Norwegian University of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Science and Technology
Professor Cecilia Marie Futsæther

Identification of biomarkers from Radiomics of brain scans for prediction of major depression using Repeated Elastic Net Technique

Krishna Mohan Shah
Master of Science in Data Science

This page is intentionally left blank.

Acknowledgements

Foremost, I express my profound gratitude to my advisor, Prof. Cecilia Marie Futsæther, for the continuous and enthusiastic feedback during this project.

Moreover, I could not have undertaken this journey without my co-supervisor, Prof. Oliver Tomic, who participated in meetings regarding my project and came with valuable feedback.

Further, PhD candidate Bao Ngoc Huynh and PhD candidate Anna Jenul helped me find my footing as I started this process and generously took time out of their schedules to participate in my research and make this project possible. Finally, I thank Professors Liesbeth Reneman, Matthan Caan, Atle Bjørnerud and PhD student Maarten Poirot for access to the dataset, without which this research would not be possible.

Next, I express thankfulness to all my friends who helped me proofread this thesis and for their thoughtful feedback.

Finally, I thank my parents for supporting me through this thesis and life.

Krishna Mohan Shah
Ås, June 15th 2022

Abstract

The application of machine learning in the field of medicine is expanding on an almost daily basis. Data from the healthcare industry typically have high dimensionality but a limited sample size. The learning process can be sped up, system performance can be improved, complexity can be minimised, and the risk of overfitting may be decreased by selecting a smaller subset of relevant features from the high-dimensional data set.

The primary objective of this investigation was to diagnose patients with major depressive disorder (MDD) using radiomics features extracted from MR images. In addition, the thesis tries to accomplish the objective by locating a collection of biomarkers that can assist in developing individualised treatment plans. Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care (EMBARC) was the source of the data.

Before beginning the classification process, the data set's dimensionality was decreased by applying a technique known as RENT, which stands for Repeated Elastic Net Technique for Feature Selection. Logistic Regression, Support Vector Machines, and Random Forest are three common classifiers utilised in computing the performance of all features and the RENT chosen features predictions. A technique known as principal component analysis (PCA) was used for the analysis of the data. Throughout the splits and the dataset, RENT chose eleven characteristics in all. According to RENT, the rostral middle frontal cortex may be a significant biomarker that can predict people who have MDD.

Contents

Appendices

List of Figures

List of Tables

Abbreviations

Abbreviation	Meaning
ACC	Accuracy
CT	Computed Tomography
DALY	Disability-adjusted Life Years
DSM	Diagonistic and Statistical Manual of Mental Disorders
DTI	Diffusion Tensor Imaging
EMBARC	Establishing Moderators and Biosignatures of Anti-depressant Response for Clinical Care
FDA	Food and Drug Administration
fMRI	function Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HDRS	Hamilton Depression Rating Measure
LR	Logistic Regression
MCC	Matthews correlation coefficient
MDD	Major Depressive Disorder
MRI	Magnetic Resonance Imaging
NIH	National Institutes of Health
PET	Positron Emmision Tomography
PLSR	Partial-Least-Squares Regression
PR	Precision
RC	Recall
RENT	Repeated Elastic Net Technique
RSKF	Repeated Stratified K-Fold
SVM	Support Vector Machine
TP	True Positive
TPR	True Positive Rate

Chapter 1

Introduction

1.1 Motivation and background

Major depressive disorder is a prevalent mental disorder characterized by poor mood, reduced mental and motor activity, pessimism, and loss of interest in life. Unlike other kinds of depression, major depression has complicated symptoms. Depression is a primary cause of disability, and a major contribution to the global illness burden [1]. Depressive symptoms rose from 27.8% in 2020 (CI: 24.9, 30.9) to 32.8% in 2021. (95 per cent CI: 29.1, 36.8) [2].

Compared to the general population, graduate students are six times as likely to experience depression [3]. Due to its relapsing and repeated nature, depression is a significant disorder that must be prevented and treated by identifying its susceptibility factors. Depression often begins in early adolescence or young adulthood. Untreated depressive people have a 20% lifetime suicide risk [4]. Stressful life experiences can trigger depression; however, this link weakens with recurring instances [5]. Not everyone who endures a traumatic incident gets depression [6].

Antidepressants are the principal treatment for moderate to severe depressive episodes, yet six decades of research have not improved their effectiveness [7] [8]. Antidepressants are commonly prescribed ‘trial-and-error’ style for depression [9]. Antidepressants require 2 to 8 weeks to take action; if none do, a new one is given. Precision medicine Precision medicine uses daily healthcare data to give the most effective treatment or preventative care to the appropriate people at the right time [10]. Precision medicine shortens depressive patients’ treatment courses [9]. Biomarkers help target precision medication. Biomarkers are molecular, anatomical, physiological, or biochemical properties [9].

MR images are converted into high-dimensional radiomics datasets. Machine learning algorithms can discover patterns in these datasets that humans cannot. Biomarkers may be used to diagnose MDD and guide treatment. Radiomics analysis can be done on medical images from several sources, allowing for an integrated cross-modality approach exploiting the potential additive value of imaging information acquired from MRI, CT, and PET scans [11]. Following the image acquisition and segmentation processes, radiomic characteristics are obtained using a radiomics framework such as pyradiomics, among others. Radiomics generates many potential image biomarkers . Nevertheless, there are several challenges involved in the process of identifying biomarkers using MR images. Even with the same subject and MR scanner, MR pictures might vary [12]. Medical imaging biomarkers must be confirmed and repeated to assure dependability [13]. During radiomics, many features are extracted; nevertheless, it is possible that the majority of those features will not supply the machine learning models with any helpful information. Therefore, the results of machine learning models are not improved. Therefore, it is necessary to cut down on the number of features by utilizing the feature selection process. The features that provide the most information to the models are the ones that are chosen via feature selection techniques. Repeated Elastic Net Technique (RENT) is used to select features. RENT is an ensemble-based feature selection technique that uses a logistic regression (LR) model with elastic net regularization trained on multiple subsets of data to discover resilient features for binary classification tasks [14]. Using the Repeated Elastic Net Technique (RENT) and radiomics characteristics generated from MR images, this thesis sought to establish a method for the early detection of individuals suffering from major depressive disorder(MDD). In addition, further research looked at the possibility that the radiomics characteristics studied may serve as biomarkers for depression.

1.2 Problem statement

This project has two distinct objectives, all of which are interrelated.The data are short and wide, meaning there is a large ratio between the number of rows and the number of columns, with the latter having a significantly greater total. Therefore, the first thing that must be done to prepare the data for the feature selection process is data preprocessing. Preparing (by cleaning and organizing) the raw data to make it appropriate for use in the development and training of machine learning models is called data preprocessing. Preprocessing will allow the data to be used more efficiently. Searching for missing data, duplicate columns, correlated characteristics, and other anomalies before deciding whether or not to eliminate them is a common definition of preprocessing. The reader ought to have prior experience with 'normal' machine learning (for example, having studied the majority of the book [15]) to understand the terms easily.

The second objective is to use the RENT feature selection approach to identify the most significant collection of features. This will assist in condensing a large dataset down to only the most significant aspects. Also, use classification models to predict whether or not the patient is suffering from a major depressive disorder. The final step is to create a framework that performs well with this dataset in the hopes that it will also perform well with new data.

1.3 Structure of the thesis

The introduction to the thesis will cover the theory behind machine learning and the many approaches taken, as outlined in Chapter 2. The data, the methods and the application of the algorithms to the data are both covered in Chapter 3. The preprocessing of the data and the findings from the study are provided in Chapter 4, followed by a more in-depth discussion with further work in Chapter 5. Finally, a conclusive statement about the analysis and the theory can be found in Chapter 6.

Chapter 2

Theory

2.1 Depression

Depressed mood, reduced interests, poor cognitive function, and vegetative symptoms such as disrupted sleep or hunger describe major depressive disorder (MDD) [16]. Antidepressants from various classes are typically given for the treatment of MDD; however, clinically applicable accurate and repeatable measures of efficacy are not yet available [17]. After cardiovascular diseases and lung cancer, major depressive disorder was among the top five causes of disability-adjusted life years (DALYs) in the United States in 2010 [18].

Depression is a prevalent disorder that affects around 280 million people worldwide, with an estimated 3.8 per cent of the population afflicted, including 5.0 per cent of adults and 5.7 per cent of persons over the age of 60 [19]. Unfortunately, more than seventy-five per cent of people living in low- and middle-income countries do not receive treatment for mental problems, even though there are established therapies that are effective [20].

Due to the inaccessibility of the human brain, there is no scientific or histological test for definite psychiatric diagnosis, unlike with cancer. Standard nosology, as represented in diagnostic manuals such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases, bases the diagnosis on a combination of symptoms alone [9].

The Hamilton Depression Rating Measure, sometimes known as the HDRS, is the depression evaluation scale used most frequently in clinical settings. There are 17 items (HDRS17) in the original form that refer to symptoms of depression that have been encountered over the previous week. As it was first designed for use with hospital inpatients, the HDRS places a strong focus on both the emotional and physical manifestations of depression [21].

When treating depressive illness, starting with the minor invasive measures is typical and adding more if necessary [8]. Current therapy for Major Depressive Disorder is based on trial-and-error, which delays response and remission for most patients. Prolonged unsuccessful therapy raises patient suffering and expenditures. Long and failed antidepressant trials may lower patient expectations, reinforce negative cognitions, and train patients not to react in subsequent trials, adding to treatment resistance. Identifying dependable indicators of antidepressant treatment response may shorten or eliminate unsuccessful trials [22]. Preventive and therapeutic measures that consider individual variability are known as personalized or precision medicine. As a result of population diversity, personalized therapy may be able to reduce the length of treatment. An individual patient's surroundings, genes, and way of life are all considered while looking at biomarkers [8].

As per Food and Drug Administration - National Institutes of Health (FDA-NIH) Biomarker Working Group, biomarkers can be defined as 'A defining characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention.' [23]. Inflammatory, neurotransmitter, neurotrophic, neuroendocrine, and metabolic indicators may help predict mental and physical health outcomes in people who are presently depressed. However, research so far has been inconsistent in its conclusions about their usefulness [24]. Using these biomarkers, an algorithm may choose the patient's therapy and give the physician data to conduct an individual evaluation of the patient. Individualized therapy has changed cancer treatment by using the tumour's genetic profile as a guide. Since there is no medical test to establish a psychological diagnosis, applying individualized therapy in psychiatry is difficult [9].

The use of medical imaging of the brain allows for the extraction of biomarkers. The quantitative imaging biomarkers, which may increase the sensitivity, specificity, accuracy, and repeatability of observed features utilized for diagnostic and therapeutic decision making, are driving a push toward the quantification of imaging data [25]. PET images the brain on a molecular level, functional MR imaging (fMR imaging) obtains physiological parameters, and MR spectroscopy extracts perfusion imaging and biochemical properties [9].

Fronto-limbic areas, including the hippocampus, prefrontal, anterior cingulate cortex, amygdala, and insula, are often predictive of therapy response for MDD patients. From 95 studies, Fonseka et al. (2018) [13] discovered numerous putative biomarkers for treatment response from structural and functional neuroimaging modalities, while Konarski et al. (2008) [26] evaluated 140 magnetic neuroimaging investigations of either bipolar disorder or MDD and found comparable findings. Brandt et al. [27] found 16 different sets of biomarkers with significant promise. Biomarkers were discovered in fronto-limbic areas, including the prefrontal cortex, anterior cingulate cortex, hippocampus, amygdala, and insula; however, the intensity and direction of the relationship varied. MDD patients had decreased striatal and amygdala sizes. Lacerda et al. (2004) [28] found less gray matter in MDD patients' OFC. In addition, MDD patients had reduced hippocampus volumes. Kristin et al. (2021) [8] revealed that the orbitofrontal gyrus may predict depression.

Obtaining useful biomarkers for psychiatric disorders is not as easy as it sounds. Biomarkers variability, medications, different diagnostic protocols, and costs are some of the problems [29]. Radiomics is a quantitative approach to medical imaging that uses complex, non-intuitive arithmetic to improve physicians' data.

2.1.1 Magnetic Resonance Imaging (MRI)

Bloch and Purcell described NMR in 1946 [30]. In 1980, Nottingham and Aberdeen created the first clinical MRIs, and MRI is today a potent clinical tool. MRI provides high-resolution pictures without ionizing radiation [31].

MRI leverages the body's intrinsic magnetic characteristics to provide detailed pictures. For imaging, the abundant hydrogen nucleus (a single proton) is employed [32]. The hydrogen proton is like a spinning planet with a north-south pole. It's like a little bar magnet. Usually, these hydrogen proton 'bar magnets' spin in the body with random axes [32]. Protons' axes align under a strong magnetic field like an MRI scanner [32]. This uniform alignment provides an MRI-axis-aligned magnetic vector. MRI scanners range from 0.5 to 1.5 Tesla [32]. Radio waves deflect the magnetic vector when applied to a magnetic field. The radio wave frequency (RF) that resonates hydrogen nuclei depends on the element (hydrogen), and magnetic field intensity [32]. The magnetic field may be electrically adjusted from head to toe using a series of gradient electric coils. By modifying the local magnetic field by tiny increments, various body slices will resonate at different frequencies [32]. Figure 2.1 shows before and after the application of a magnetic field and how the proton is oriented in space.

When the radiofrequency source is turned off, the magnetic vector returns to its

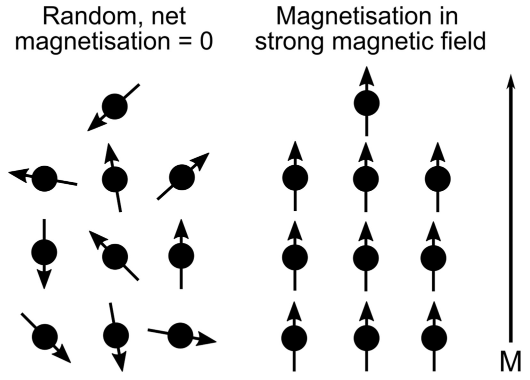
Random, net magnetisation = 0
Magnetisation in strong magnetic field
M

Figure 2.1: Before and after the application of a magnetic field and how the proton is oriented in space [33]

resting condition, emitting a radio wave signal. This signal produces MR pictures. Receiver coils are wrapped around the body portion to increase signal detection [32]. Cross-sectional pictures are created by plotting the received signal's intensity on a grey scale [32].

Sequential radiofrequency pulses can highlight specific tissues or problems [32]. Various tissues relax at different speeds when the radiofrequency pulse is turned off. Protons relax in two ways. First, the magnetic vector must come to rest before the axial spin can [32]. In magnetic resonance (MR) imaging, a T1-weighted (T1W) image reveals signal changes based on the tissue's intrinsic T1 relaxation time [34]. When creating contrast in images, repetition time (TR) and echo time (TE) play a critical role [35]. With short TE and TR periods, T1-weighted pictures can be generated. In contrast, longer TE and TR periods yield T2-weighted pictures [34].

MR exams use pulse sequences. Various tissues (fat and water) have different relaxing periods. By utilizing a 'fat suppression' pulse sequence, the signal from fat is eliminated, leaving just abnormalities [32].

fMRI studies structure and function simultaneously [31]. fMRI uses magnetic field inhomogeneities caused by oxygenated and deoxygenated haemoglobin. Oxygenated haemoglobin is less paramagnetic than deoxyhemoglobin; hence no exogenous agent is needed. Because oxygenated blood flows to a tissue, an fMRI will look different before and after (change in blood oxygenation) [31]. This is because activated brain regions have increased blood flow. fMRI provides the same functional information as PET without radionuclides [31].

MRI is sensitive to illness because most diseases increase the water content. Infection and tumour might seem similar, making pathology difficult to determine

[32]. The image analysis performed by a radiologist is often superior to that of a non-radiologist [32].

MRI has no known biological dangers since, unlike x-rays and CT scans, it employs harmless radiofrequency radiation [32]. However, pacemakers, metal clips, and metal valves can move dangerously in MRI scanners. MRI is becoming more common in clinical practice as costs and availability drop [32].

