted age with significant precision was not found. For example, Table 6 displays the MSE achieved by training a Neural Network with the same architecture as the sex predicting network.

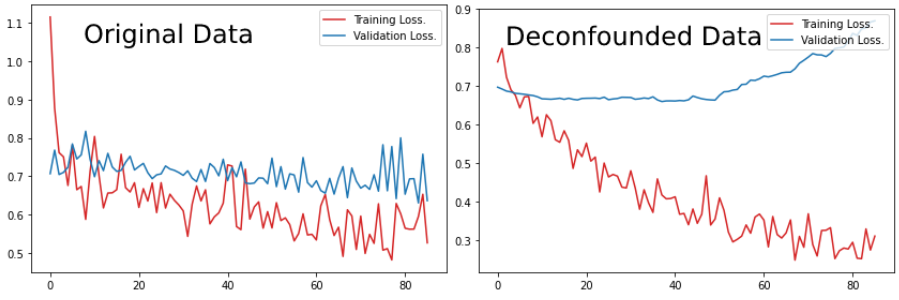

Figure 5: Training curves for the experiments of predicting sex.

Sex Predicting Network			
Data	**Accuracy**	**Sensitivity**	**Specificity**
Original, Training Set	0.95	0.93	1
Original, Test Set	0.8	0.8	0.8
Deconfounded, Training Set	1	1	1
Deconfounded, Test Set	0.6	1	0.2

Table 5: Evaluation metrics of the sex predicting network.

Age Predicting Network	
Data	**MSE ($Years^2$)**
Original, Training Set	189.21
Original, Test Set	295.05
Deconfounded, Training Set	164.24
Deconfounded, Test Set	329.67

Table 6: Evaluation metrics of the age predicting network.

4.2 Machine Learning Models

To measure the performance of the models, accuracy, sensitivity, and specificity measures are reported. In the case of the evaluation of Random Forests models in the training set, cross-validated versions of these metrics are presented. Tables 7 and 8 display the results for the HCs/Patients problem and Tables 9 and 10 show the results for the TRS/NTR problem. It is clear that in both cases the Neural Networks models outperform their Random Forest counterparts.

In the case of the HCs/Patients problem, it is hard to pick a best performing model. All the Neural Networks achieved 0.7 in test set accuracy, with sensitivities and specificities varying in performance. NN_{orig} and NN_{decon} show above average but not excellent sensitivity and specificity, while $NN_{orig\_reduced}$ and $NN_{decon\_reduced}$ have almost perfect score in one of these metrics with below average performance on the other.

The three best performing networks in the TRS/NTR problem, $NN_{orig\_reduced}$, NN_{decon}, and $NN_{decon\_reduced}$ achieve 0.73 test set accuracy and have identical sensitivity and specificity scores. It is worth noting that the Random Forest models completely fail to learn to differentiate between TRS and TNR subjects, which is evidenced by the perfect specificity and null sensitivity in the test set. The best performing Neural Networks are not significantly better than the Random Forests in this regard, registering perfect test set specificity but only 0.25 sensitivity.

Training Set, HCs/Patients Problem			
Model	**Accuracy**	**Sensitivity**	**Specificity**
$\boldsymbol{RF_{orig}}$	0.7	0.79	0.59
$\boldsymbol{RF_{decon}}$	0.71	0.79	0.62
$\boldsymbol{NN_{orig}}$	0.79	0.74	0.85
$\boldsymbol{NN_{orig\_reduced}}$	0.84	0.77	0.94
$\boldsymbol{NN_{decon}}$	1	1	1
$\boldsymbol{NN_{decon\_reduced}}$	0.96	0.95	0.97

Table 7: Evaluation metrics of the models in the HCs/Patients problem, training set.

Test Set, HCs/Patients Problem			
Model	**Accuracy**	**Sensitivity**	**Specificity**
$\boldsymbol{RF_{orig}}$	0.6	0.82	0.33
$\boldsymbol{RF_{decon}}$	0.6	0.82	0.33
$\boldsymbol{NN_{orig}}$	0.7	0.64	0.78
$\boldsymbol{NN_{orig\_reduced}}$	0.7	0.45	1.0
$\boldsymbol{NN_{decon}}$	0.7	0.73	0.67
$\boldsymbol{NN_{decon\_reduced}}$	0.7	0.91	0.44

Table 8: Evaluation metrics of the models in the HCs/Patients problem, test set.

Training Set, TRS/NTR Problem			
Model	**Accuracy**	**Sensitivity**	**Specificity**
$\boldsymbol{RF_{orig}}$	0.71	0.14	1.0
$\boldsymbol{RF_{decon}}$	0.71	0.36	0.89
$\boldsymbol{NN_{orig}}$	0.71	0.14	1.0
$\boldsymbol{NN_{orig\_reduced}}$	0.69	0.14	0.96
$\boldsymbol{NN_{decon}}$	1	1	1
$\boldsymbol{NN_{decon\_reduced}}$	0.76	0.71	0.79

Table 9: Evaluation metrics of the models in the TRS/NTR problem, training set.

Test Set, TRS/NTR Problem			
Model	**Accuracy**	**Sensitivity**	**Specificity**
$\boldsymbol{RF_{orig}}$	0.64	0.0	1.0
$\boldsymbol{RF_{decon}}$	0.64	0.0	1.0
$\boldsymbol{NN_{orig}}$	0.64	0.25	0.86
$\boldsymbol{NN_{orig\_reduced}}$	0.73	0.25	1.0
$\boldsymbol{NN_{decon}}$	0.73	0.25	1.0
$\boldsymbol{NN_{decon\_reduced}}$	0.73	0.25	1.0

Table 10: Evaluation metrics of the models in the TRS/NTR problem, test set.

The Random Forests implementation in sklearn allows to extract the importance of each feature fed to the models. Tables 15, 16, 17, and 18 in the Appendix summarise the most important features of each model in each one of the classification problems, showing up to 60 features. In turn, the location and type of the most important features for each model, up to 60, are pictured in Figures 6, 7, 8, and 9. For better understanding of these results, Tables 11, 12, 13, and 14 show the brain regions where most important features of each model are located, ranked by number of features per region.

RF_{orig}, HC/Patients Problem		
Anatomical Region	**Count Number**	**Percentage in First 52 Features**
superior parietal	9	17.3%
superior temporal	6	11.5%
insula	5	9.6%
lateral occipital	4	7.7%
precuneus	4	7.7%
middle temporal	3	5.8%
rostral middle frontal	3	5.8%
post central	3	5.8%
enthorinal	2	3.8%
caudal middle frontal	2	3.8%
inferior parietal	2	3.8%
inferior temporal	2	3.8%
medial orbitofrontal	2	3.8%
supramarginal	1	1.9%
precentral	1	1.9%
superior frontal	1	1.9%
lingual	1	1.9%
lateral orbitofrontal	1	1.9%

Table 11: Counts of location of most important features given by RF_{orig} in the HC/Patients problem.