2.1.2 Radiomics

The growing translational field of study known as ‘radiomics’ has the primary objective of eliciting high-dimensional data that can be mined from clinical pictures [36]. Images are analyzed by software that uses mathematical techniques to extract quantitative characteristics, which are then used as descriptors. Image acquisition is the first phase in the radiomics pipeline, followed by segmentation, feature extraction, feature selection, and finally, modelling and assessment [37]. The list below describes the steps in the radiomics pipeline.

- Image Acquisition: Computed Tomography (CT) scans, Positron Emission Tomography (PET) scans, and Magnetic Resonance Imaging (MRI) scans are the most prevalent types of medical imaging techniques (MRI) [37].
- Segmentation: The process of determining the perimeter of a lesion based on a picture or a sequence of images, which may be done manually with the help of interactive computer tools or automatically with the use of image segmentation algorithms [38].
- Feature Extraction: Extraction of high-dimension feature data to quantitatively describe characteristics of volumes of interest is at the heart of the field of radiomics. Agnostic features are those that aim to capture lesion heterogeneity through quantitative descriptors. In contrast, semantic features are routinely used in the radiology vocabulary to characterize regions of interest. Semantic features are used to describe regions of interest [39].
- Feature Selection: As a result of the feature extraction stage, the radiomics pipeline contains a vital step called feature selection. This is because the feature extraction step produces a large number of features. In addition, due to the constraints imposed by clinical trials on the collection of samples, the dataset only contains a limited number of samples while having a wealth of attributes. For the following three reasons, feature selection is critical for

challenges involving short-wide datasets: 1) to solve the problem known as the ‘curse of dimensionality’; 2) to condense the input data to shorten the amount of time it takes for the model to run; and 3) to make the outcome more easily understandable [40].

- Feature selection, a dimensionality reduction strategy, removes unnecessary, redundant, or noisy characteristics to choose a limited group of valuable features. Feature selection can improve learning accuracy, computational cost, and model interpretability [41]. Methods can be categorized according to feature selection categories: Filters choose features without a classifier, Wrapper models use classifiers to discover the best features, and Embedded approaches look for the optimum model feature subset [42].
- Modelling and assessment: The remaining features, which are significant and independent of one another, may be utilized to train the model for the reasonable prediction of classification using different machine learning algorithms.

2.2 RENT

High-dimensional biological datasets may contain duplicate, noisy, and irrelevant information, lowering classification performance and raising processing costs. Feature selection is used to reduce noisy information and find diagnostic patterns [43]. Repeated Elastic Net Technique (RENT) is an ensemble-based feature selection strategy that seeks to identify resilient features for binary classification problems by utilizing a logistic regression (LR) model with elastic net regularization that is trained on different subsets of data [14]. The workflow of RENT is shown in the figure 2.2.

The distinct subsets are produced by randomly selecting the primary training data while simultaneously replacing some samples. This results in creating one-of-a-kind subsets for each model in the ensemble. By determining the frequency with which a feature is picked across numerous models, using multiple models enables a more accurate evaluation of the relevance of the features being considered [37]. Elastic Net decides which characteristics should be included in each model. The characteristics that are not selected weight zero, whereas the features that are picked have a weight that is not zero [44].

Each trained model is equipped with a vector of feature weights denoted by n that is then included in a weight matrix denoted by **B**. In a space with N dimensions

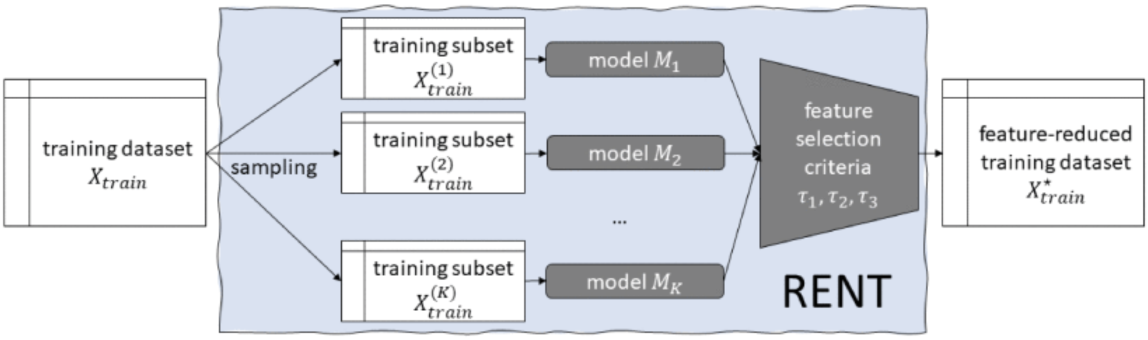

Figure 2.2: The workflow of RENT. RENT splits and trains the input dataset across K submodels and selects the features based on three criteria that quantify the feature selection percentage, stability, and weight. The output is the set of selected features [14].

for features, the weights matrix $\mathbf{B}$ will have dimensions of the form $(K * N)$. A threshold $(\tau 1)$ that the user provides controls with the frequency with which the feature should be chosen from among all K models.

$$\tau\left(\beta_n\right) = c\left(\beta_n\right) \tag{2.1}$$

where $c\left(\beta_n\right)$ determines how important a characteristic is based on how often it occurs on average [44], given by

$$c\left(\beta_n\right) = \frac{1}{k}\sum_{k=1}^{K} 1, [\beta_{k,n} \neq 0] \tag{2.2}$$

The feature is considered stable if just a few instances of the weights' signals switch between positive and negative values $(\tau 2)$. It would be ideal for a feature to have weights that are all the same sign, either all positive or all negative. Only when all of the non-zero weights have the same polarity can the value for $\tau 2$ reach its most significant potential of being equal to or greater than the value for $\tau 1$. The user is given the option of indicating the desired number of proportions of feature weights with the same sign $(\tau 2)$ [44].

$$\tau 2\left(\beta_n\right) = \frac{1}{k}|\sum_{k=1}^{K} sign\left(\beta_{k,n}\right)| \tag{2.3}$$

The feature routinely exhibits non-zero weights across all K submodels while having a very low variance ($\tau 3$). The $\tau 3$ criterium is defined as

$$\tau 3\left(\beta_n\right) = t_{K-1}\left(\frac{\left|\mu\left(\beta_n\right)\right|}{\sqrt{\frac{\sigma^2\left(\beta_n\right)}{K}}}\right) \tag{2.4}$$

where $/mu$ is feature specific mean , $/sigma$ is the variance and t_{K-1} is is the cumulative density function of students t-distribution with $K1$ degrees of freedom. The user has the ability to set a significant level threshold value ($\tau 3$) for the analysis. The value $\tau 3$ denotes the cumulative distribution function of the Student's tdistribution when $K1$ degrees of freedom are included [44]. Every one of the quality measurements is confined inside the range of 0 to 1 ([$\tau 1$, $\tau 2$, $\tau 3$] $\in$ [0, 1]) [37]. $\tau 3$(n) is a t-test, therefore $\tau 3 = 0.975$ yields a 5% significance level [8]. These selection criteria help the user to define the strictness of the feature selection process.

2.3 Correlation

Correlation evaluates whether two variables fluctuate and reflects the degree of their relationship. Finally, covariance defines the linear connection between two properties and how they change together [15].

The correlation coefficient may be calculated by dividing covariance by feature standard deviations. The Equation shows this,

$$CORR\left(x_j, x_k\right) = \frac{\sigma_{jk}}{\sigma_{x_j}\sigma_{x_j}} \tag{2.5}$$

where, σ_{x_j} and σ_{x_j} are the sample standard deviations, and σ_{jk} is the sample covariance. Correlation ranges from -1 to 1. The two features are connected if their correlation is closer to -1 or 1. When the coefficient is negative, a rise in one property indicates a reduction in the other feature. The correlation coefficient will be 0 [15] if there's no link.

2.4 One-hot encoding

The characteristics of a dataset might be of several data kinds. Similar data can also be grouped into a restricted number of groups by a characteristic. Most algorithms demand numerical input. Hence categorical data should be translated into numeric data. For each category value, One-Hot Encoding produces a new column. These columns are given the values 0 and 1 [45].

2.5 Outliers

Outliers are considered suspicious because they are so far out of the norm. The issue is that even a small number of outliers can significantly skew the overall results (by altering the mean performance, increasing variability, etc.) [46]. Most machine learning algorithms perform poorly when an outlier is present. It is therefore desirable to identify and eliminate any outliers. Every dataset contains some data that stands out from the rest for some reasons; some of the most prevalent explanations are: malicious activity, instrumentation error, change in the environment, and human error [47]. It is essential to eliminate any outliers that are the product of improper misrepresentation. The interquartile range, also known as IQR, is a method that can be utilized to assist in locating outliers in data that is continuously distributed. This method orders the dataset into four equal parts by dividing it into quartiles and then dividing each of those quartiles into the dataset. For example, Q1, Q2, and Q3 are first, second, and third quartiles [48].

$$IQR = Q3 - Q1 \tag{2.6}$$

$$upper\_bound = Q3 + 1.5 * IQR \tag{2.7}$$

$$lower\_bound = Q1 - 1.5 * IQR \tag{2.8}$$

The figure 2.3 shows the range of Q1, Q3 and their upper bound and lower bound.The data points considered to be outliers are those that either fall below the lower or above the upper bound.

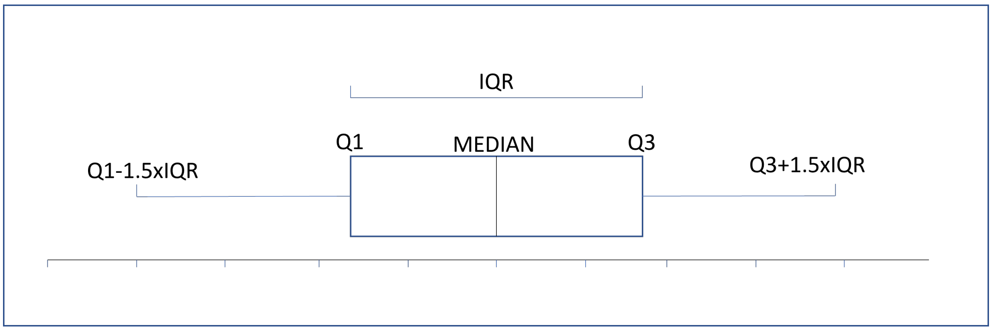

Figure 2.3: The workflow of outliers detection and how is it calculated. The value above $Q3 + 1.5 * IQR$ and below $Q1 - 1.5 * IQR$ is regarded as outliers [48].

2.6 Variance Threshold

Constant features have the same or comparable values across the dataset. Machine learning algorithms cannot make reliable predictions about the target based on these features as they provide little to no information. High predictor variance is beneficial, but low predictor variance is not.

By utilizing Sklearn's Variance Threshold, we can eliminate constant features [45]. The Variance Threshold algorithm is a feature selector that removes from the dataset all of the low variance features that are of little value when it comes to modelling [45]. It concentrates solely on the characteristics (x), ignoring the desired response (y) [45]. For example, using a criterion of 0.01 would eliminate the column in which 99.9 per cent of the values are identical.

2.7 Data Scaling

The data may be transformed in several ways, one of which is by scaling the data to restrict the value range. The data ($/X$) are then scaled to be centred around the mean ($/mu$) with a standard deviation ($/sigma$) of one unit as part of the standardization process [44], shown in Figure 2.4. The numbers do not need to fall inside a specific range.

$$X' = \frac{X - \mu}{\sigma} \tag{2.9}$$

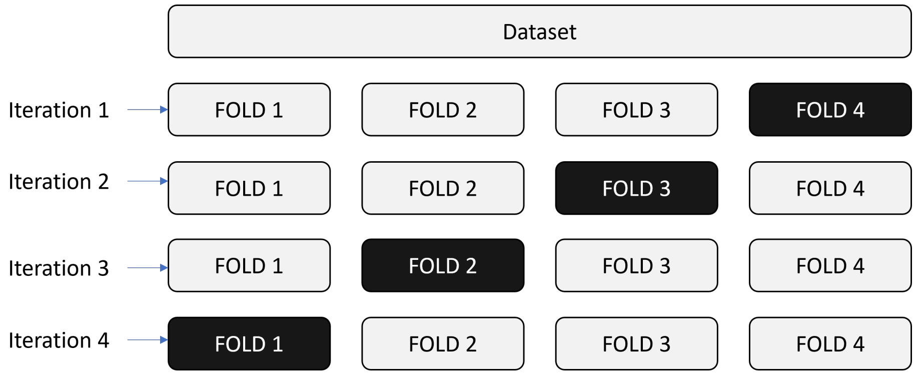

Figure 2.4: The basic representation of the split using RSKF with 4 folds and one repetition. At each iteration, three folds (lightly shaded) are used for training, and the remaining fold (dark shaded) is used for testing.

2.8 Model validation

2.9 Splitting the dataset

When there is a very small sample size, it is exceedingly challenging to partition the dataset into training and test datasets. Therefore, the best method for dividing the data for training and testing is to employ Repeated Stratified K-Fold [45].

2.9.1 Repeated Stratified K Fold

Medical datasets have few samples for many reasons. With few samples, the splitting approach fails because of inadequate training and validation data. Cross-validation increases the model's generalizability in this case. A modest change on the K Fold cross-validation approach is developed such that each fold has around the same percentage of target class samples as the whole set. In case of prediction difficulties, the mean response value is about the same in all folds. Stratified K Fold describes this variant. Repeated Stratified K-Fold repeats Stratified K-Fold with different randomizations, which divide data into training and test sets to examine model generalizability as shown in figure 2.4. K is the model's training frequency [45].

2.10 Score Metrics

Scoring metrics are measurements that represent a model's performance. Accuracy, F1 score, Matthews correlation coefficient (MCC), and Area under curve (AUC) were the measures utilized in the thesis to calculate the model's performance. True positive, true negative, false positive, and false negative can be used to determine all measurements [49].

True positives are actual positives that are accurately anticipated positives (TP). False negatives are actual positives that were incorrectly forecasted as negatives (FN). True negatives are actual negatives that are accurately anticipated negatives (TN). False positives are actual negatives incorrectly forecasted as positives (FP).

Accuracy is calculated by

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.10}$$

Another accuracy indicator is precision (PR), which evaluates how many depressed patients were properly predicted. Recall (RC) is the ratio of how many patients were depressed compared to how many people were projected to be depressed [49]. The weighted average of the precision and recall scores is the F1 measure.

$$PR = \frac{TP}{TP + FP} \tag{2.11}$$

$$RC = \frac{TP}{TP + FN} \tag{2.12}$$

$$F1 = 2\frac{PR.RC}{PR + RC} \tag{2.13}$$

The Matthews correlation coefficient (MCC) is a more reliable statistical rate that produces a high score only if the prediction performed well in all four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally to the size of positive ($\hat{P}$) and negative ($\hat{N}$) elements in the dataset [50].

$$MCC = \frac{TP.TN + FP.FN}{\sqrt{P.(TP + FN).(TN + FP).N}} \tag{2.14}$$

The receiver operating curve (ROC) is shown in AUC based on the true positive rate (TPR) and false positive rate (FPR) (FPR). The larger the ROC, the better the model's performance [49].

$$TPR = \frac{TP}{TP + FN} \tag{2.15}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.16}$$

2.11 Model algorithm

2.11.1 Logistic Regression

Logistic Regression (LR) is a linear classification approach. It presupposes linearity between the dependent variable's logit and the independent variable (predictor). Logistic Regression employs the logistic sigmoid function (z) defined as:

$$\phi\left(z\right) = \frac{1}{1 + e^{-z}} \tag{2.17}$$

The activation function's net input is z. It translates net input to [0, 1], representing the sample's class probability. A threshold function converts probability to binary. Logistic Regression's threshold is mathematically described as:

$$z = \begin{cases} 1, & \text{if } \phi\left(z\right) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{2.18}$$

2.11.2 Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning technique for classifying data. The classification method finds a hyperplane that divides the two classes, termed the decision boundary, as shown in Figure 2.5. It maximizes the distance between decision boundaries and samples. Large margins reduce generalization errors in models [15].