RF_{decon}, HC/Patients Problem		
Anatomical Region	**Count Number**	**Percentage in First 60 Features**
precentral	8	13.3%
insula	8	13.3%
post central	4	6.7%
lateral orbitofrontal	4	6.7%
superior parietal	3	5%
lateral occipital	3	5%
rostral middle frontal	3	5%
precuneus	3	5%
fusiform	3	5%
paracentral	3	5%
superior temporal	2	3.3%
supra marginal	2	3.3%
pars opercularis	2	3.3%
cuneus	2	3.3%
caudal middle frontal	1	1.7%
inferior parietal	1	1.7%
superior frontal	1	1.7%
ishtmus cingulate	1	1.7%
rostral anterior cingulate	1	1.7%
temporal pole	1	1.7%
parahippocampal	1	1.7%
pericalcarine	1	1.7%
pars triangularis	1	1.7%
bankssts	1	1.7%

Table 12: Counts of location of most important features given by RF_{decon} in the HC/Patients problem.

RF_{orig}, TRS/NTR		
Anatomical Region	**Count Number**	**Percentage in First 60 Features**
supra marginal	5	8.3%
lateral occipital	5	8.3%
para central	5	8.3%
superior temporal	4	6.7%
precentral	4	6.7%
inferior parietal	4	6.7%
pars triangularis	4	6.7%
superior parietal	3	5%
rostral middle frontal	2	3.3%
inferior temporal	2	3.3%
medial orbitofrontal	2	3.3%
post central	2	3.3%
superior frontal	2	3.3%
insula	2	3.3%
isthmus cingulate	2	3.3%
lateral orbitofrontal	2	3.3%
para hippocampal	2	3.3%
pars orbitalis	2	3.3%
caudal middle frontal	1	1.7%
precuneus	1	1.7%
lingual	1	1.7%
pars opercularis	1	1.7%
temporal pole	1	1.7%
posterior cingulate	1	1.7%

Table 13: Counts of location of most important features given by RF_{orig} in the TRS/NTR problem.

RF_{decon}, TRS/NTR		
Anatomical Region	**Count Number**	**Percentage in First 42 Features**
supra marginal	4	9.5%
precuneus	4	9.5%
precentral	4	9.5%
post central	4	9.5%
inferior parietal	3	7.1%
insula	3	7.1%
para central	3	7.1%
superior parietal	2	4.8%
rostral middle frontal	2	4.8%
lateral occipital	2	4.8%
superior frontal	2	4.8%
pars triangularis	2	4.8%
posterior cingulate	2	4.8%
entorhinal	1	2.4%
caudal middle frontal	1	2.4%
lingual	1	2.4%
isthmus cingulate	1	2.4%
temporal pole	1	2.4%

Table 14: Counts of location of most important features given by RF_{decon} in the TRS/NTR problem.

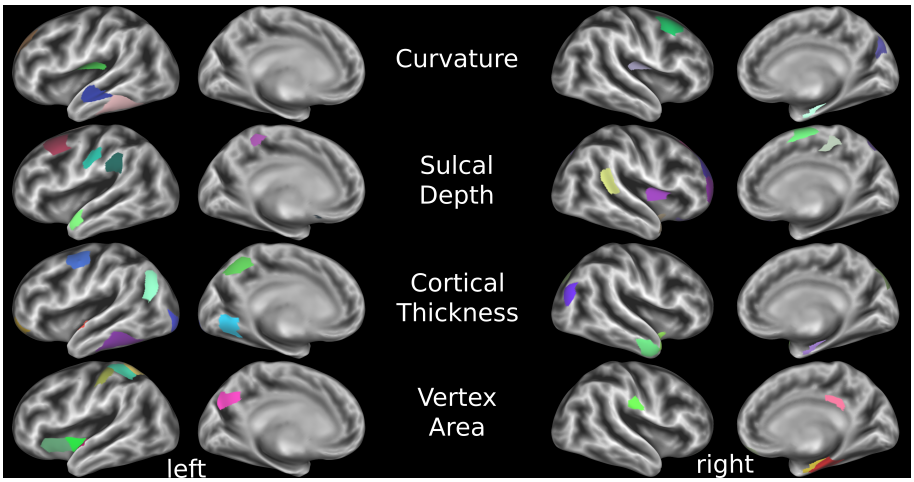

Figure 6: Most important features given by RF_{orig} in the HCs/Patients problem.

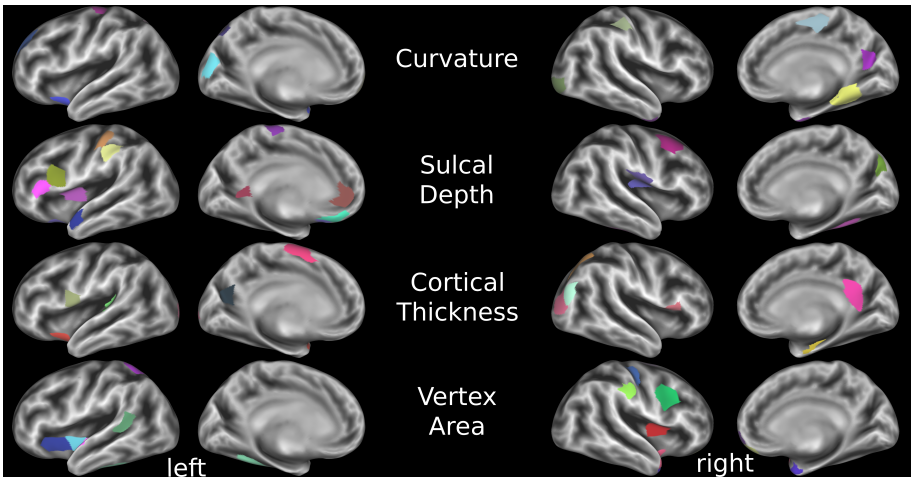

Figure 7: Most important features given by RF_{decon} in the HCs/Patients problem.

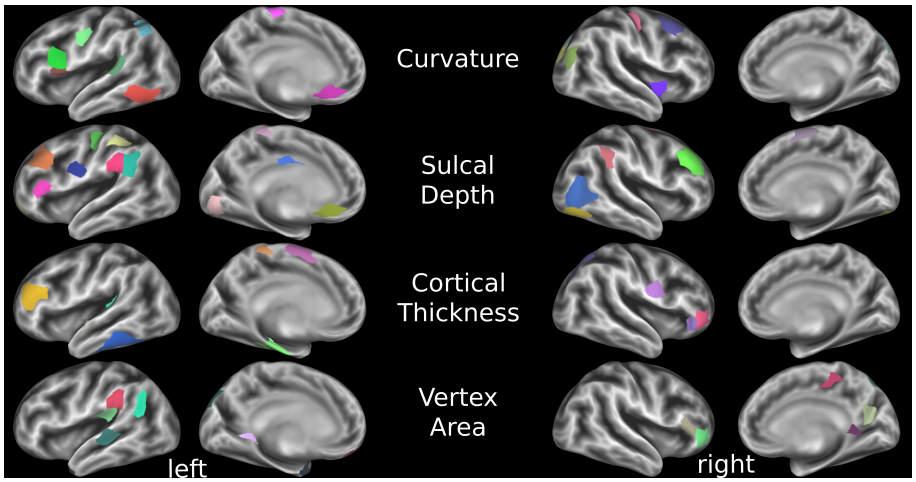

Figure 8: Most important features given by RF_{orig} in the TRS/NTR problem.