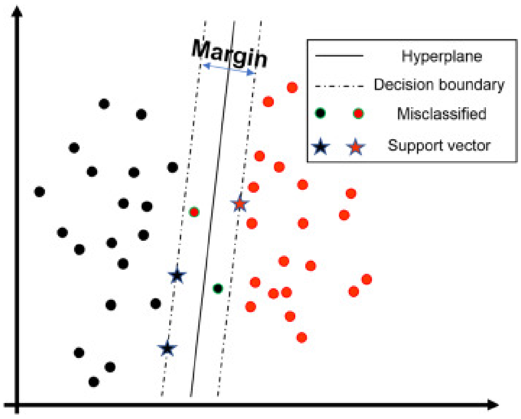

Figure 2.5: The ideal hyperplane, support vectors, and margin employed in the SVM method are illustrated here. Figure found in [51]

In this particular instance, we are attempting to differentiate between the dark and the samples samples, which we may refer to as hyperplane. The ideal hyperplane, support vectors, and margin employed in the SVM method are illustrated in figure 2.5 [51]. The support vectors are the samples nearest to the decision border [44].

2.11.3 Random Forest

The random forest model is an example of an ensemble tree-based learning algorithm; the method takes the predictions from many individual trees and calculates an overall average [52]. It is feasible to construct a model that generalizes better than one tree on its own by integrating numerous trees into a single model. The risk of the model becoming overly dependent on the training data may be mitigated by ensuring an adequate number of trees [44]. After the tree has been created, a set of bootstraps that do not include any specific record from the original data [out-of-bag (OOB) samples] are used as the test set [53]. In the figure, one can find a concise explanation of the algorithm behind the random forest.

Random forest is more stable in the presence of outliers and very high dimensional parameter spaces when compared to other machine learning algorithms because it adheres to certain principles for tree building, tree combination, self-testing, and post-processing [54] [55]. The pseudocode can be seen in figure 2.6. In comparison to decision trees, the estimation of the error rate produced by the random forest method is far more precise [52].

```
for $i \leftarrow 1$ to $B$ do
    Draw a bootstrap sample of size $N$ from the training data;
    while node size != minimum node size do
        randomly select a subset of $m$ predictor variables from total $p$;
        for $j \leftarrow 1$ to $m$ do
            if $j$th predictor optimizes splitting criterion then
                split internal node into two child nodes;
                break;
            end
        end
    end
end
return the ensemble tree of all B subtrees generated in the outer for loop;
```

Figure 2.6: The psuedocode of random forest classifier. [52]

.

2.12 PCA

PCA is an unsupervised, non-parametric data analysis approach [56]. Principal components can decrease a dataset's dimensionality while keeping systematic information. The orthogonal transformation turns correlated properties into linearly uncorrelated ones [15]. PCA creates additional axes, termed principal components (PC), along the direction of greatest variance. PCA functions as follows: First, normalize the data since principal component scaling is sensitive. The covariance matrix is next. Calculating covariance between characteristics yields the covariance matrix. Eigenvalue decomposition decomposes the covariance matrix into eigenvectors and eigenvalues [44]. Eigenvector elements are the original data's weight coefficients or loadings. The following condition holds for an eigenvector, $\vec{v}$, with eigenvalue, λ.

$$\sum \vec{v} = \vec{v} \tag{2.19}$$

For the reduction, choose the subset of eigenvectors $\vec{v}$ that contributes most to variance, then use the eigenvalues λ to compute explained variance. λ_j is the specific eigenvalue. For example, If $d = 2$, the covariance matrix gives 3X3 dimension matrix where $_{12}$ is the covariance for $j = 1$ and $k = 2$.

$$Explained\_variance\_ratio = \frac{\lambda_j}{\sum_{j=1}^{d} \lambda_j} \tag{2.20}$$

which is the fraction of the given eigenvalue divided by the total sum of all the eigenvalues[44]. Similarly to eigenvalues, eigenvectors are ordered. The k top related eigenvectors may be picked from this to represent the new feature subspace. k must be less than d to reduce dimension. It is best to choose the subset of eigenvectors that includes most of the data's information or the number of major components that most of the variance. Next, build a transformation matrix, **W**, from the top k eigenvectors. This matrix can convert the original data set into a new feature subspace [44].

2.13 PLSR

Partial least squares regression (PLSR) is the statistical technique used to investigate the nature and magnitude of the connections between variables. Regression analysis estimates characteristics based on past data [57]. The equation that can be used to describe the regression model is

$$y = a + bx \tag{2.21}$$

in which a represents the intercept of the line that best fits the data, and b represents the slope of the line. The individual sample's divergence from the line is referred to as the residuals. The best-fitting line to the data is determined by selecting a line that minimizes the sum of the squared residuals. The line that achieves the lowest possible value for the sum of the squared residuals is known as the least-squares line, and the equation used to determine its slope is shown below.

$$b = \frac{SS_{xy}}{SS_{xx}} \tag{2.22}$$

In this equation, the total of the cross products is denoted by SS_{XY}, while the sum of the squares for variable x is denoted by SS_{XX} [44]. A supervised method for conducting exploratory data analysis, partial least squares regression, or PLSR, uses partial least squares. It projects the data into a new subspace, just like principal component analysis (PCA); however, in contrast to PCA, it projects both the X and Y data at the same time [44].

Chapter 3

Materials and Methods

3.1 Data

The purpose of the EMBARC (Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care) study [58] was to discover why antidepressant medication works for some people with MDD, but not for others. During the course of the EMBARC project, participants were observed to monitor the development of their symptoms, and they also had MRI scans, the data from which will ideally enable researchers to identify a collection of biosignatures that can assist in the process of specific treatment planning. Recent improvements in machine learning allow for the study of vast sets of factors, even if no single MRI variable has demonstrated adequate predictive ability to predict treatment response when analyzed individually.

Sertraline is a commonly used medicine in treating a major depressive disorder, obsessive-compulsive disorder, panic disorder, PTSD, premenstrual dysphoria, and social anxiety disorder [59]. For sixteen weeks, 309 participants were randomly allocated to either a placebo or sertraline arm. Baseline MRI scans included T1-weighted and Diffusion Tensor Imaging (DTI). The degree of anisotropy in the brain microstructure may be assessed using DTI. MDD and white matter alterations are linked, according to a number of studies [60]. Analytical and processing methods for DTI data have been detailed at great length [61]. Images of fractional anisotropy (FA) were created for each participant by fitting diffusion tensor models. The Desikan–Killiany–Tourville atlas was utilized to construct cortical and subcortical segmentations using FastSurfer [62]. PyRadiomics was used to construct regional image characteristics called radiomics from these segmentations and FA-maps [63].

3.1.1 Description of datasets

PhD student Maarten Poirot at Amsterdam University Medical Center provided the processed data extracted from MR images. There are 297 patients in the datasets. A total of 28 patients were eliminated because of missing data, resulting in a dataset of 268 subjects, 138 of which were in the placebo group and 130 of whom were in the sertraline group (1). In these datasets, we solely used seratline-treated class (1). Only seven demographic data (age, gender, etc.) were included in the 10175 attributes; the rest were radiomics features characterizing the anatomy of the brain (shape, texture, etc.) in dwi-dataset. In anatomical dataset, there are 20340 features for the same patients. Table 3.1 shows the dimension of the data provided for this thesis. Table 3.2 shows decription of non-radiomics features in the data.

Data	**Dimention**
dwi dataset (Clinical)	(296,10175)
anat dataset (Anatomical)	(296,20340)

Table 3.1: The initial size and dimension of the data.

Table 3.2 shows the description of non-radiomics features in the data.

Feature name	Description	Data Type
w8_responder	Depression or not	Categorical
age	Age of the patients	Continuous
race	Race [White, Asian, Black or African American and Other]	Categorical
gender	Gender [Male, Female]	Categorical
hispanic	Hispani [Hisp-No, Hisp-Yes]	Categorical
Stage1X	Medicine given on the first stage of trail [sertaline, Placebo]	Categorical
w0_score_17	Initial Score patients according to HDRS17	Continuous
w8_score_17	End of week 8 trial Score patients according to HDRS17	Continuous
w16_score_17	End of Week 16 trial Score of the patients according to HDRS17	Continuous

Table 3.2: The decription of non-radiomics features in the data.

3.2 Software

In the study, we utilized the Anaconda open source distribution to implement the study's software [64]. Python 3.7.13 was used in this project. As far as the data processing goes, we used Pandas [65] version 1.3.5 and NumPy [66] version 1.21.5. PCA and PLSR exploratory analysis and visualization were performed using the Hoggorm [67] and Hoggmplot [68] packages. Matplotlib [69] version 3.5.1, Plotly [70] version 5.8.0, and Seaborn [71] version 0.11.2 are also used for visualization. Data processing and machine learning were handled using Scikit-learn [45] and RENT [14], respectively, versions 0.0.1 and 1.0.2.

3.3 Workflow

The workflow that was utilized for the study may be seen in the figure 3.1. Data were preprocessed to ensure robustness and reliability. The samples in the test set can have a big influence on the model's prediction when using a short wide dataset. Using the train-test split just once would not give us a legitimate result since the

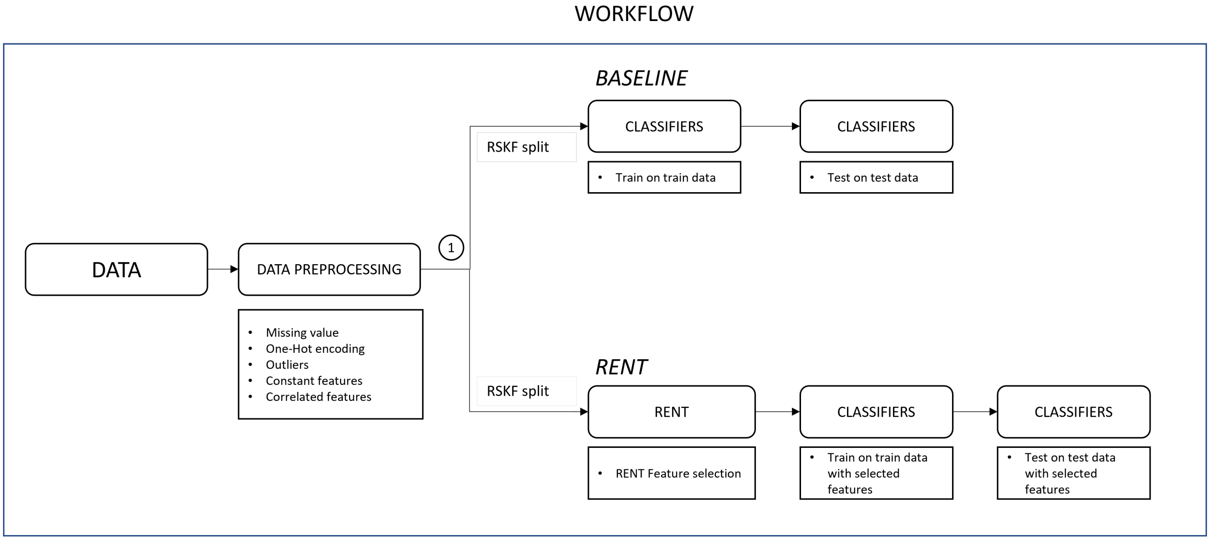

Figure 3.1: The workflow used in the study. PCA and PLSR were done at point 1.

split may sometimes produce high performance, and other times may give poor performance. The data was divided into a training and a test set using RSKF cross validation with four folds and three repetition to overcome the problem.

3.4 Data preprocessing

In the first stage of the experiment, there were two groups: sertraline and placebo. First, one hundred forty-six samples, who were given sertraline, were chosen for the sertraline group. Next, we eliminated 26 samples with the most missing values in the data. The study's target, labelled as '*w8_responder*' in the data, shows whether or not the participants were diagnosed with depression. As the target column was already binary, it required no preprocessing. The target feature was separated into its data frame before the rest of the features were subjected to preprocessing. The $get_dummies()$ function in Pandas uses One-hot-encoding to convert category labels into numerical values, as discussed in Section 2.1. In addition, pandas' *drop_duplicates*() technique aids in analysing duplicate values and removing duplicate features from a data frame. The constant features were then deleted using Variance Threshold, as mentioned in section 2, using a 0.1 threshold. Next, outliers were identified and removed. Finally, as discussed in section 2, the next step removed correlated features with greater than 90 percent correlation. The anatomical dataset also employed the same patients and the same steps as the clinical dataset.

3.5 Baseline Model

The PCA analysis on the splits was carried out with the help of the nipalsPCA function, which is part of the Hoggorm package. The purpose of the exploratory analysis was to review the data to see if there were any interesting patterns or patterns of systematic change. Using the dataset of RENT-selected features from each split, PCA was performed for each split. Visualizations of the scores, loadings and correlation loadings for each of the various blocks were created using the Hoggormplot package to seek outliers and feature clustering. Finally, the analysis was performed on the training data to check the systematic variation and patterns both before and after splitting and comparable to one another hyperparameter search.

Table 3.3 shows a list of classifiers together with their hyperparameters and a brief description of each.

Model	**Hyperparameter**	**Description**
Logistic Regression	C	Inverse of regularisation strength $(float)$
	solver	Algorithm for optimization problem $['lbfgs','liblinear']$
Random Forest	criterion	Function to measure the quality of a split $['gini','entropy']$
	max depth	Maximum depth of the tree (int)
	max features	Maximum number of features to consider $['auto','sqrt','log']$
	n estimators	No. of trees in the forest (int)
SVM	C	Regularisation parameter $(float)$
	kernel	Kernel type $['linear','rbf']$
	gamma	Kernel coefficient $(float)$

Table 3.3: The list of classifiers together with their hyperparameters and a brief description of each.

A baseline is a point of reference that helps put the results of trained models

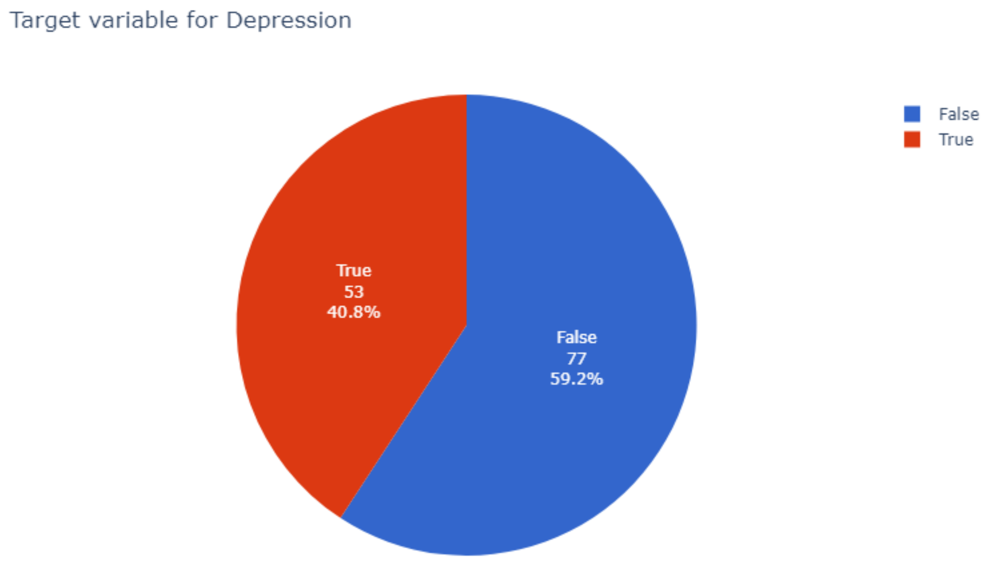

Figure 3.2: The distribution of the target column 'w8_responder'.

into context. The models are accessible as part of the Scikit-Learn package. The hyperparameters of the models can be altered to improve the models' overall performance. Grid searches were done on each model to discover the optimum hyperparameters using the scikit-learn function GridSearchCV. The scikit-learn function GridSearchCV was used, which did a five-fold cross-validated grid search over a defined grid of parameters to find the combination of hyperparameters that made the best predictions on the training data. The estimator, parameter grid, score, and the number of folds utilized in the cross-validation were all inputs for the GridSearchCV program. The baseline and feature selected by RENT models were grid searched to find the optimal parameters.

There were 130 targets, 53 of which had the class label 1, while the other 77 had the class label 0 as shown in figure 3.2. The scoring measure was the Matthews correlation coefficient (MCC), as detailed in Section 2. The parameters that produced the best results in the grid search were used to initialize the various classifiers. We used Scikit-Learn's RepeatedStratifiedKFold function as cross validation for evaluation. Finally, we assessed the effectiveness of each classifier by taking the mean of the scores obtained after the classification process.

3.6 Feature selection using RENT

The RENT method was applied as a feature selection algorithm on every split. After several rounds of trial and error, the hyperparameter was optimized to its optimal value. The C and L1 hyperparameters were adjusted to the values of 0.1 and 0.7, respectively. The values for tau, denoted by $\tau 1$, $\tau 2$, and $\tau 3$, were each initialized to a value of 0.4, 0.4, and 0.975, respectively. The characteristics picked on each split are preserved in their unique data frame for use in the future.

The remaining fourth fold, with only the selected features, was tested with different classifiers. GridsearchCV was used to find the best hyperparameter for each classifier. The scores are the mean of each classifier in each fold. MCC score was used as the performance metric to compare the prediction performance of the classifiers.

Chapter 4

Results

This research aimed to identify the radiomics features and brain regions that were predictive of class labels. The dataset was then divided into a training set and a test set. First, we performed the Repeated Elastic Net Technique (RENT) technique on each split to discover the ideal feature set, which was accomplished by training an ensemble of one hundred elastic net regularized models. And then selecting features based on the weight distributions of features across all models. Next, the model made calculations to determine the average level of performance across all models as well as the frequency the model chose a feature. Following that, logistic regression was carried out on the test set to validate the effectiveness of the RENT model. In addition, a validation study was conducted to determine whether the RENT model performed significantly better than a random model.

4.1 Data preprocessing

The dataset consisted of 296 samples with 10160 columns consisting of radiomics and demographic features and a target column. The total samples were reduced to 146 by choosing just the sertraline group, as shown in fig. Out of 146 samples, 26 had the most missing values. Following their exclusion, there were a total of 130 samples. The figure 3.2 depicts the target class distribution, with 77 people classified as not depressed and 53 as depressed. As indicated in figure 4.6, the patients were diagnosed in four different centres. Most patients were between the ages of 20 and 30, followed by 30 to 40 years and 50-60 years. The figure 4.2 and 4.1 display total number of patients in that age group. The table 4.1 reflects the dimension of the data at every step of preprocessing. Between session 1 and 2, session 1 features were only selected from anatomical data for the study.

Pre-process	**dwi dataset (Clinical)**	**anat dataset (Anatomical)**
Initial Dimension	(296,10175)	(296,20340)
Sertraline Group and Session 1 (free from missing values)	(130,10166)	(146, 10171)
Only radiomics features	(130,10165)	(130, 10171)
After duplicate features removal	(130, 4767)	(130, 10165)
After constant features removal	(130, 2346)	(130, 5195)
After outliers removal	(130, 393)	(130, 1487)
After correlated features removal	(130, 344)	(130, 729)

Table 4.1: Initial dimension of the data and change in dimension after every preprocessing process.

The figure 4.3, 4.4 and 4.5 shows the pie chart of the columns ‘race’, ‘hispanic’, and ‘gender’ in the data recpectively.

The dataset and target were split. The columns ‘w8_score_17’ and ‘w16_score_17’ have been removed since doctors utilized them to determine the ultimate goal which was highly correlated to the target. In addition, ‘Stage2TX’ and ‘Stage1TX’ columns were removed since they specified the study strategy for separating placebo and sertraline groups. After removing duplicates, constants, outliers, and correlated columns, data preprocessing reduced the data to 130 rows by 344 for clinical data whereas 130 rows by 729 columns for anatomical data. The MR images were taken in four different centers as shown in figure 4.6.

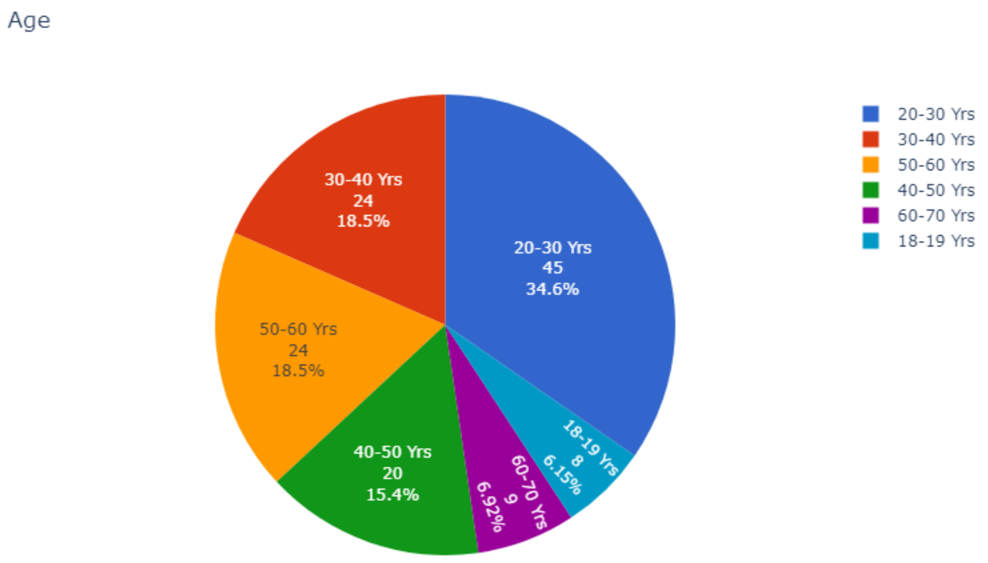

Figure 4.1: Distribution of age group in the data.

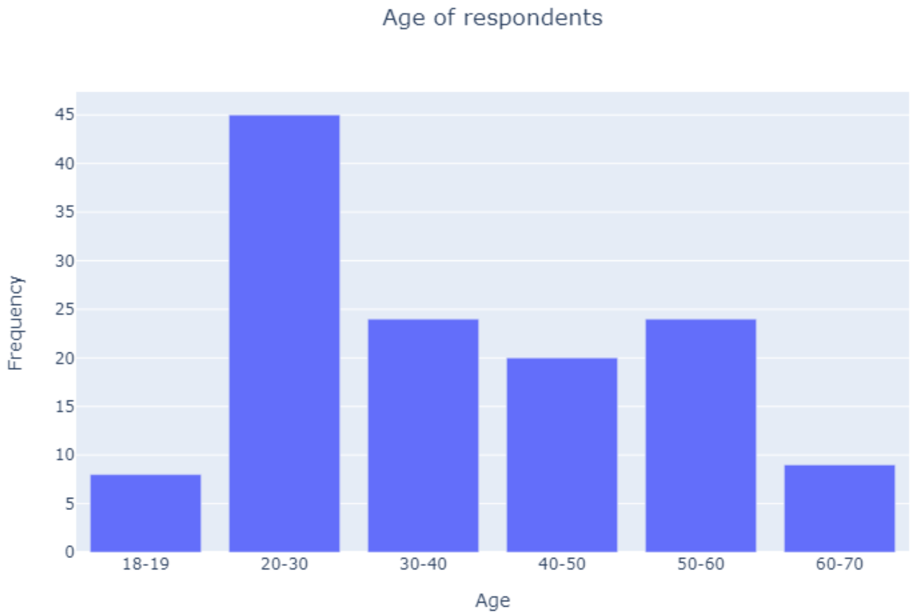

Figure 4.2: Bar plot of the distribution of age group in the data.

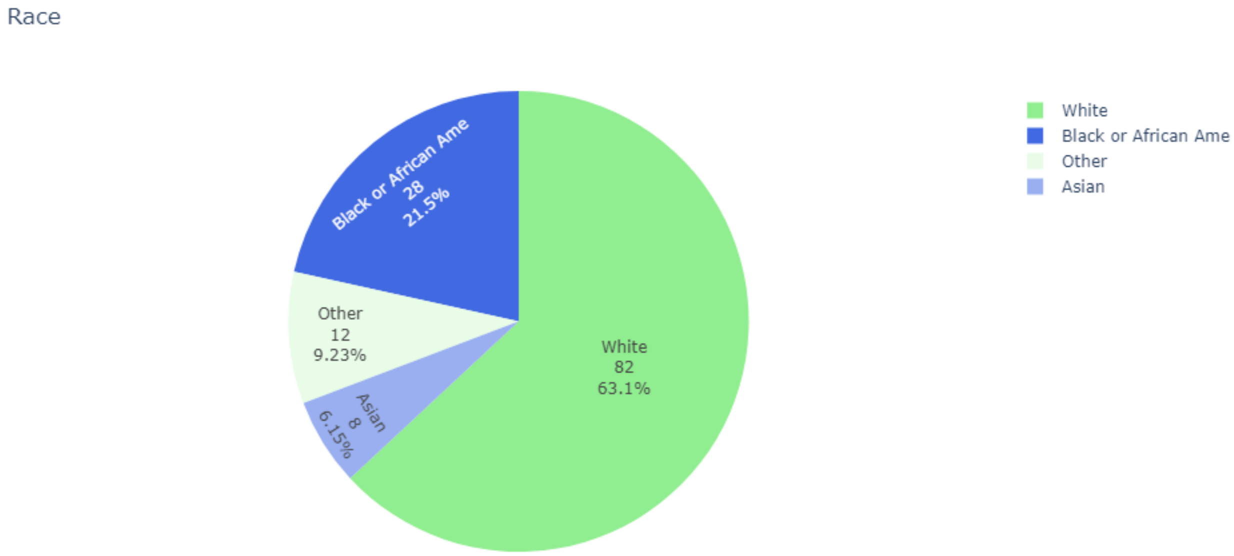

Figure 4.3: Distribution of values in column 'race' in the data. Four different values wer used in race feature column.

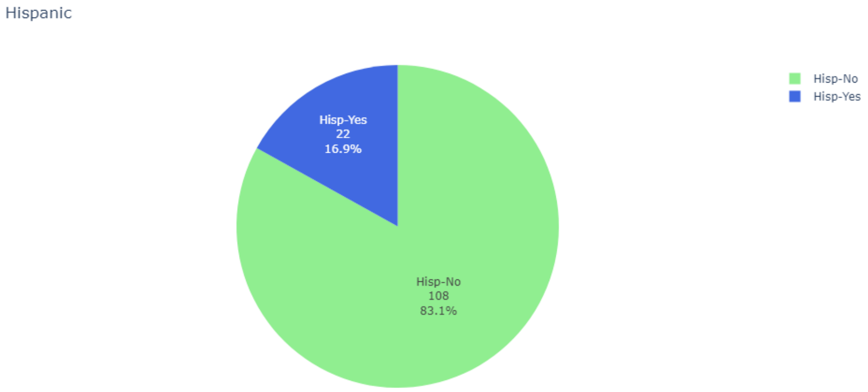

Figure 4.4: Distribution of variables in 'hispanic' columns in the data.

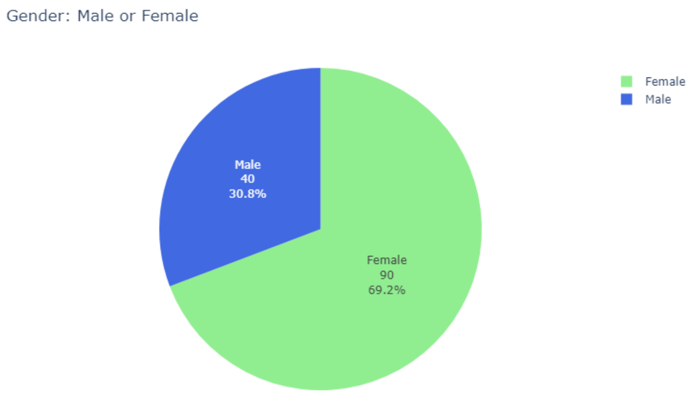

Figure 4.5: Bar plot of the distribution of gender, here male or female, in the data.

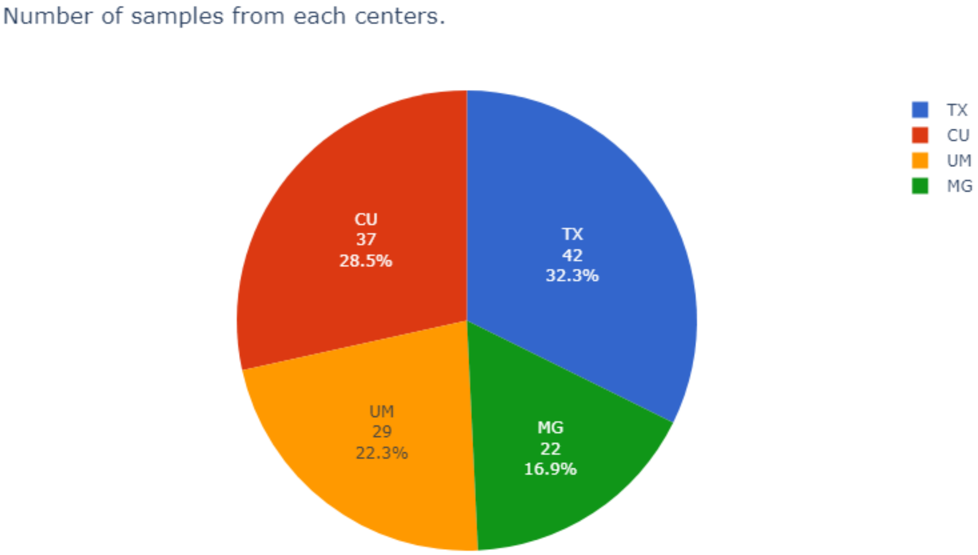

Figure 4.6: The number of samples from each of the centers

4.2 DWI dataset

4.2.1 PCA

Figures showing the scores, loadings, and explained variance are shown as results from an analysis performed with principal component analysis (PCA) on whole data after data preprocessing. The plot of the PCA scores can be found in the figure 4.16.

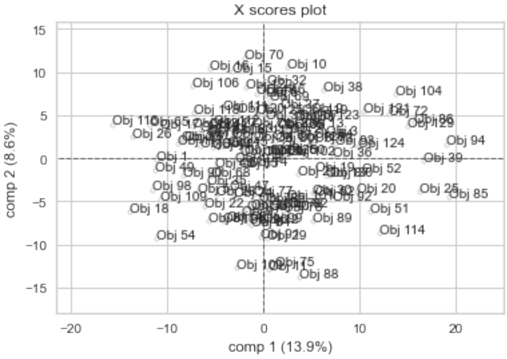

Figure 4.7: PCA scores of clinical data. The first principal component is along the horizontal axis, while the second principal component is along the vertical axis. The proportion of the explained variance each component contributes to is indicated in parentheses after the components' respective axes.

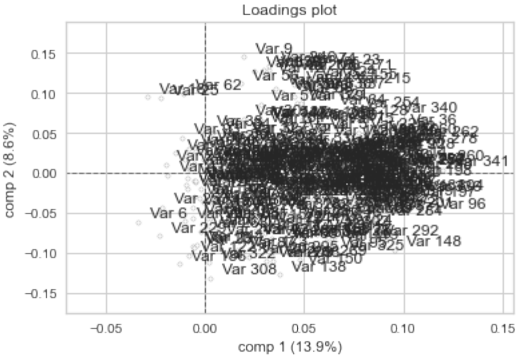

Figure 4.8: The figure displays loading plot from the PCA analysis.