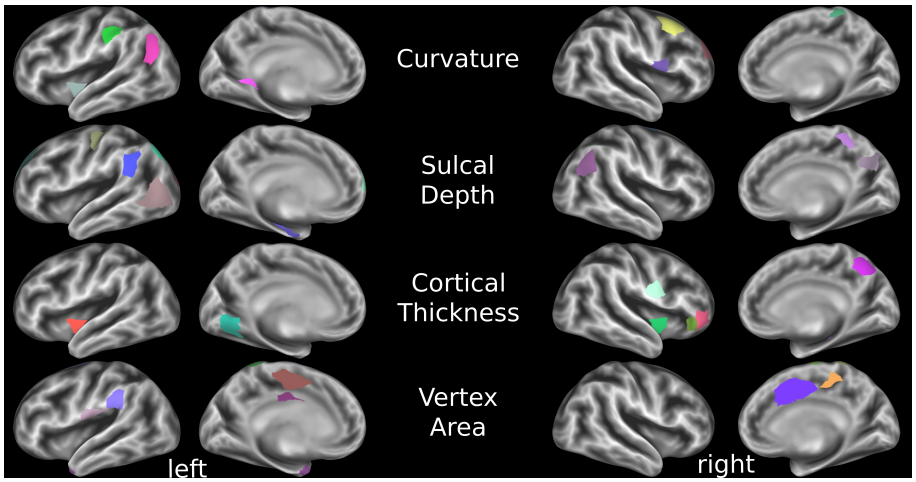

Figure 9: Most important features given by RF_{decon} in the TRS/NTR problem.

4.3 PALM Results

Output of statistical analysis on curvature, sulcus depth, and thickness of both hemispheres and asymmetry have been indicated in Figure 10. The colour scales demonstrate the logarithm of p-values. The yellow-red colour scales show the distribution of values across controls with higher mean than patients while blue-light blue colour scales show the distribution of values across patients with higher mean than controls. Apart from the thickness of the left hemisphere, no statistically significant result is found in the rest of metrics with significant level

lower than 0.05.

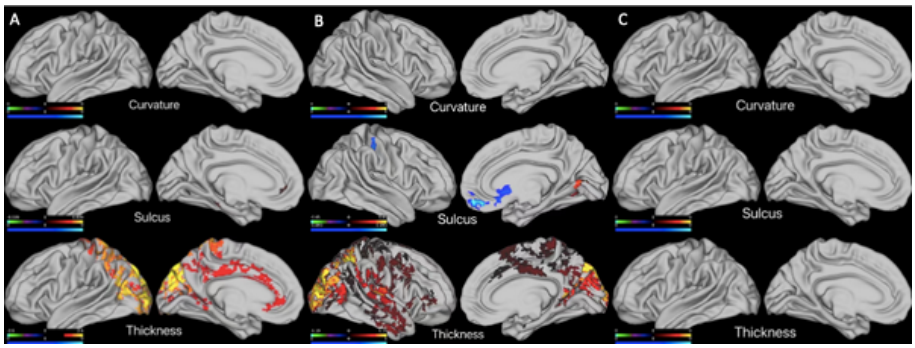

Figure 10: Statistical analysis of curvature, sulcus, and thickness for left hemisphere (A),right hemisphere (B) and asymmetry (C). Colour scales represent the logarithm of p-values.

5 Discussion

In this project we extracted features from structural and functional brain MRI images of patients with schizophrenia and healthy controls, and used these features to train Machine Learning models to classify the subjects. We tackled two classification tasks: classifying patients from healthy controls (HCs/Patients problem) and classifying treatment resistant [TRS] and non-treatment resistant [NTR] patients from eachother (TRS/NTR problem). Being able to successfully distinguish TRS and NTR patients at an early stage (i.e. before treatment with anti-psychotics starts) could allow more tailored and holistic approach to treatment for TRS patients that could help to improve clinical outcomes. Further, understanding what drived the predictions of such a model could highlight structural and functional features in the brains of TRS and NTR patients which drive these differences in response to anti-psychotics. This knowledge can help to inform the development of the next generation of anti-psychotic medication.

We implemented two types of Machine Learning model to classify patients based on structural brain features: Random Forest classifiers and Fully-connected Neural Network classifiers. For both classification problems we found that Neural Network models outperformed Random Forest models. For the HCs/Patients problem, the best predictive accuracy achieved with data that had been linearly deconfounded with respect to age and sex was 73% on a test set comprising 11 controls and 9 schizophrenia patients (resulting in a random accuracy of 55%). This result is similar to other studies focused on the same task of classifying schizophrenia patients and HCs from features derived from structural MRI brain images using Machine Learning models. For example, Greenstein et al. [19] achieved predictive accuracy of 73.7%, Iwabuchi et al. [20] achieved 77% and Lu et al. [21] achieved 88.4% classification accuracy in distinguishing

patients from HCs on a class-balanced dataset. Whilst classifying patients from HCs was not the primary objective of this project, the similarity in predictive accuracy achieved to other current literature provides evidence that our Machine Learning models were tuned and trained correctly. This is important in evaluating the outcomes of the second task (TRS/NTR problem), as we must be able to distinguish poor model performance from poor model implementation.

For the TRS/NTR patients classification problem, none of our models achieved significantly better performance than random accuracy. This could be due either to a poor model implementation, meaning that the relevant features for prediction were not exploited by the model, or because of a lack of signal in the feature set. Given that the same set of models were able to achieve reasonable predictive accuracy on the HCs/Patients problem, it is likely that the latter is true to some extent. At the time of writing this report, we were aware of only one other team that had addressed the TRS/NTR problem using features derived from structual MRI such as regional measures of cortical thikness, surface area, and mean curvature. Ambrosen et al. [22] were unable to extrapolate models that could successfully differentiate patients and healthy controls to the task of identifying TRS and NTR patients.

There is evidence that incorporation of functional data into the structural feature set that we used to predict treatment response may improve the classification ability of our Machine Learning models. Doucet et al. [63] acquired structural and functional resting-state MRI from patients in the early stages of schizophrenia diagnosis, and compared this data to clinical response after 24 weeks of anti-psychotic medication. They found no significant relationship between clinical response and measures of subcortical thickness and volume, however they did find a strong covariance between baseline functional connectivity in several brain networks and treatment outcome. Ambrosen et al. [22] also conclude that the incorporation of functional data may have significantly improved the classification ability of their models.