The figure does not illustrate any unique clustering of the data into two groups, nor does it show any extreme outliers. Following each component's corresponding axis is a set of parentheses containing an indication of the percentage of the explained variance to the component contributes. Finally, a figure 4.17 of the loadings illustrates the degree to which each characteristic exerts its impact on the components. For example, it is clear from the figure that the variable 6 referred to as 'ses-1_Left-Lateral-Ventricle_original-gldm-DependenceEntropy' has a greater propensity to be in the initial component. The figure 4.20 displays explained variance of the principal components. The blue line shows the calibrated variance, while the red line represents the validated variance. The figure 4.10 shows the cumulative explained variance as the principal components are included.

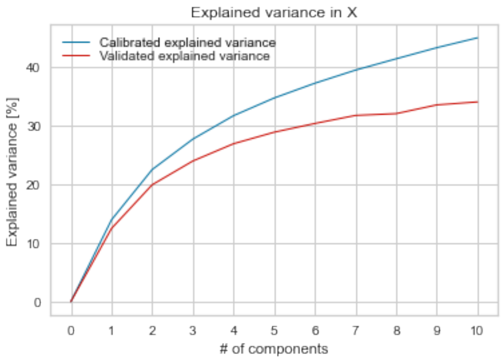

Figure 4.9: Explained variance of the principal components.The blue line shows the calibrated variance, while the red line represents the validated variance.

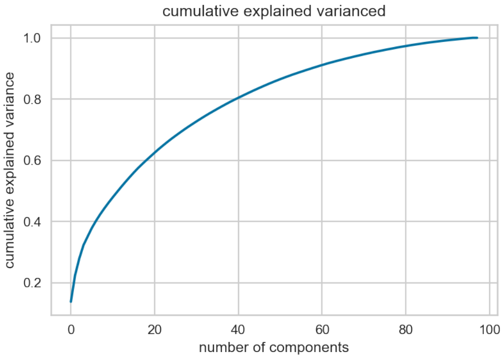

Figure 4.10: The cumulative explained variance from the PCA analysis.

4.2.2 PLSR

After data preparation, analysis using partial least squares regression (PSLR) was carried out on the entire dataset. As a consequence of this analysis, figures displaying the scores, loadings, and explained variance were generated and shown. The results of PLSR are not the same as those of PCA since PLSR is a supervised approach and incorporates the target feature into the analysis. The plot of the PCA scores can be found in the figure 4.16. The first number reflects the explained variance in x, the features in the data, while the second indicates the explained variance in y, which is the target variable.

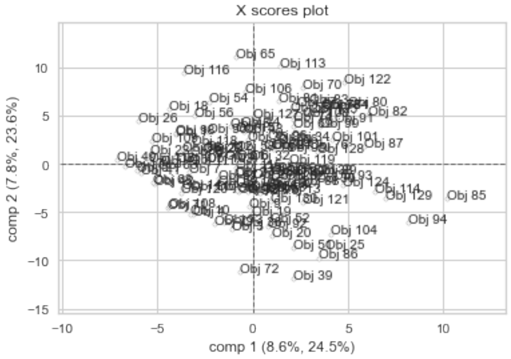

Figure 4.11: PLSR scores of clinical data. The first principal component is along the horizontal axis, while the second principal component is along the vertical axis. The proportion of the explained variance each component contributes to is indicated in parentheses after the components' respective axes for both x and y.

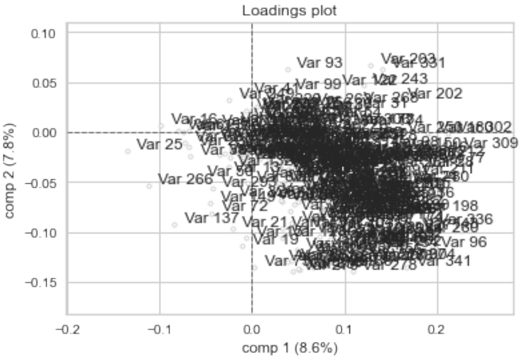

Figure 4.12: Loading plot from the PLSR analysis on clinical data.

No distinct grouping of the data into two groups is depicted in the figure, and no extreme outliers are shown either. Following each component's corresponding axis is a set of parentheses containing an indication of the percentage of the explained variance to the component contributes. Finally, the figure 4.17 of the loadings illustrates the degree to which each characteristic exerts its impact on the components. For example, it is clear from the figure that the variable 25 referred to as 'ses-1_Left-Putamen_original-glszm-LargeAreaEmphasis' has a greater propensity to be the initial component.

4.2.3 Classification modelling and evaluation

Baseline

After preprocessing, the data was split with RSKF with four folds and three repetitions. GridsearchCV was used to determine the best hyperparameters and performance for each split employing Logistic Regression, Support Vector Machine, and Random Forest classifiers. The table 4.2 shows the model with the greatest MCC score in each split with repeated stratified k-fold cross validation.

Overall, Logistic Regression outperformed all other classifiers when computing the average performance on all splits. The table 4.3 shows the model with the greatest MCC score in each split with repeated stratified k-fold cross validatoin.

RENT Hyperparameters selection

RSKF was used to partition the data into four folds and one repeat. RENT with hyperparameters $C = 0.1$ and $L1 = 0.7$ was run at each split. After trial and error, $\tau 1$,$\tau 2$, and $\tau 3$ were set at 0.4, 0.4, and 0.975, respectively. RENT chose a set of features for each split. The classifiers were trained and tested on the selected subset of features. The MCC score was employed as the primary performance metric on both RENT and classifiers. Through many rounds of stratified k-folding, an average MCC score was calculated.

The table 4.5 lists the features selected by RENT with RSKF four folds and three repetition, totalling to 12 splits with the number of times the feature was selected.

The figure 4.13 is the plot generated by RENT. The list of features and the number of times RENT picked each feature are shown on the horizontal and vertical axes. After 100-fold repetition, the one above the 0.8 horizontal line, for example, was picked more than 80% of the time. We need to know if the specified characteristics produce models that outperform random ones to verify them. The figure 4.22 displays validation plot RENT. Data on MCC scores from 100 runs in two validation experiments are shown in the blue and green graphs.There are as many random characteristics drawn in Validation Study 1 (VS1) as RENT has chosen [8]. In the

Split	**Model**	**Hyperparameter**	**Value**	**MCC Score**
1	Logistic Regression	C solver	0.0001 'lbfgs'	0.0000
2	Logistic Regression	C solver	0.1 'liblinear'	0.2714
3	Logistic Regression	C solver	0.1 'liblinear'	0.1102
4	Logistic Regression	C solver	0.1 'liblinear'	0.3181
5	Logistic Regression	C solver	0.1 'liblinear'	0.2930
6	Logistic Regression	C solver	0.1 'liblinear'	0.2556
7	Logistic Regression	C solver	0.1 'liblinear'	0.3181
8	Logistic Regression	C solver	10.0 'liblinear'	0.1479
9	Logistic Regression	criterion max_depth	'gini' 30	0.0551
10	Logistic Regression	C solver	100.0 'lbfgs'	0.2714
11	Logistic Regression	criterion max_depth	'gini' 60	0.345
12	SVM	C gamma kernel	10.0 0.01 'rbf'	0.1102

Table 4.2: The model with the greatest MCC score in each split on clinical data.

second validation study (VS2), the target labels are permuted randomly, but the sample characteristics are kept. The MCC score for the RENT model is shown in the red line [8].

Model	F1_1		F1_0		ACC		MCC		ROC	
	mean	std	mean	std	mean	std	mean	std	mean	std
LR	0.45	0.17	0.69	0.04	0.61	0.07	0.15	0.19	0.58	0.09
RF	0.26	0.18	0.69	0.08	0.57	0.11	0.03	0.29	0.51	0.11
SVM	0.03	0.07	0.74	0.01	0.59	0.02	0.01	0.15	0.5	0.03

Table 4.3: Average performance of classifiers in all split on all features of clinical dataset.

Model	F1_1		F1_0		ACC		MCC		ROC	
	mean	std	mean	std	mean	std	mean	std	mean	std
SVM	0.4158	0.1036	0.6963	0.0622	0.6048	0.0627	0.1497	0.1273	0.5657	0.062
LR	0.477	0.0856	0.635	0.0617	0.5717	0.0646	0.1164	0.1341	0.5584	0.0674
RF	0.4971	0.0886	0.6898	0.0631	0.6202	0.0624	0.2053	0.1322	0.5966	0.0618

Table 4.4: Average score of classifiers on RENT selected features set at every split on clinical dataset .

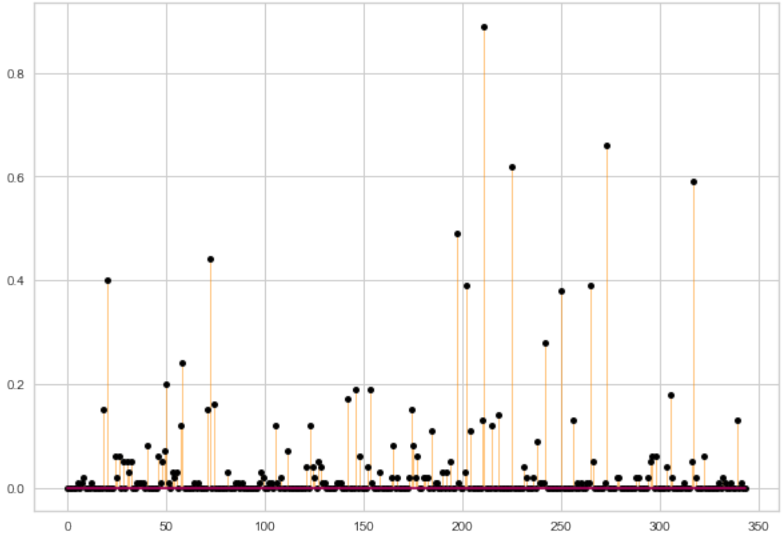

Figure 4.13: The figure displays feature selected after 100 fold repetition by RENT on clinical data. The list of features and the number of times RENT picked each feature are shown on the horizontal and vertical axes.

Feature Name	**Freq.**
ses-1_ctx-rh-rostralmiddlefrontal_original-shape-LeastAxisLength	11
ses-1_ctx-lh-superiorfrontal_original-shape-LeastAxisLength	8
ses-1_ctx-rh-lingual_original-glszm-ZoneEntropy	8
ses-1_ctx-lh-supramarginal_original-shape-Maximum3DDiameter	7
ses-1_ctx-rh-caudalmiddlefrontal_original-glszm-ZoneEntropy	6
ses-1_ctx-rh-middletemporal_original-glszm-ZoneEntropy	6
ses-1_Right-Thalamus-Proper_original-shape-Maximum2DDiameterRow	6
ses-1_ctx-rh-parsorbitalis_original-shape-Maximum2DDiameterRow	5
ses-1_Right-Inf-Lat-Vent_original-shape-MinorAxisLength	5
ses-1_Right-Inf-Lat-Vent_original-gldm-LargeDependenceEmphasis	5
ses-1_3rd-Ventricle_original-shape-Maximum2DDiameterRow	3
ses-1_ctx-rh-isthmuscingulate_original-glszm-ZoneEntropy	3
ses-1_ctx-rh-entorhinal_original-shape-LeastAxisLength	3
ses-1_ctx-rh-fusiform_original-shape-Maximum2DDiameterRow	3
ses-1_ctx-lh-precentral_original-shape-LeastAxisLength	2
ses-1_ctx-rh-superiorparietal_original-firstorder-Maximum	2
ses-1_ctx-lh-parstriangularis_original-glszm-ZoneVariance	2
ses-1_ctx-rh-superiorfrontal_original-glszm-ZoneVariance	2
ses-1_ctx-lh-pericalcarine_original-shape-Maximum2DDiameterColumn	2
ses-1_Left-Thalamus-Proper_original-shape-LeastAxisLength	2
ses-1_Right-Lateral-Ventricle_original-gldm-LargeDependenceEmphasis	2
ses-1_4th-Ventricle_original-gldm-LargeDependenceEmphasis	2
ses-1_ctx-lh-parstriangularis_original-glszm-ZoneEntropy	2
ses-1_ctx-lh-fusiform_original-shape-Maximum2DDiameterSlice	2
ses-1_Right-Hippocampus_original-glszm-LargeAreaEmphasis	2
ses-1_ctx-lh-caudalmiddlefrontal_original-shape-Maximum2DDiameterRow	1
ses-1_ctx-lh-parahippocampal_original-shape-MajorAxisLength	1
ses-1_ctx-lh-superiortemporal_original-gldm-LargeDependenceEmphasis	1
ses-1_Right-Hippocampus_original-shape-MinorAxisLength	1
ses-1_Left-VentralDC_original-firstorder-Energy	1
ses-1_Right-Cerebellum-Cortex_original-shape-MajorAxisLength	1
ses-1_ctx-rh-pericalcarine_original-gldm-LargeDependenceEmphasis	1
ses-1_Right-Thalamus-Proper_original-shape-LeastAxisLength	1
ses-1_ctx-lh-superiorparietal_original-firstorder-Maximum	1
ses-1_ctx-rh-inferiorparietal_original-shape-LeastAxisLength	1

Table 4.5: List of RENT selected features with their selection frequency from RSKF split of 12 splits on clinical data.

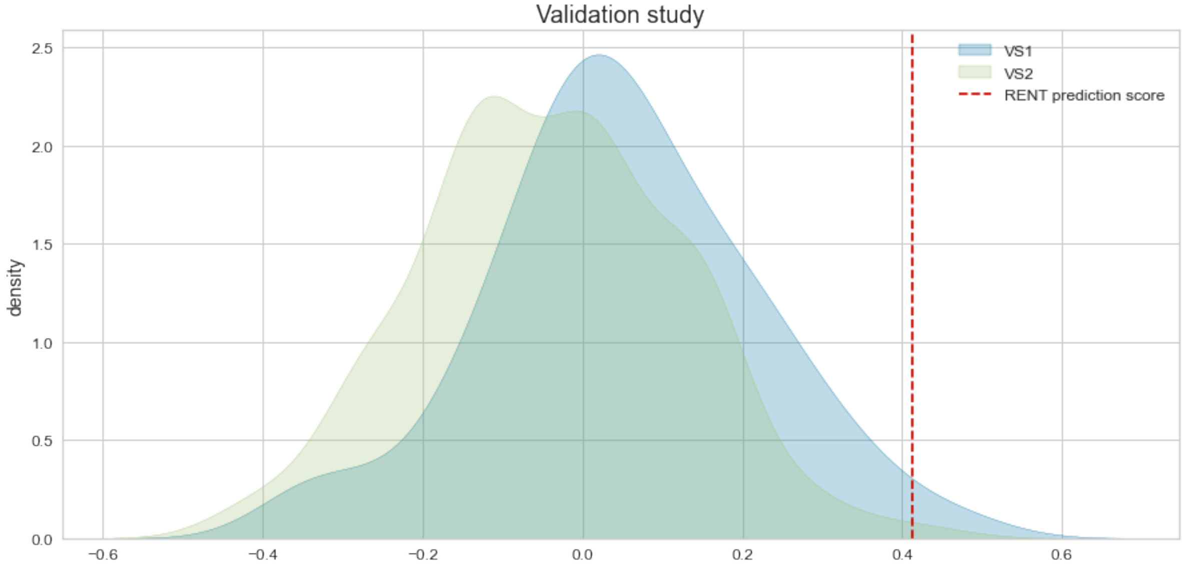

Figure 4.14: Validation plot RENT. Data on MCC scores from 100 runs in two validation experiments are shown in the blue and green graphs.There are as many random characteristics drawn in Validation Study 1 (VS1) as RENT has chosen. In the second validation study (VS2), the target labels are permuted randomly, but the sample characteristics are kept. The MCC score for the RENT model is shown in the red line.

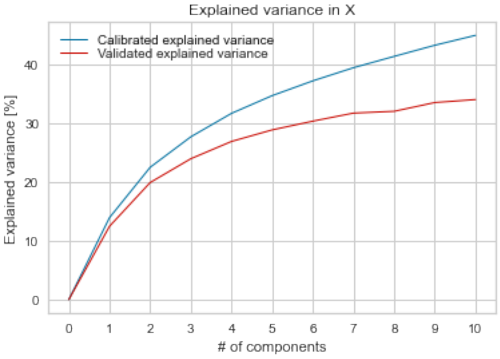

Figure 4.15: Explained variance of the principal components of anatomical data. The blue line shows the calibrated variance, while the red line represents the validated variance for x.

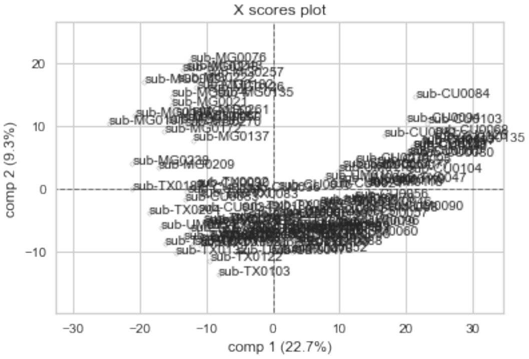

Figure 4.16: PCA scores of anatomical data. The first principal component is along the horizontal axis, while the second principal component is along the vertical axis. The proportion of the explained variance each component contributes to is indicated in parentheses after the components' respective axes.

4.3 ANAT dataset

4.3.1 PCA

Figures showing the scores, loadings, and explained variance are shown as results from an analysis performed with principal component analysis (PCA) on whole data after data preprocessing. The plot of the PCA scores can be found in the figure 4.16. The figure 4.20 displays explained variance of the principal components. The blue line shows the calibrated variance, while the red line represents the validated variance for x, the columns in the data.

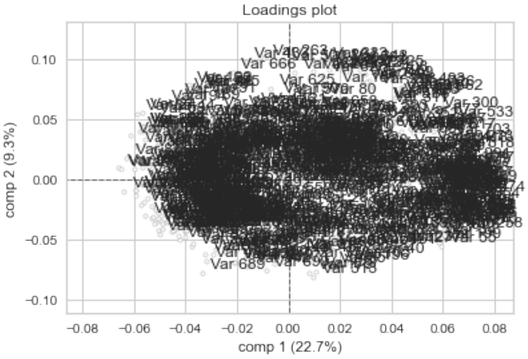

Figure 4.17: Loading plot from the PCA analysis of anatomical data.

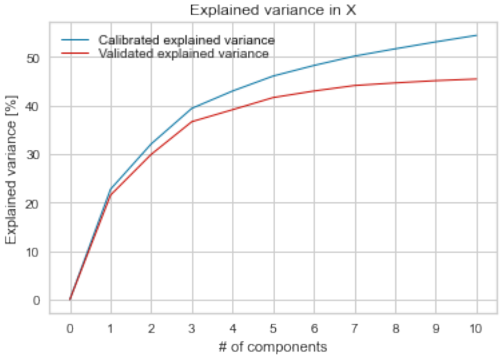

Figure 4.18: Explained variance of the principal components of anatomical data.The blue line shows the calibrated variance, while the red line represents the validated variance.

The figure does not illustrate any unique clustering of the data into two groups, nor does it show any extreme outliers. Additionally, the plot does not display any extreme outliers. Following each component's corresponding axis is a set of parentheses containing an indication of the percentage of the explained variance to the component contributes.

Finally, a figure 4.17 of the loadings illustrates the degree to which each characteristic exerts its impact on the components. For example, it is clear from the figure that the variable referred to as 'ses-1_Left-Lateral-Ventricle_original-gldm-DependenceEntropy' has a greater propensity to be the initial component.

The figure 4.20 displays explained variance of the principal components.The blue line shows the calibrated variance, while the red line represents the validated variance.

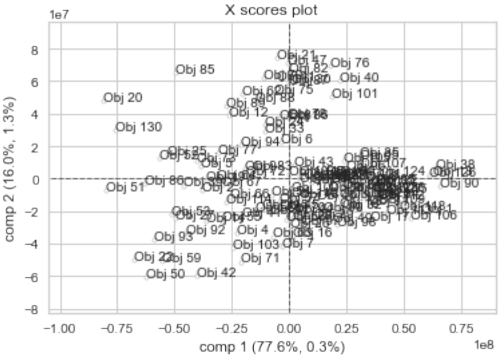

Figure 4.19: PLSR scores of anatomical data. The first principal component is along the horizontal axis, while the second principal component is along the vertical axis. The proportion of the explained variance each component contributes to is indicated in parentheses after the components' respective axes for both x and y.

4.3.2 PLSR

After data preparation, analysis using partial least squares regression (PSLR) was carried out on the entire dataset. As a consequence of this analysis, figures displaying the scores, loadings, and explained variance were generated and shown. The plot of the PCA scores can be found in the figure 4.19. the first number reflects the explained variance in x, while the second number indicates the explained variance in y.

No distinct grouping of the data into two groups is depicted in the figure 4.19, and no extreme outliers are shown either. Following each component's corresponding axis is a set of parentheses containing an indication of the percentage of the explained variance to the component contributes.

The figure 4.20 displays explained variance of the principal components.The blue line shows the calibrated variance, while the red line represents the validated variance.

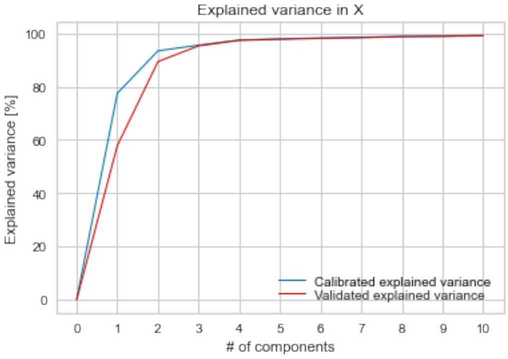

Figure 4.20: The figure displays explained variance of the principal components.The blue line shows the calibrated variance, while the red line represents the validated variance for x.

4.3.3 Classification modelling and evaluation

Baseline

The processes done in clinical data are comparable to steps in clinical data. RSKF was used to divide the data into four groups with three repeats after preprocessing for the data splits. For each split, GridsearchCV employed Logistic Regression, Support Vector Machine, and Random Forest classifiers to identify the optimum hyperparameters and performance. The table 4.6 The model with the greatest MCC score in each split with repeated stratified k-fold is shown.

Overall, Logistic Regression outperformed all other classifiers when computing the average performance on all splits.

Model	F1_1		F1_0		ACC		MCC		AUC	
	mean	std	mean	std	mean	std	mean	std	mean	std
LR	0.37	0.11	0.67	0.04	0.57	0.05	0.06	0.12	0.53	0.06
RF	0.27	0.12	0.69	0.04	0.56	0.06	0.02	0.16	0.51	0.06
SVM	0.06	0.09	0.72	0.03	0.57	0.03	-0.05	0.13	0.49	0.03

Table 4.6: Average performance of classifiers in all split on all features of anat dataset.

RENT Hyperparameters selection

RSKF was used to partition the data into four folds and one repeat. RENT with hyperparameters C = 0.1 and L1 = 0.7 was run at each split. After trial and error, $\tau1$,$\tau2$, and $\tau3$ were set at 0.4, 0.4, and 0.975, respectively. RENT chose a set of features for each split. The classifiers were trained and tested on the selected subset of features. The MCC score was employed as the primary performance metric on both RENT and classifiers. Through many rounds of stratified k-folding, an average MCC score was calculated.

The table 4.9 shows the score of each classification model with the highest MCC score in each split with repeated stratified k-fold and RENT selected features on anatomical data. The tables 4.7 and 4.8 lists the features selected by RENT with RSKF four folds and three repetition, totalling to 12 splits with the number of times the feature was selected.

The table 4.7 and 4.8 has list of features selected with the number of times the feature was selected by RENT after RSKF 12 splits on anatomical dataset.

From the table 4.9, we can see that performance score of different classifiers on RENT selected features for anatomical data. The figure 4.21 shows feature selected after 100 fold repetition by RENT on anatomical data. The list of features and the number of times RENT picked each feature are shown on the horizontal and vertical axes. The figure 4.22 displays the validation plot RENT on anatomical dataset. The figure displays validation plot RENT. Data on MCC scores from 100 runs in two validation experiments are shown in the blue and green graphs.There are as many random characteristics drawn in Validation Study 1 (VS1) as RENT has chosen. In the second validation study (VS2), the target labels are permuted randomly, but the sample characteristics are kept. The MCC score for the RENT model is shown in the red line.

The table 4.10 has the list of features selected at least 50% times of the time by RENT on both clinical and anatomical data. Seven features from clinical data and four features from anatomical data were selected.

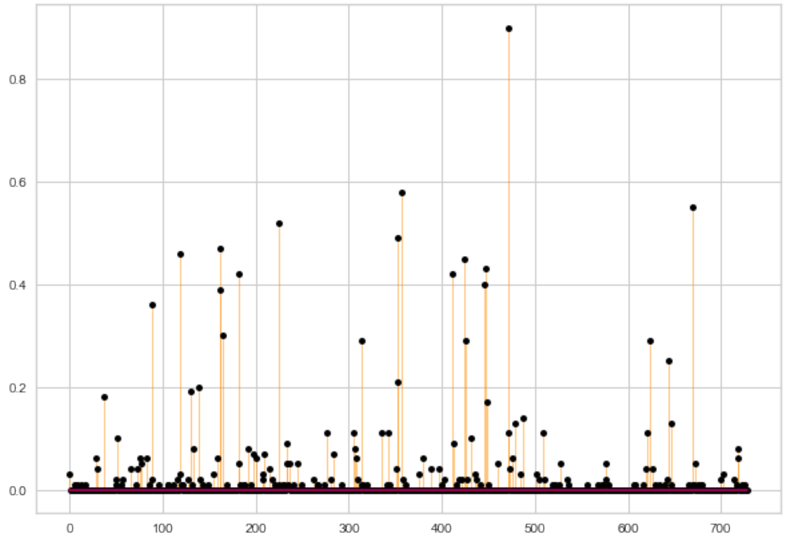

Figure 4.21: Feature selected after 100 fold repetition by RENT on anatomical data. The list of features and the number of times RENT picked each feature are shown on the horizontal and vertical axes.

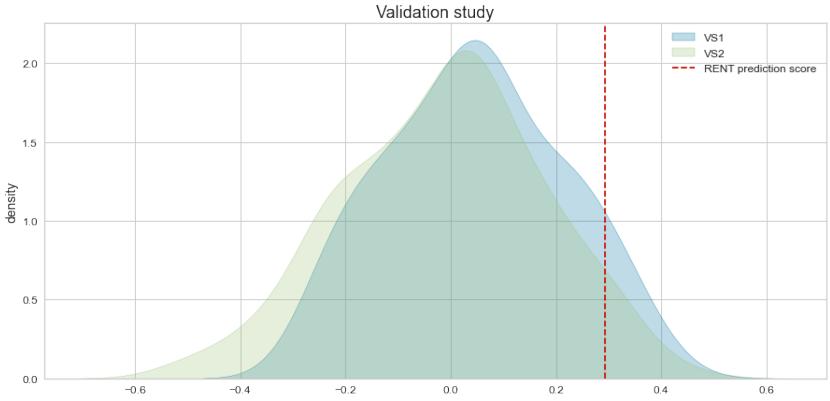

Figure 4.22: The validation plot RENT on anatomical dataset. The figure displays validation plot RENT. Data on MCC scores from 100 runs in two validation experiments are shown in the blue and green graphs.There are as many random characteristics drawn in Validation Study 1 (VS1) as RENT has chosen. In the second validation study (VS2), the target labels are permuted randomly, but the sample characteristics are kept. The MCC score for the RENT model is shown in the red line.

Feature Name	**Freq.**
ses-1_ctx-rh-rostralmiddlefrontal_original-shape-LeastAxisLength	12
ses-1_Right-Inf-Lat-Vent_original-firstorder-Energy	10
ses-1_ctx-lh-posteriorcingulate_original-glszm-ZoneEntropy	8
ses-1_ctx-lh-supramarginal_original-shape-Maximum3DDiameter	7
ses-1_Right-Thalamus-Proper_original-shape-Maximum2DDiameterRow	5
ses-1_ctx-lh-superiorfrontal_original-shape-LeastAxisLength	5
ses-1_Right-Inf-Lat-Vent_original-shape-MinorAxisLength	4
ses-1_ctx-rh-parsorbitalis_original-shape-Maximum2DDiameterRow	4
ses-1_CSF_original-glszm-ZoneEntropy	3
ses-1_ctx-rh-entorhinal_original-shape-LeastAxisLength	3
ses-1_ctx-lh-middletemporal_original-gldm-DependenceVariance	3
ses-1_Left-Caudate_original-glszm-HighGrayLevelZoneEmphasis	3
ses-1_ctx-rh-fusiform_original-shape-Maximum2DDiameterRow	3
ses-1_ctx-lh-superiorfrontal_original-gldm-LargeDependenceHighGrayLevelEmphasis	3
ses-1_ctx-lh-middletemporal_original-glszm-ZoneEntropy	3
ses-1_Right-Amygdala_original-firstorder-Minimum	3
ses-1_4th-Ventricle_original-gldm-LargeDependenceHighGrayLevelEmphasis	3
ses-1_ctx-lh-precuneus_original-glcm-Autocorrelation	3

Table 4.7: List of RENT selected features with their selection frequency from RSKF split of 12 splits on anatomical data.

Feature Name	**Freq.**
ses-1_4th-Ventricle_original-firstorder-Energy	2
ses-1_ctx-lh-inferiorparietal_original-glszm-ZoneEntropy	2
ses-1_ctx-lh-rostralanteriorcingulate_original-gldm-DependenceVariance	2
ses-1_Right-Inf-Lat-Vent_original-glcm-Autocorrelation	2
ses-1_Right-Amygdala_original-glcm-Autocorrelation	1
ses-1_Left-Accumbens-area_original-glszm-ZoneEntropy	1
ses-1_ctx-rh-posteriorcingulate_original-glszm-ZoneVariance	1
ses-1_ctx-rh-transversetemporal_original-glszm-ZoneEntropy	1
ses-1_ctx-lh-precuneus_original-gldm-LargeDependenceHighGrayLevelEmphasis	1
ses-1_ctx-rh-transversetemporal_original-ngtdm-Complexity	1
ses-1_ctx-lh-cuneus_original-gldm-DependenceVariance	1
ses-1_ctx-lh-postcentral_original-firstorder-Maximum	1
ses-1_ctx-lh-superiorparietal_original-glszm-ZoneEntropy	1
ses-1_Right-Hippocampus_original-shape-MinorAxisLength	1
ses-1_3rd-Ventricle_original-shape-Maximum2DDiameterRow	1
ses-1_ctx-lh-parsorbitalis_original-ngtdm-Complexity	1
ses-1_ctx-lh-rostralanteriorcingulate_original-glszm-HighGrayLevelZoneEmphasis	1
ses-1_ctx-lh-rostralanteriorcingulate_original-glszm-SmallAreaHighGrayLevelEmphasis	1
ses-1_ctx-lh-caudalmiddlefrontal_original-shape-Maximum2DDiameterRow	1
ses-1_Left-Accumbens-area_original-glszm-SmallAreaHighGrayLevelEmphasis	1
ses-1_ctx-lh-parahippocampal_original-firstorder-Kurtosis	1
ses-1_Brain-Stem_original-glszm-GrayLevelNonUniformity	1
ses-1_ctx-lh-superiorfrontal_original-glcm-Autocorrelation	1
ses-1_Left-Thalamus-Proper_original-shape-LeastAxisLength	1
ses-1_CSF_original-glszm-GrayLevelNonUniformity	1
ses-1_ctx-rh-caudalmiddlefrontal_original-glszm-ZoneEntropy	1

Table 4.8: List of RENT selected features with their selection frequency from RSKF split of 12 splits on anatomical data.

Model	F1_1		F1_0		ACC		MCC		AUC	
	mean	std	mean	std	mean	std	mean	std	mean	std
SVM	0.4746	0.1425	0.6852	0.0763	0.61	0.0898	0.1744	0.195	0.5846	0.0941
LR	0.4503	0.1033	0.6401	0.0617	0.5691	0.0601	0.1019	0.1282	0.5509	0.0641
RF	0.4334	0.0959	0.6575	0.0572	0.5767	0.0555	0.1039	0.1187	0.5504	0.0567

Table 4.9: Average score of classifiers on RENT selected features set at every split on anat dataset .