In order to derive a set of features for training the Machine Learning models to classify subjects, the cortical surface mesh was extracted from the T1w brain MRI image of each subject. This surface mesh was then used to generate a set of features based on average cortical thickness, sulcal depth and curvature across 125 regions of interest. These regions of interest were divided between 30 anatomical regions from the Desikan-Killiany atlas [64]. For both classification problems, we extracted the 60 most important features from the Random Forest classifier model and then ranked the brain regions in order of those with the highest number of features in these top 60. Given that the predictive performance for the TRS/NTR problem was not significantly above random, it does not make sense to analyse these regions as the model has not managed to pick up on signals relevant to the classification task in the data, so feature importances are likely random too. However, for the HCs/Patients problem it appears that the model was able to differentiate groups to some degree, so analysis of the features driving these predictions could provide interesting insight.

For the HCs/Patients classification problem with deconfounded data, the three brain anatomical regions that contained the most features of importance

were pre-central (13.3% of features), insula (13.3% of features) and postcentral (6.7% of features). The influence of these regions related to symptoms of schizophrenia corresponds to their usual functions. The precentral gyrus controls volitional movements of the contralateral side of the body [65]. The insula plays an important role in controlling emotion processing and arousal, including body awareness as well as decision making [66]. The postcentral is involved in the body's sensory inputs including touch, pain and temperature [67].

In 2018 the Enhancing Neuro Imaging Genetics through Meta Analysis [ENIGMA] group conducted a large meta-analysis of cortical thickness and surface area abnormalities in schizophrenia patients [49]. The study included 4474 patients with schizophrenia, and 5098 healthy volunteers with a similar mean age. T1w structural MRI brain images were processed using Freesurfer to extract cortical thickness and surface area for 70 DK atlas regions for each subject. The group differences for the DK atlas regions were compared using univariate statistical analysis. It was found that patients with schizophrenia had a significantly thinner cortex in all DK regions compared to the healthy controls, with the largest effect sizes in the bilateral fusiform, superior, middle and inferior temporal, left superior frontal gyri and bilateral insula regions. These results have some overlap with our findings based on feature importances, mainly with the insula region, but there are also some regions such as the superior, and middle gyri that have fairly low importance for our model. However, the fact that the ENIGMA meta-analysis identified significant differences in cortical thickness in all DK anatomical regions between groups suggests that our model may have utilised these differences in any of the regions.

6 Limitations and Future Work

In this section we will raise some of the limitations of our methodology, and make suggestions for improvements that could be carried out as future work.

It is important to note that feature importances derived from Random Forest models have some limitations, which may explain the differences between our results and that of the ENIGMA meta-analysis. The splitting criteria for decision tree models is to essentially pick the features and thresholds which result in the greatest decrease in mixing of categories at each node of the tree. As Random Forest models are ensembles of decision trees, feature importances can be derived from how often a particular feature is used to split the data as the more a feature is used the more important it is to the prediction. This can cause problems when two or more features are correlated with each other, as the feature which is chosen to split the data at a given node is essentially interchangeable with other features. As a result, feature importances derived from a Random Forest model can significantly vary between model training runs. Correlations between variables can also cause Random Forest models to underestimate the importance of variables [68]. In order to address this issue Verdinelli et al. introduce a method for calculating feature importances that is unaffected by correlations between variables. [69]. Chavent et al. also present a

method for calculating Random Forest feature importances that is not sensitive to correlations between variables [70].

One potential limitation of our use of confound regression to deconfound the data with respect to age and sex is that this method is only capable of compensating for linear relationships between the confounds and the features of the dataset. This means that there may be some non-linear relationships remaining in the deconfounded data, which could be exploited by our non-linear Machine Learning models for prediction. In future work it would be of interest to implement a non-linear deconfounding method such as an adversarial deconfounding autoencoder [71, 72] and observe whether this results in more robust deconfounding than confound regression.

As outlined in the Discussion, functional imaging data could hold the key to being able to differentiate between treatment resistant and treatment non-resistant patients. The study from which we derived the structural features to train our models also included fMRI data, but as mentioned in Section 3.3.1 we were unable to complete the processing of this data within the time frame of the project due to a failed registration step. An obvious next step for this project is to incorporate this data once processed into the feature set for training the models. This will hopefully improve the accuracy of the models on the TRS/NTR classification problem.

References

[1] Shitij Kapur and Gary Remington. Dopamine d2 receptors and their role in atypical antipsychotic action: still necessary and may even be sufficient. *Biological psychiatry*, 50(11):873–883, 2001.

[2] Katherine Beck, Robert McCutcheon, Lucy Stephenson, Marcela Schilderman, Natasha Patel, Rosalind Ramsay, and Oliver D Howes. Prevalence of treatment-resistant psychoses in the community: A naturalistic study. *Journal of Psychopharmacology*, 33(10):1248–1253, 2019.

[3] Oliver D Howes, Rob McCutcheon, Ofer Agid, Andrea De Bartolomeis, Nico JM Van Beveren, Michael L Birnbaum, Michael AP Bloomfield, Rodrigo A Bressan, Robert W Buchanan, William T Carpenter, et al. Treatment-resistant schizophrenia: treatment response and resistance in psychosis (trrip) working group consensus guidelines on diagnosis and terminology. *American Journal of Psychiatry*, 174(3):216–229, 2017.

[4] Jean-Pierre Lindenmayer. Treatment refractory schizophrenia. *Psychiatric Quarterly*, 71(4):373–384, 2000.

[5] Hubert J Coppens, Cees J Slooff, Anne MJ Paans, Tonnie Wiegman, Willem Vaalburg, and Jakob Korf. High central d2-dopamine receptor occupancy as assessed with positron emission tomography in medicated but therapy-resistant schizophrenic patients. *Biological psychiatry*, 29(7):629–634, 1991.

[6] Adam Wolkin, Faouzia Barouche, Alfred P Wolf, John Rotrosen, Joanna S Fowler, Chyng-Yann Shiue, Thomas B Cooper, and Jonathan D Brodie. Dopamine blockade and clinical response: evidence for two biological subgroups of schizophrenia. *Am J Psychiatry*, 146(7):905–908, 1989.

[7] Thomas H McGlashan. A selective review of recent north american long-term followup studies of schizophrenia. *Schizophrenia bulletin*, 14(4):515–542, 1988.

[8] Carol A Tamminga, Robert W Buchanan, and James M Gold. The role of negative symptoms and cognitive dysfunction in schizophrenia outcome. *International clinical psychopharmacology*, 1998.

[9] James A Waltz, Michael J Frank, Benjamin M Robinson, and James M Gold. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological psychiatry*, 62(7):756–764, 2007.