Feature Name	**Freq.**
Clinical data	
ses-1_ctx-rh-rostralmiddlefrontal_original-shape-LeastAxisLength	11
ses-1_ctx-lh-superiorfrontal_original-shape-LeastAxisLength	8
ses-1_ctx-rh-lingual_original-glszm-ZoneEntropy	8
ses-1_ctx-lh-supramarginal_original-shape-Maximum3DDiameter	7
ses-1_ctx-rh-caudalmiddlefrontal_original-glszm-ZoneEntropy	6
ses-1_ctx-rh-middletemporal_original-glszm-ZoneEntropy	6
ses-1_Right-Thalamus-Proper_original-shape-Maximum2DDiameterRow	6
Anatomical data	
ses-1_ctx-rh-rostralmiddlefrontal_original-shape-LeastAxisLength	12
ses-1_Right-Inf-Lat-Vent_original-firstorder-Energy	10
ses-1_ctx-lh-posteriorcingulate_original-glszm-ZoneEntropy	8
ses-1_ctx-lh-supramarginal_original-shape-Maximum3DDiameter	7

Table 4.10: List of RENT selected features selected at least 50% of the times with their selection frequency from RSKF split of 12 splits on clinical and anatomical data.

Chapter 5

Discussion

5.1 Data

The research uses two different datasets: the dwi-dataset and the anatomical-dataset. Patients were divided into two categories: those who were diagnosed with depression after the experiment and those who were not. The class balance of the variable in question showed that class 0 contributed 40.8% of the total, while class 1 contributed 59.2%. Utilizing all of the readily available data with no feature set gaps was a primary concern. Because there were missing values in both datasets, the samples with the fewest missing values were chosen for the analysis. Additionally, the same samples (130 in total) were chosen from both datasets for further examination; therefore, the total number of samples used in the study was 130. Patients who had received sertraline during the initial phase of their treatment were selected for this investigation. The study was not repeated using stage 2 for each patient since there was a constraint on the amount of time available; nonetheless, doing such a test would be interesting in further analyses.

5.2 Preprocessing

Before the data was preprocessed, the target column 'w8_responder' was isolated. The data were examined for the presence of associated features, duplicate features, and constant features. A feature can't deliver helpful information if most of the samples it is applied to have the same value spreading. As a result, it is essential to eliminate the continuous traits. It was decided to exclude features that correlated greater than 90 per cent.

5.3 Analysis

Exploratory analysis can be carried out with both principal component analysis and partial least squares regression approaches. Principal component analysis (PCA), on the other hand, is an unsupervised method. Partial least squares regression (PLSR), on the other hand, is a supervised method that incorporates the goal into the study. The figure 4.16 does not emphasize any severe outliers or demonstrate any prominent grouping of the data into two groups for each data set. Also, the figure does not indicate any extreme outliers. Finally, the figure 4.20 demonstrates that no one point can be used to describe the primary components in a way that adequately explains the data.

5.4 Classification

The performance of multiple different classifiers was evaluated during the analysis using three distinct classifiers, LR, SVM and Random Forest, which were trained and tested separately. Each classifier was utilized throughout the training and testing phases on the baseline and the RENT processes. By adjusting a parameter in the kernel, the support vector machine (SVM) can handle both linear and non-linear problems, and this capability was tested in both cases. After doing a grid search, it was discovered that the 'liblinear' solver produced the best results for the baseline data regarding Logistic Regression.

5.5 RENT

The Repeated Elastic Net Technique performed well with feature selection. Every possible combination of C and L1 values was examined on the data sets because the RENT-selected values for C and L1 did not provide satisfactory performance. After analysing the performance, it was decided that 0.1 would be a good number for C, while 0.7 would be an appropriate value for L1. After analysing each possible combination on the datasets, the values 0.4, 0.4, and 0.975 were selected as the optimal values for $\tau 1$, $\tau 2$, and $\tau 3$, respectively. The final call on the cutoff parameter was reached after consultation with the advisors who were engaged in developing this thesis. RENT effectively cut the number of features down to fewer features. RENT selected thirty-five features from the clinical data and forty-two features from the anatomical data. It was unclear if all of the features chosen by RENT ought to be included or whether a limit needs to be imposed on the number of times a feature could be chosen before deciding whether or not a feature needs to be included. A comparison was made with the prior research and their features.

For feature selection, RENT uses Logistic Regression in conjunction with elastic net regularisation. The performance of the classifiers was tested by using a repeated stratified k-fold, and RENT was used to choose the features used in the evaluation. When the characteristics suggested by RENT were included, performance increased across the board for all classifiers. As seen in the table 5.1 and 5.2 there is significant improvement in the performance of the classifiers. Random Forest performed the best among other classifiers in clinical data while SVM performed better in anatomical data.

Classifier	**All Features**	**Selected features**
RF	0.03	0.21
LR	0.15	0.12
SVM	0.01	0.15

Table 5.1: Average MCC score with for all the features in the test data and using the features selected by RENT on dwi data.

Classifier	**All Features**	**Selected features**
RF	0.06	0.10
LR	0.06	0.10
SVM	-0.05	0.17

Table 5.2: Average MCC score with for all the features in the test data and using the features selected by RENT on anatomical dataset.

The MCC score has the maximum value when evaluated using a random forest, which is valid for both datasets. After using RENT for feature selection, both SVM and Random Forest performance saw a substantial improvement. LR, on the other hand, does not exhibit any signs of performance enhancement. However, the slight decrease in performance is tolerable because of the simplification of the characteristics and the improvement in their interpretability.

The RENT algorithm chose seven features from the DWI data and four features from the anatomical data, each chosen at least 50% of the time as shown in the table 4.10. The table 4.5, 4.7 and 4.8 include information on the number of times each feature was chosen. RENT considered characteristics based on their textures as well as their shapes.

According to the specified characteristics, the shape of the rostral middle frontal gyrus may be an important brain area to consider when attempting to diagnose someone with depression. Because the attributes chosen were from the frontal gyrus, it may be deduced that the characteristics found in the frontal areas have the potential to act as biomarkers. The areas of the brain known as the posterior cingulate cortex, the rostral middle frontal cortex, and the middle temporal cortex have been suggested as potential biomarker regions by RENT selected features. Several studies concluded that multiple possible biomarkers in frontolimbic regions, such as the prefrontal cortex, anterior cingulate cortex, hippocampus, amygdala, and insula, most frequently influenced response outcome [8]. However, the strength and direction of the biomarker's association with clinical response varied, most likely due to differences in the studies themselves [8]. Because of this, it is necessary to duplicate and validate brain biomarkers several times in large independent sets of samples before they can be employed for medical purposes [13].

Two separate validation experiments were carried out to check and make certain that the characteristics chosen by RENT were crucial to the high level of the model's performance [8]. The response target was permuted in validation study 2 (VS2), whereas random features were chosen in validation study 1 (VS1) [14]. RENT developed 100 logistic regression models and tested their accuracy by making predictions based on validation data that had not yet been observed[14]. After that, RENT compared the MCC scores obtained from these tests to predictions made by RENT based on features that it had chosen. For the purpose of making a comparison between the MCC scores, a one-sided Student's t-test was carried out. It was hypothesized that the RENT MCC would be lower than the average MCC derived from VS1 and VS2; however, this was not the case [14]. Because the null hypothesis was not accepted for most of the splits, we can deduce that RENT chose factors that are pertinent and significant for determining whether or not a patient suffers from depression [8]. In all of the validation experiments, the MCC scores for each split, which are indicated by the red line, were often rather high. The fact that the red line was continuously located further to the right than the majority of the VS1 and VS2 distributions in most cases is evidence that RENT worked well even [8].

5.6 Further work

This section provides an outline of the further work that may be done in the future.

- Analysis using both stage of the trail: Since this study only utilized data from stage 1, it would be interesting to analyze data from stage 2. Since a unique medicine was administered to each patient during stage 2, as detailed in the column labelled 'Stage2X,' there is room for more investigation into the possibility of accurately predicting the effects of the medication.

- Further analysis regarding the feature selected: The data were incomplete and broad, and there were not enough patients. Therefore, before considering this feature in a clinical trial, it will need to have its viability established using a substantial amount of data.

- Standardization MR images in different centers: When taking MR images, various centres utilize a wide variety of MR equipment, each of which has its unique configuration. Therefore, there ought to be appropriate standards to make collecting data less chaotic.

- Investigating preprocessing methods, especially normalisation and tackling missing values: It is not uncommon for healthcare data to have missing values. However, it is difficult to decide whether the missing values should be eliminated or imputed because the data pertains to healthcare, and the outcome of the prediction is affected by either choice.

- Evaluating RENT and other feature selection methods on different data. For validating RENT selected features and other feature selection strategies, additional study and research are necessary mainly when they are utilized on data on healthcare.

Chapter 6

Conclusion

The short-wide dataset, which consists of radiomics features taken from MR images, was effectively reduced from thousands of features to eleven for this thesis. In addition, RENT cut down the number of features as a feature selector, and different classifiers were utilized for the prediction process.

The thesis utilized the clinical dataset and the anatomical dataset to select the features to use. For example, specific properties of the rostral middle frontal gyrus may be a significant brain location to examine when seeking to identify someone with major depression. Therefore, according to RENT's chosen features, the regions of the brain known as the posterior cingulate cortex, the rostral middle frontal cortex, and the middle temporal cortex are all possible candidates for biomarker regions.

Random forest outperformed Logistic regression, SVM, and other classifiers in terms of overall performance in clinical data. Whereas SVM performed better in anatomical data. After using RENT's feature selection, the performance was greatly improved. Repeated stratified K fold was utilized for train test splits and cross-validation. On the other hand, the predictive performance obtained from this investigation was not particularly impressive. RENT illuminated the significance of every aspect, as well as the significance of certain features. Biomarkers mentioned in this study must be studied further before they can be used in medicine.

Bibliography

[1] WHO, 'Depression,' 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] C. Ettman, G. Cohen, P. Vivier and S. Galea, 'Savings, home ownership, and depression in low-income us adults,' *Social Psychiatry and Psychiatric Epidemiology*, vol. 56, pp. 1–9, Jul. 2021. DOI: 10.1007/s00127-020-01973-y.

[3] L. Gin, N. Wiesenthal, I. Ferreira and K. Cooper, 'Phdepression: Examining how graduate research and teaching affect depression in life sciences phd students,' *CBE life sciences education*, vol. 20, ar41, Sep. 2021. DOI: 10.1187/cbe.21-03-0077.

[4] T. Petersen, 'Handbook of depression. edited by i. h. gotlib and c. l. hammen. guilford press: London. 2002.,' *Psychological Medicine*, vol. 33, pp. 1130–1131, Sep. 2003. DOI: 10.1017/S0033291703248510.

[5] K. Kendler, L. Thornton and C. O. Gardner, 'Genetic risk, number of previous depressive episodes, and stressful life events in predicting onset of major depression.,' *The American journal of psychiatry*, vol. 158 4, pp. 582–6, 2001.

[6] S. Monroe, G. Slavich and K. Georgiades, 'The social environment and depression: The importance of life stress,' *Handbook of Depression, 3rd Ed.*, 2014.

[7] Y. Fang and Y. Fang, 'Depressive disorders: Mechanisms, measurement and management,' *Depressive Disorders: Mechanisms, Measurement and Management*, 2019.

[8] K. Tukun, 'Diagnosing patients with major depressive disorder using radiomics features extracted from mr scans of the brain,' *Norwegian University of Life Sciences*, 2021.

[9] A. Schrantee, H. Ruhé and L. Reneman, 'Psychoradiological biomarkers for psychopharmaceutical effects,' *Neuroimaging Clinics of North America*, vol. 30, no. 1, pp. 53–63, 2020, Psychoradiology, ISSN: 1052-5149. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1052514919300851.

[10] G. Gameiro, V. Sinkunas, G. Liguori and J. Auler-Júnior, 'Precision medicine: Changing the way we think about healthcare,' *Clinics*, vol. 73, Dec. 2018. DOI: 10.6061/clinics/2017/e723.

[11] J. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi and B. Baeßler, 'Radiomics in medical imaging—"how-to" guide and critical reflection,' *Insights into Imaging*, vol. 11, Dec. 2020. DOI: 10.1186/s13244-020-00887-2.

[12] J. Goya-Outi, F. Orlhac, R. Calmon *et al.*, 'Computation of reliable textural indices from multimodal brain MRI: Suggestions based on a study of patients with diffuse intrinsic pontine glioma,' *Physics in Medicine &amp Biology*, vol. 63, no. 10, p. 105 003, May 2018. DOI: 10.1088/1361-6560/aabd21. [Online]. Available: https://doi.org/10.1088/1361-6560/aabd21.

[13] T. Fonseka, G. MacQueen and S. Kennedy, 'Neuroimaging biomarkers as predictors of treatment outcome in major depressive disorder,' *Journal of Affective Disorders*, pp. 21–35, 2018, Are there Biomarkers for Mood Disorders? DOI: https://doi.org/10.1016/j.jad.2017.10.049. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032717310431.

[14] A. Jenul, S. Schrunner, K. H. Liland, U. G. Indahl, C. M. Futsæther and O. Tomic, 'Rent—repeated elastic net technique for feature selection,' *IEEE Access*, vol. 9, pp. 152 333–152 346, 2021. DOI: 10.1109/ACCESS.2021.3126429.

[15] S. Raschka and v. Mirlalili, *Python machine learning, Third Edition*. Packt, 2019. [Online]. Available: https://www.packtpub.com/product/python-machine-learning-third-edition/9781789955750.

[16] C. Otte, S. Gold, B. Penninx *et al.*, 'Major depressive disorder,' *Nature reviews. Disease primers*, vol. 2, p. 16 065, Sep. 2016. DOI: 10.1038/nrdp.2016.65.

[17] C. Fabbri, L. Hosak, R. Mössner *et al.*, 'Consensus paper of the wfsbp task force on genetics: Genetics, epigenetics and gene expression markers of major depressive disorder and antidepressant response,' *The World Journal of Biological Psychiatry*, vol. 18, no. 1, pp. 5–28, 2017. DOI: 10.1080/15622975.2016.1208843. [Online]. Available: https://doi.org/10.1080/15622975.2016.1208843.

[18] C. Murray, C. Atkinson, K. Bhalla *et al.*, 'The state of us health, 1990-2010: Burden of diseases, injuries, and risk factors,' *US Burden of Disease Collaborators*, 2013. DOI: https://doi.org10.1001/jama.2013.13805. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23842577/.

[19] G. H. D. E. (GHDx), 'Institute of health metrics and evaluation,' last accessed on 2022-05-22, 2018. [Online]. Available: https://openreview.net/forum?id=B1Yy1BxCZ.

[20] S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi and et al., 'Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys,' *Psychol Med.*, 2018, ISSN: 1560-1571. DOI: https://doi.org/10.1017/S0033291717003336. [Online]. Available: https://www.https://pubmed.ncbi.nlm.nih.gov/29173244/.

[21] U. o. F. College of Medicine, 'Hamilton depression rating scale (hdrs),' last accessed on 2022-05-22, 2011. [Online]. Available: https://https://dcf.psychiatry.ufl.edu/files/2011/05/HAMILTON-DEPRESSION.pdf.

[22] A. Leuchter and et al., 'A new paradigm for the prediction of antidepressant treatment response.,' *Dialogues in clinical neuroscience*, 2009. DOI: https://doi.org/10.31887/DCNS.2009.11.4/afleuchter.

[23] F.-N. B. W. Group, *BEST (Biomarkers, EndpointS, and other Tools) Resource*. Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US), 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK326791/.

[24] B. Jani, G. McLean, B. Nicholl *et al.*, 'Risk assessment and predicting outcomes in patients with depressive symptoms: A review of potential role of peripheral blood based biomarkers,' *Frontiers in Human Neuroscience*, vol. 9, p. 18, 2015.