[10] Elena Antonova, Tonmoy Sharma, Robin Morris, and Veena Kumari. The relationship between brain structure and neurocognition in schizophrenia: a selective review. *Schizophrenia research*, 70(2-3):117–145, 2004.

[11] Stephen M Lawrie, Andrew M McIntosh, Jeremy Hall, David GC Owens, and Eve C Johnstone. Brain structure and function changes during the development of schizophrenia: the evidence from studies of subjects at increased genetic risk. *Schizophrenia bulletin*, 34(2):330–340, 2008.

[12] Tomas Kasparek, Radovan Prikryl, Daniel Schwarz, Hana Kucerova, Radek Marecek, Michal Mikl, Jiri Vanicek, and Eva Ceskova. Gray matter morphology and the level of functioning in one-year follow-up of first-episode schizophrenia patients. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 33(8):1438–1446, 2009.

[13] E Jääskeläinen, P Juola, J Kurtti, M Haapea, M Kyllönen, J Miettunen, P Tanskanen, GK Murray, S Huhtaniska, A Barnes, et al. Associations between brain morphology and outcome in schizophrenia in a general population sample. *European Psychiatry*, 29(7):456–462, 2014.

[14] Lucy D Vanes, Elias Mouchlianitis, Tracy Collier, Bruno B Averbeck, and Sukhi S Shergill. Differential neural reward mechanisms in treatment-responsive and treatment-resistant schizophrenia. *Psychological medicine*, 48(14):2418–2427, 2018.

[15] Paola Dazzan. Neuroimaging biomarkers to predict treatment response in schizophrenia: the end of 30 years of solitude? *Dialogues in Clinical Neuroscience*, 2022.

[16] Erica F Barry, Lucy D Vanes, Derek S Andrews, Krisna Patel, Charlotte M Horne, Elias Mouchlianitis, Peter J Hellyer, and Sukhi S Shergill. Mapping

cortical surface features in treatment resistant schizophrenia with in vivo structural mri. *Psychiatry research*, 274:335–344, 2019.

[17] Erin C Dowd, Michael J Frank, Anne Collins, James M Gold, and Deanna M Barch. Probabilistic reinforcement learning in patients with schizophrenia: relationships to anhedonia and avolition. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(5):460–473, 2016.

[18] Wolfram Schultz. Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27, 1998.

[19] Deanna Greenstein, James D Malley, Brian Weisinger, Liv Clasen, and Nitin Gogtay. Using multivariate machine learning methods and structural mri to classify childhood onset schizophrenia and healthy controls. *Frontiers in psychiatry*, 3:53, 2012.

[20] Sarina Iwabuchi, Peter F Liddle, and Lena Palaniyappan. Clinical utility of machine-learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. *Frontiers in psychiatry*, 4:95, 2013.

[21] Xiaobing Lu, Yongzhe Yang, Fengchun Wu, Minjian Gao, Yong Xu, Yue Zhang, Yongcheng Yao, Xin Du, Chengwei Li, Lei Wu, et al. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural mri images. *Medicine*, 95(30), 2016.

[22] Karen S Ambrosen, Martin W Skjerbæk, Jonathan Foldager, Martin C Axelsen, Nikolaj Bak, Lars Arvastson, Søren R Christensen, Louise B Johansen, Jayachandra M Raghava, Bob Oranje, et al. A machine-learning framework for robust and reliable prediction of short-and long-term treatment response in initially antipsychotic-naïve schizophrenia patients based on multimodal neuropsychiatric data. *Translational psychiatry*, 10(1):1–13, 2020.

[23] John McGrath, Sukanta Saha, David Chant, and Joy Welham. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic reviews*, 30(1):67–76, 2008.

[24] Lisa Dixon. Dual diagnosis of substance abuse in schizophrenia: prevalence and impact on outcomes. *Schizophrenia research*, 35:S93–S100, 1999.

[25] Lukas Snoek, Steven Miletić, and H Steven Scholte. How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage*, 184:741–760, 2019.

[26] Kai Görgen, Martin N Hebart, Carsten Allefeld, and John-Dylan Haynes. The same analysis approach: Practical protection against the pitfalls of novel neuroimaging analysis methods. *Neuroimage*, 180:19–30, 2018.

[27] Ahmed Abdulkadir, Olaf Ronneberger, Sarah J Tabrizi, and Stefan Klöppel. Reduction of confounding effects with voxel-wise gaussian process regression in structural mri. In *2014 International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4. IEEE, 2014.

[28] Juergen Dukart, Matthias L Schroeter, Karsten Mueller, and Alzheimer's Disease Neuroimaging Initiative. Age correction in dementia–matching to a healthy brain. *PloS one*, 6(7):e22193, 2011.

[29] Daniel Kostro, Ahmed Abdulkadir, Alexandra Durr, Raymund Roos, Blair R Leavitt, Hans Johnson, David Cash, Sarah J Tabrizi, Rachael I Scahill, Olaf Ronneberger, et al. Correction of inter-scanner and within-subject variance in structural mri based automated diagnosing. *NeuroImage*, 98:405–415, 2014.

[30] Anil Rao, Joao M Monteiro, Janaina Mourao-Miranda, Alzheimer's Disease Initiative, et al. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, 150:23–49, 2017.

[31] Michael T Todd, Leigh E Nystrom, and Jonathan D Cohen. Confounds in multivariate pattern analysis: theory and rule representation case study. *Neuroimage*, 77:157–165, 2013.

[32] Lujain Ibrahim, Munib Mesinovic, Kai-Wen Yang, and Mohamad A. Eid. Explainable prediction of acute myocardial infarction using machine learning and shapley values. *IEEE Access*, 8:210410–210417, 2020.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[35] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[36] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[37] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie

Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15(11):e1002699, 2018.

[38] Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S Corrado, Lily Peng, Dale R Webster, Naama Hammel, Yun Liu, and Avinash V Varadarajan. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1):18–27, 2020.

[39] N.J.S. Morch, U. Kjems, L.K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother, and K. Rehm. Visualization of neural networks using saliency maps. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 2085–2090 vol.4, 1995.

[40] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[41] Stanley R Kay, Abraham Fiszbein, and Lewis A Opler. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261–276, 1987.

[42] Robert R Conley and Deanna L Kelly. Management of treatment resistance in schizophrenia. *Biological psychiatry*, 50(11):898–911, 2001.

[43] Nancy C Andreasen, William T Carpenter Jr, John M Kane, Robert A Lasser, Stephen R Marder, and Daniel R Weinberger. Remission in schizophrenia: proposed criteria and rationale for consensus. *American Journal of Psychiatry*, 162(3):441–449, 2005.

[44] Lucy Vanes, Elias Mouchlianitis, Krisna Patel, Erica Barry, Katie Wong, Megan Thomas, Timea Szentgyorgyi, Dan Joyce, and Sukhi Shergill. Neural correlates of positive and negative symptoms through the illness course: an fmri study in early psychosis and chronic schizophrenia. *Scientific Reports*, 9, 10 2019.

[45] James Schmitt, Rhoshel Lenroot, Sarah Ordaz, Gregory Wallace, Jason Lerch, Alan Evans, Elizabeth Prom-Wormley, Kenneth Kendler, Michael Neale, and Jay Giedd. Variance decomposition of mri-based covariance maps using genetically-informative samples and structural equation modeling. *NeuroImage*, 47:56–64, 08 2008.

[46] Ilwoo Lyu, Hakmook Kang, Neil Woodward, and Bennett Landman. Sulcal depth-based cortical shape analysis in normal healthy control and schizophrenia groups. volume 10574, page 1, 03 2018.

[47] Jace B. King, Melissa Lopez-Larson, and Deborah A. Yurgelun-Todd. Mean cortical curvature reflects cytoarchitecture restructuring in mild traumatic brain injury. *NeuroImage : Clinical*, 11:81 – 89, 2016.

[48] Bruce Fischl and Anders Dale. Fischl b, dale am. measuring the thickness of the human cerebral cortex from magnetic resonance images. proc natl acad sci usa 97: 11050-11055. *Proceedings of the National Academy of Sciences of the United States of America*, 97:11050–5, 10 2000.

[49] Theo GM Van Erp, Esther Walton, Derrek P Hibar, Lianne Schmaal, Wenhao Jiang, David C Glahn, Godfrey D Pearlson, Nailin Yao, Masaki Fukunaga, Ryota Hashimoto, et al. Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (enigma) consortium. *Biological psychiatry*, 84(9):644–654, 2018.

[50] D.C. Essen, K Ugurbil, Edward Auerbach, Deanna Barch, T.E.J. Behrens, Richard Bucholz, A Chang, Liyong Chen, Maurizio Corbetta, Sandra Curtiss, Stefania Della Penna, David Feinberg, Matthew Glasser, Noam Harel, A.C. Heath, Linda Larson-Prior, Daniel Marcus, Georgios Michalareas, Steen Moeller, and Essa Yacoub. The human connectome project: A data acquisition perspective. *NeuroImage*, 62:2222–31, 02 2012.

[51] Erin Dickie, Alan Anticevic, Dawn Smith, Timothy Coalson, Mathuvanthi Manogaran, Navona Calarco, Joseph Viviano, Matthew Glasser, David Van Essen, and Aristotle Voineskos. Ciftify: A framework for surface-based analysis of legacy mr acquisitions. *NeuroImage*, 197, 05 2019.

[52] Emma C Robinson, Saad Jbabdi, Matthew F Glasser, Jesper Andersson, Gregory C Burgess, Michael P Harms, Stephen M Smith, David C Van Essen, and Mark Jenkinson. Msm: a new flexible framework for multimodal surface matching. *Neuroimage*, 100:414–426, 2014.

[53] Lukas Snoek. lukassnoek/skbold: First official release. *Zenodo*, Nov 2017.

[54] Abdulah Fawaz, Logan Z. J. Williams, Amir Alansary, Cher Bass, Karthik Gopinath, Mariana da Silva, Simon Dahan, Chris Adamson, Bonnie Alexander, Deanne Thompson, Gareth Ball, Christian Desrosiers, Hervé Lombaert, Daniel Rueckert, A. David Edwards, and Emma C. Robinson. Benchmarking geometric deep learning for cortical segmentation and neurodevelopmental phenotype prediction. *bioRxiv*, 2021.

[55] Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

[56] Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009.

[57] Anderson M Winkler, Gerard R Ridgway, Gwenaëlle Douaud, Thomas E Nichols, and Stephen M Smith. Faster permutation inference in brain imaging. *Neuroimage*, 141:502–516, 2016.

[58] Sophia Frangou, Amirhossein Modabbernia, Steven CR Williams, Efstathios Papachristou, Gaelle E Doucet, Ingrid Agartz, Moji Aghajani, Theophilus N Akudjedu, Anton Albajes-Eizagirre, Dag Alnæs, et al. Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3–90 years. *Human brain mapping*, 43(1):431–451, 2022.

[59] Carol E Franz, Hong Xian, Daphne Lew, Sean N Hatton, Olivia Puckett, Nathan Whitsel, Asad Beck, Anders M Dale, Bin Fang, Christine Fennema-Notestine, et al. Body mass trajectories and cortical thickness in middle-aged men: a 42-year longitudinal study starting in young adulthood. *Neurobiology of aging*, 79:11–21, 2019.

[60] DH Salat, DS Tuch, DN Greve, AJW Van Der Kouwe, ND Hevelone, AK Zaleta, BR Rosen, B Fischl, S Corkin, H Diana Rosas, et al. Age-related alterations in white matter microstructure measured by diffusion tensor imaging. *Neurobiology of aging*, 26(8):1215–1227, 2005.

[61] Naftali Raz, Paolo Ghisletta, Karen M Rodrigue, Kristen M Kennedy, and Ulman Lindenberger. Trajectories of brain aging in middle-aged and older adults: regional and individual differences. *Neuroimage*, 51(2):501–511, 2010.

[62] Andreas B Storsve, Anders M Fjell, Christian K Tamnes, Lars T Westlye, Knut Overbye, Hilde W Aasland, and Kristine B Walhovd. Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *Journal of Neuroscience*, 34(25):8488–8498, 2014.

[63] Gaelle E Doucet, Dominik A Moser, Maxwell J Luber, Evan Leibu, and Sophia Frangou. Baseline brain structural and functional predictors of clinical outcome in the early course of schizophrenia. *Molecular psychiatry*, 25(4):863–872, 2020.

[64] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[65] Department of Neurobiology and Anatomy at McGovern Medical School. Neuroanatomy online: An open-access electronic laboratory for the neurosciences. https://nba.uth.tmc.edu/neuroanatomy/. Accessed: 13-05-2022.

[66] Elsevier. Clinicalkey. https://www.clinicalkey.com/. Accessed: 13-05-2022.

[67] John Bigbee. Postcentral gyrus. In Jeffrey S. Kreutzer, John DeLuca, and Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 1972–1972. Springer New York, 2009.

[68] Sean Stijven, Wouter Minnebo, and Katya Vladislavleva. Separating the wheat from the chaff: on feature selection and feature importance in regression random forests and symbolic regression. In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*, pages 623–630, 2011.

[69] Isabella Verdinelli and Larry Wasserman. Decorrelated variable importance. *arXiv preprint arXiv:2111.10853*, 2021.

[70] Marie Chavent, Jerome Lacaille, Alex Mourer, and Madalina Olteanu. Handling correlations in random forests: which impacts on variable importance and model interpretability? In *ESANN*, 2021.

[71] Ayse B Dincer, Joseph D Janizek, and Su-In Lee. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Supplement_2):i573–i582, 2020.

[72] George Cevora. Fair adversarial networks. *arXiv preprint arXiv:2002.12144*, 2020.

7 Appendix

7.1 Feature Importances

RF_{orig}, HC/Patients Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
1	L_1_LABEL_39	0.040318	superior parietal
2	R_3_LABEL_237	0.037658	entorhinal
3	R_3_LABEL_241	0.037175	middle temporal
4	L_1_LABEL_77	0.036618	caudal middle frontal
5	L_3_LABEL_37	0.033148	superior temporal
6	R_2_LABEL_237	0.033034	superior temporal
7	R_0_LABEL_226	0.030422	caudal middle frontal
8	A_1_LABEL_75	0.030107	supra marginal
9	L_3_LABEL_74	0.028800	superior parietal
10	R_1_LABEL_148	0.028604	rostral middle frontal
11	L_2_LABEL_2	0.028405	lateral occipital
12	R_2_LABEL_244	0.027536	lateral occipital
13	L_1_LABEL_81	0.027089	superior temporal
14	A_1_LABEL_121	0.025806	precuneus
15	A_2_LABEL_68	0.025689	lateral orbitofrontal
16	R_2_LABEL_129	0.024148	insula
17	L_2_LABEL_86	0.023737	lingual
18	A_2_LABEL_53	0.022247	inferior parietal
19	R_0_LABEL_237	0.021887	precuneus
20	R_3_LABEL_128	0.020824	superior parietal
21	L_2_LABEL_37	0.020803	superior temporal
22	A_1_LABEL_65	0.019632	superior parietal
23	A_2_LABEL_1	0.019630	precentral
24	R_1_LABEL_236	0.019623	inferior parietal
25	A_0_LABEL_55	0.019318	inferior temporal
26	L_3_LABEL_16	0.018861	superior parietal
27	A_1_LABEL_100	0.017745	medial orbitofrontal
28	L_3_LABEL_60	0.017657	post central
29	R_0_LABEL_202	0.017533	entorhinal
30	A_2_LABEL_55	0.017050	inferior temporal
31	R_1_LABEL_191	0.016821	superior frontal
32	R_2_LABEL_148	0.016505	lateral occipital
33	R_3_LABEL_153	0.015929	middle temporal
34	L_1_LABEL_80	0.013109	postcentral
35	L_0_LABEL_99	0.013041	medial orbitofrontal
36	L_3_LABEL_108	0.012948	insula
37	R_1_LABEL_232	0.012607	middle temporal
38	L_0_LABEL_92	0.012533	post central
39	R_3_LABEL_235	0.012100	lateral occipital

RF_{orig}, HC/Patients Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
40	R_0_LABEL_217	0.011720	rostral middle frontal
41	A_0_LABEL_66	0.010921	rostral middle frontal
42	L_3_LABEL_17	0.010787	precuneus
43	R_1_LABEL_143	0.009791	superior parietal
44	L_2_LABEL_54	0.009746	precuneus
45	R_2_LABEL_168	0.009249	superior parietal
46	R_1_LABEL_165	0.008198	superior parietal
47	L_0_LABEL_50	0.006228	superior temporal
48	R_1_LABEL_210	0.006020	insula
49	R_1_LABEL_151	0.005938	insula
50	R_1_LABEL_213	0.005635	superior parietal
51	L_3_LABEL_58	0.004775	insula
52	R_1_LABEL_249	0.004299	superior temporal

Table 15: Most important features given by RF_{orig} in the HC/Patients problem.

RF_{decon}, HC/Patients Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
1	R_3_LABEL_148	0.033256	precental
2	R_2_LABEL_237	0.029226	lateral occipital
3	R_2_LABEL_188	0.019324	inferior parietal
4	A_1_LABEL_6	0.019110	isthmus cingulate
5	R_3_LABEL_172	0.014409	precental
6	A_1_LABEL_11	0.013618	lateral orbitofrontal
7	R_1_LABEL_217	0.013150	postcentral
8	R_1_LABEL_226	0.012456	caudal middle frontal
9	R_0_LABEL_169	0.012410	postcentral
10	L_1_LABEL_20	0.012067	rostral anterior cingulate
11	A_0_LABEL_98	0.011019	insula
12	L_2_LABEL_97	0.010805	lateral occipital
13	L_2_LABEL_89	0.010585	superior temporal
14	L_1_LABEL_26	0.010372	precental
15	R_2_LABEL_194	0.010325	pars opercularis
16	A_2_LABEL_93	0.009601	precentral
17	L_2_LABEL_98	0.009428	precentral
18	L_3_LABEL_122	0.008883	fusiform
19	R_2_LABEL_244	0.008669	superior parietal
20	R_0_LABEL_243	0.008400	lateral occipital
21	A_2_LABEL_98	0.008178	insula
22	R_2_LABEL_238	0.008173	precuneus
23	L_0_LABEL_66	0.007693	rostral middle frontal
24	L_0_LABEL_111	0.007647	superior parietal
25	R_3_LABEL_208	0.007305	supra marginal
26	L_1_LABEL_39	0.007273	lateral orbitofrontal
27	L_2_LABEL_96	0.006760	precuneus
28	L_1_LABEL_10	0.006699	lateral orbitofrontal
29	R_3_LABEL_196	0.006647	insula
30	A_1_LABEL_10	0.006584	para central
31	R_3_LABEL_163	0.006472	insula
32	R_3_LABEL_154	0.006364	insula
33	R_3_LABEL_151	0.006333	temporal pole
34	R_0_LABEL_136	0.006314	para hippocampal
35	L_1_LABEL_119	0.006213	pericalcarine
36	R_3_LABEL_190	0.006174	para central
37	R_1_LABEL_138	0.006132	fusiform
38	R_0_LABEL_176	0.006064	para central
39	L_0_LABEL_79	0.006051	cuneus

RF_{decon}, HC/Patients Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
40	L_1_LABEL_9	0.005965	pars triangularis
41	A_1_LABEL_60	0.005956	postcentral
42	R_1_LABEL_202	0.005867	cuneus
43	R_0_LABEL_221	0.005843	precuneus
44	R_1_LABEL_247	0.005812	postcentral
45	R_0_LABEL_127	0.005795	fusiform
46	A_0_LABEL_69	0.005739	superior parietal
47	L_3_LABEL_58	0.005676	insula
48	R_3_LABEL_128	0.005639	rostral middle frontal
49	R_0_LABEL_219	0.005600	insula
50	R_2_LABEL_189	0.005566	lateral orbitofrontal
51	L_3_LABEL_85	0.005538	bankssts
52	L_1_LABEL_29	0.005366	supramaginal
53	L_3_LABEL_108	0.005344	insula
54	A_0_LABEL_83	0.005333	rostral middle frontal
55	A_2_LABEL_40	0.005332	superior frontal
56	L_1_LABEL_70	0.005324	pars opercularis
57	L_3_LABEL_37	0.005221	precentral
58	A_0_LABEL_36	0.005203	precentral
59	L_1_LABEL_81	0.005112	superior temporal
60	L_3_LABEL_74	0.005098	precentral

Table 16: Most important features given by RF_{decon} in the HC/Patients problem.

RF_{orig}, TRS/NTR Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
1	R_2_LABEL_174	0.024907	precental
2	R_1_LABEL_198	0.020820	lateral occipital
3	R_0_LABEL_172	0.018875	lateral occipital
4	L_3_LABEL_3	0.017318	lateral orbitofrontal
5	R_1_LABEL_147	0.017289	lateral occipital
6	R_2_LABEL_212	0.016667	pars triangularis
7	R_0_LABEL_202	0.016648	lateral occipital
8	L_3_LABEL_43	0.016000	superior parietal
9	R_0_LABEL_229	0.015949	inferior parietal
10	L_0_LABEL_113	0.015510	para hippocampal
11	R_1_LABEL_190	0.015231	inferior parietal
12	R_0_LABEL_233	0.015111	postcentral
13	A_2_LABEL_28	0.014866	para hippocampal
14	R_3_LABEL_134	0.014194	pars triangularis
15	A_2_LABEL_61	0.013622	rostral middle frontal
16	R_3_LABEL_174	0.013554	pars triangularis
17	R_1_LABEL_165	0.013402	superior frontal
18	A_2_LABEL_46	0.013198	supra marginal
19	L_3_LABEL_75	0.012821	supra marginal
20	R_3_LABEL_221	0.012810	para central
21	A_0_LABEL_76	0.012785	inferior temporal
22	R_1_LABEL_132	0.012482	caudal middle frontal
23	L_1_LABEL_59	0.012426	posterior cingulate
24	R_2_LABEL_194	0.012426	pars triangularis
25	L_1_LABEL_33	0.011864	lingual
26	A_2_LABEL_109	0.011462	superior temporal
27	L_3_LABEL_38	0.011429	temporal pole
28	A_1_LABEL_47	0.011383	superior temporal
29	A_1_LABEL_17	0.011313	rostral middle frontal
30	R_2_LABEL_166	0.011200	superior parietal
31	L_3_LABEL_125	0.010712	superior temporal
32	A_0_LABEL_120	0.010562	precentral
33	L_0_LABEL_97	0.010467	lateral occipital
34	L_1_LABEL_7	0.010268	supra marginal
35	R_3_LABEL_249	0.010244	precuneus
36	L_0_LABEL_70	0.009888	pars opercularis
37	L_2_LABEL_55	0.009768	inferior temporal
38	R_0_LABEL_168	0.009524	insula
39	R_1_LABEL_144	0.009474	superior frontal

RF_{orig}, TRS/NTR Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
40	L_0_LABEL_10	0.009318	para central
41	A_2_LABEL_84	0.009288	para central
42	L_1_LABEL_9	0.009176	pars orbitalis
43	L_3_LABEL_21	0.009127	para central
44	L_1_LABEL_47	0.009079	precentral
45	L_3_LABEL_103	0.008938	isthmus cingulate
46	L_1_LABEL_84	0.008869	para central
47	L_1_LABEL_41	0.008802	pre central
48	A_1_LABEL_37	0.008791	lateral orbitofrontal
49	A_2_LABEL_40	0.008713	insula
50	L_3_LABEL_114	0.008696	inferior parietal
51	L_0_LABEL_45	0.008525	superior temporal
52	L_1_LABEL_118	0.008400	postcentral
53	A_0_LABEL_32	0.008219	inferior parietal
54	A_1_LABEL_68	0.008163	pars orbitalis
55	L_1_LABEL_75	0.008143	supra marginal
56	L_1_LABEL_90	0.007925	medial orbitofrontal
57	R_0_LABEL_244	0.007554	superior parietal
58	L_0_LABEL_90	0.007153	medial orbitofrontal
59	R_3_LABEL_131	0.007085	isthmus cingulate
60	R_3_LABEL_236	0.007068	supra marginal

Table 17: Most important features given by RF_{orig} in the TRS/NTR problem.

RF_{decon}, TRS/NTR Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
1	R_2_LABEL_174	0.106037	precentral
2	A_1_LABEL_112	0.057195	entorhinal
3	L_3_LABEL_59	0.045193	posterior cingulate
4	A_1_LABEL_104	0.043622	lateral occipital
5	R_1_LABEL_190	0.043615	inferior parietal
6	L_3_LABEL_75	0.039904	supra marginal
7	L_1_LABEL_22	0.038860	lateral occipital
8	R_2_LABEL_212	0.037150	insula
9	R_0_LABEL_202	0.036869	caudal middle frontal
10	L_3_LABEL_51	0.035626	para central
11	L_1_LABEL_7	0.032370	supra marginal
12	R_3_LABEL_140	0.030604	posterior cingulate
13	R_2_LABEL_179	0.028770	pars triangularis
14	L_3_LABEL_38	0.028022	temporal pole
15	R_3_LABEL_249	0.025367	para central
16	L_1_LABEL_47	0.025210	postcentral
17	R_1_LABEL_142	0.021550	precuneus
18	A_1_LABEL_57	0.021024	superior frontal
19	R_2_LABEL_166	0.019940	postcentral
20	L_3_LABEL_13	0.019517	postcentral
21	A_2_LABEL_108	0.018908	insula
22	R_0_LABEL_191	0.018112	precentral
23	A_2_LABEL_67	0.016821	superior frontal
24	R_0_LABEL_218	0.016544	rostral middle frontal
25	R_2_LABEL_171	0.015329	pars triangularis
26	A_1_LABEL_14	0.015054	inferior parietal
27	R_1_LABEL_178	0.014493	precuneus
28	R_3_LABEL_167	0.014375	precentral
29	R_2_LABEL_233	0.013750	preceneus
30	R_1_LABEL_246	0.013148	precentral
31	A_1_LABEL_66	0.013118	rostral middle frontal
32	R_3_LABEL_240	0.011635	superior parietal
33	L_0_LABEL_108	0.011295	insula
34	A_2_LABEL_106	0.011211	lingual
35	L_2_LABEL_55	0.011184	preceneus
36	A_0_LABEL_29	0.010714	supra margional
37	A_0_LABEL_72	0.009200	postcentral
38	L_3_LABEL_42	0.007463	supra margional
39	L_0_LABEL_69	0.007353	superior parietal

RF_{decon}, TRS/NTR Problem			
Feature Ranking	**Region Label**	**Feature Importance**	**Anatomical Region**
40	L_0_LABEL_103	0.006061	isthmus cingulate
41	R_0_LABEL_209	0.005614	para central
42	A_0_LABEL_53	0.002174	inferior parietal

Table 18: Most important features given by RF_{decon} in the TRS/NTR problem.