[25] J. Prescott, 'Quantitative imaging biomarkers: The application of advanced image processing and analysis to clinical and preclinical decision making,' *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, vol. 26, Mar. 2012. DOI: 10.1007/s10278-012-9465-7.

[26] J. Konarski, R. McIntyre, S. Kennedy, S. Rafi-Tari, J. Soczynska and T. Ketter, 'Volumetric neuroimaging investigations in mood disorders: Bipolar disorder versus major depressive disorder.,' *Bipolar disorders.*, 2008. DOI: https://doi.org/10.1111/j.1399-5618.2008.00435.x.

[27] S. Brand, M. Moller and B. Harvey, 'A review of biomarkers in mood and psychotic disorders: A dissection of clinical vs. preclinical correlates,' *Current neuropharmacology*, vol. 13, no. 3, pp. 324–368, 2015.

[28] A. Lacerda, M. Keshavan, A. Hardan *et al.*, 'Anatomic evaluation of the orbitofrontal cortex in major depressive disorder,' *Biological Psychiatry*, vol. 55, no. 4, pp. 353–358, 2004, ISSN: 0006-3223. DOI: https://doi.org/10.1016/j.biopsych.2003.08.021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006322303009491.

[29] R. Strawbridge, A. Young and A. Cleare, 'Biomarkers for depression: Recent insights, current challenges and future prospects,' *Neuropsychiatric Disease and Treatment*, vol. Volume 13, pp. 1245–1262, May 2017. DOI: 10.2147/NDT.S114542.

[30] F. Bloch, W. Hansen and M. Packard, 'Nuclear induction,' *Phys Rev.*, 1946.

[31] R. Hobbie, 'Intermediate physics for medicine and biology,' *Springer*, 2007.

[32] A. Berger, 'Magnetic resonance imaging,' *BMJ (Clinical research ed.)*, 2002. DOI: 10.1136/bmj.324.7328.35.

[33] V. Vassiliou, D. Cameron, S. Prasad and P. Gatehouse, 'Magnetic resonance imaging: Physics basics for the cardiologist,' *JRSM Cardiovascular Disease*, vol. 7, p. 2 048 004 018 772 237, 2018, PMID: 30128147. DOI: 10.1177/2048004018772237. [Online]. Available: https://doi.org/10.1177/2048004018772237.

[34] Y. Chen, S. J. Almarzouqi and A. G. Morgan Michael L.and Lee, 'T1-weighted image,' in *Encyclopedia of Ophthalmology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 1747–1750, ISBN: 978-3-540-69000-9. DOI: 10.1007/978-3-540-69000-9_1228. [Online]. Available: https://doi.org/10.1007/978-3-540-69000-9_1228.

[35] R. Bitar, G. Leung, R. Perng *et al.*, 'Mr pulse sequences: What every radiologist wants to know but is afraid to ask,' *Radiographics*, vol. 26, no. 2, pp. 513–537, 2006.

[36] S. Rizzo, F. Botta, S. Raimondi *et al.*, 'Radiomics: The facts and the challenges of image analysis,' *European Radiology Experimental*, vol. 2, Dec. 2018. DOI: 10.1186/s41747-018-0068-z.

[37] N. Mohammadi, 'Radiomics using mr brain scans and rent for identifying patients receiving adhd treatment,' *Norwegian University of Life Sciences*, 2021.

[38] T. Exarchos, A. Papadopoulos and D. Fotiadis, *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications.* 2009, ISBN: 9781605663142. DOI: 10.4018/978-1-60566-314-2.

[39] R. Gillies, P. Kinahan and H. Hricak, 'Radiomics: Images are more than pictures, they are data,' *Radiology*, vol. 278, no. 2, pp. 563–577, 2016, PMID: 26579733. DOI: 10.1148/radiol.2015151169. [Online]. Available: https://doi.org/10.1148/radiol.2015151169.

[40] A. Destrero, S. Mosci, C. Mol, A. Verri and F. Odone, 'Feature selection for high-dimensional data,' *Computational Management Science*, vol. 6, pp. 25–40, Feb. 2009. DOI: 10.1007/s10287-008-0070-7.

[41] J. Miao and L. Niu, 'A survey on feature selection,' *Procedia Computer Science*, vol. 91, pp. 919–926, Dec. 2016. DOI: 10.1016/j.procs.2016.07.111.

[42] V. Bolon-Canedo, N. Sánchez-Maroño and A. Alonso-Betanzos, 'A review of feature selection methods on synthetic data,' *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[43] H. Singh, B. Singh and M. Aneja, 'An efficient feature selection method based on improved elephant herding optimization to classify high-dimensional biomedical data,' *Expert Systems*, May 2022. DOI: 10.1111/exsy.13038.

[44] C. Olofsson, 'Using machine learning and repeated elastic net techniquefor identifcation of biomarkers of early alzheimer's disease,' *Norwegian University of Life Sciences*, 2021.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, 'Scikit-learn: Machine learning in Python,' *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/stable/about.html#citing-scikit-learn.

[46] D. Cousineau and S. Chartier, 'Outliers detection and treatment: A review,' *International Journal of Psychological Research*, vol. 3, Jun. 2010. DOI: 10.21500/20112084.844.

[47] V. Chandola, A. Banerjee and V. Kumar, 'Outlier detection: A survey,' *ACM Computing Surveys*, vol. 14, p. 15, 2007.

[48] H. P. Vinutha, B. Poornima and B. M. Sagar, 'Detection of outliers using interquartile range technique from intrusion dataset,' in *Information and Decision Sciences*, S. C. Satapathy, J. M. R. Tavares, V. Bhateja and J. R. Mohanty, Eds., Singapore: Springer Singapore, 2018, pp. 511–518, ISBN: 978-981-10-7563-6.

[49] G. Hackeling, *Mastering machine learning with scikit-learn : apply effective learning algorithms to real-world problems usung scikit-learn, Second ed.* Packt, 2017, ISBN: 9781788299879. [Online]. Available: https://www.packtpub.com/product/mastering-machine-learning-with-scikit-learn-second-edition/9781788299879.

[50] B. Matthews, 'Comparison of the predicted and observed secondary structure of t4 phage lysozyme,' *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975, ISSN: 0005-2795. DOI: https://doi.org/10.1016/0005-2795(75)90109-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005279575901099.

[51] S. Misra, H. Li and J. He, *Machine Learning for Subsurface Characterization*. Elsevier Science, 2019, ISBN: 9780128177372. [Online]. Available: https://books.google.no/books?id=WdO1DwAAQBAJ.

[52] M. Schonlau and R. Yuyan Zou, 'The random forest algorithm for statistical learning,' *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020. DOI: 10.1177/1536867X20909688. eprint: https://doi.org/10.1177/1536867X20909688. [Online]. Available: https://doi.org/10.1177/1536867X20909688.

[53] A. Sarica, A. Cerasa and A. Quattrone, 'Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: A systematic review,' *Frontiers in Aging Neuroscience*, vol. 9, p. 329, Oct. 2017. DOI: 10.3389/fnagi.2017.00329.

[54] R. Caruana and A. Niculescu-Mizil, 'An empirical comparison of supervised learning algorithms,' in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.

[55] B. Menze, B. Kelm, R. Masuch, U. Himmelreich, P. Bachert and W. Petrich, 'And fa hamprecht (2009). a comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data,' *BMC Bioinformatics*, vol. 10, p. 213,

[56] V. Gupta, G. Singh, R. Singh, R. Singh and H. Singh, 'An introduction to principal component analysis and its importance in biomedical signal processing,' May 2022.

[57] A. Sykes, *An Introduction to Regression Analysis*, ser. Coase lecture. Law School, University of Chicago, 1993. [Online]. Available: https://books.google.no/books?id=hxyaHAAACAAJ.

[58] M. Trivedi, P. McGrath, M. Fava *et al.*, 'Establishing moderators and biosignatures of antidepressant response in clinical care (embarc): Rationale and design,' in *J Psychiatr Res. 2016*, 2016.

[59] H. Singh and A. Saadabadi, 'Sertraline,' *StatPearls*, vol. 26, no. 2, pp. 513–537, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK547689/.

[60] M. Korgaonkar, L. Williams, Y. Song, T. Usherwood and S. Grieve, 'Diffusion tensor imaging predictors of treatment outcomes in major depressive disorder,' in *Br J Psychiatry 2014*, 2014. DOI: 10.1192/bjp.bp.113.140376.

[61] M. Korgaonkar, S. Grieve, S. Koslow, J. Gabrieli, E. Gordon and L. Williams, 'Loss of white matter integrity in major depressive disorder: Evidence using tract-based spatial statistical analysis of diffusion tensor imaging,' in *Hum Brain Mapp 2011*, 2011.

[62] L. Henschel, 'Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline,' in *Neuroimage 2019*, 2020.

[63] J. J. van Griethuysen, A. Fedorov, C. Parmar *et al.*, 'Computational Radiomics System to Decode the Radiographic Phenotype,' *Cancer Research*, vol. 77, no. 21, e104–e107, Oct. 2017, ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-17-0339. eprint: https://aacrjournals.org/cancerres/article-pdf/77/21/e104/2934659/e104.pdf. [Online]. Available: https://doi.org/10.1158/0008-5472.CAN-17-0339.

[64] *Anaconda software distribution*, version Vers. 2-2.4.0, 2020. [Online]. Available: https://docs.anaconda.com/.

[65] W. McKinney, 'Data Structures for Statistical Computing in Python,' in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[66] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, 'Array programming with NumPy,' *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[67] O. Tomic, T. Graff, K. H. Liland and T. Næs, 'Hoggorm: A python library for explorative multivariate statistics,' *The Journal of Open Source Software*, vol. 4, no. 39, 2019. DOI: 10.21105/joss.00980. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00980.

[68] O. Tomic, T. Graff, K. Liland and T. Næs, 'Hoggorm: A python library for explorative multivariate statistics,' *Journal of Open Source Software*, vol. 4, p. 980, Jul. 2019. DOI: 10.21105/joss.00980.

[69] J. D. Hunter, 'Matplotlib: A 2d graphics environment,' *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[70] P. T. Inc. ‘Collaborative data science.’ (2015), [Online]. Available: https://plot.ly.

[71] M. L. Waskom, ‘Seaborn statistical data visualization,’ *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. [Online]. Available: https://doi.org/10.21105/joss.03021.

Appendix A

Appendix

Appendix A: Figures from exploratory PCA during RENT selected features Exploratory PCA of first two and last two splits;first, second, second to last and last split on clinical dataset. Images of only the first two and last two splits are featured in this part due to a large number of images overall.

CLINICAL DATASET & ANATOMICAL DATASET

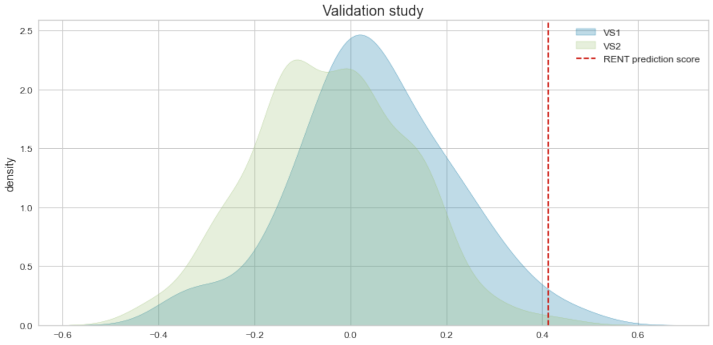

Figure A.1: Validation study on Split 1 by RENT selected features based on $K = 100$ models on clinical data.

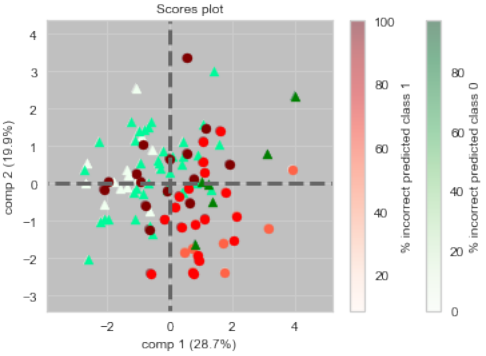

Figure A.2: Classification and misClassification on Split 1 by RENT selected features on clinical data.

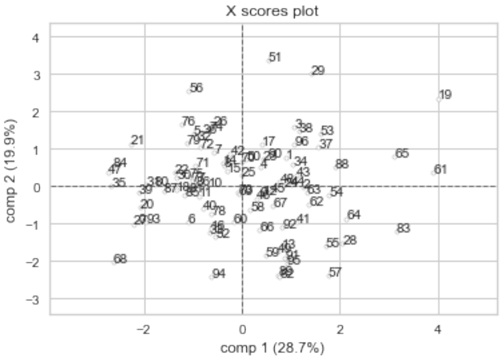

Figure A.3: PCA score plot on Split 1 by RENT selected features on clinical data.

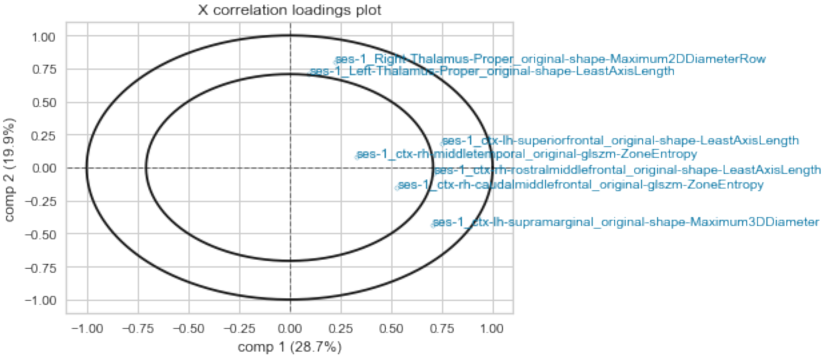

Figure A.4: PCA correlation plot on Split 1 by RENT selected features on clinical data.

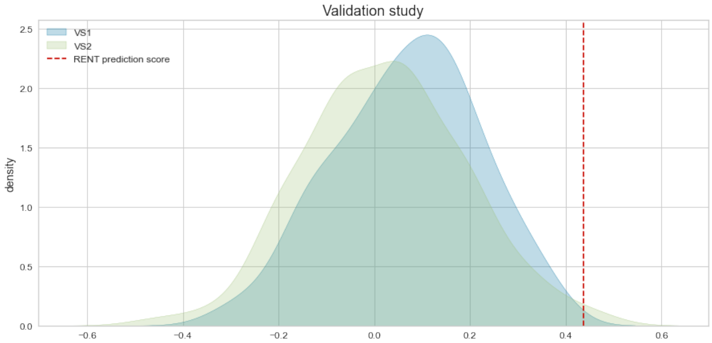

Figure A.5: Validation study on Split 2 by RENT selected features based on $K = 100$ models on clinical data.

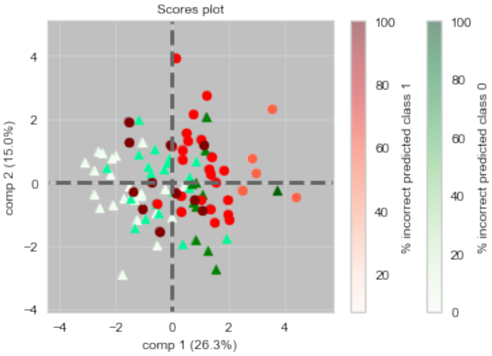

Figure A.6: Classification and misClassification on Split 2 by RENT selected features on clinical data.

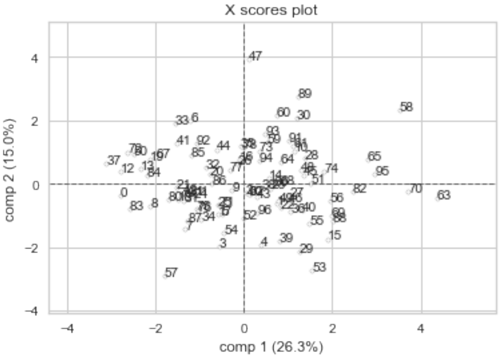

Figure A.7: PCA score plot on Split 2 by RENT selected features on clinical data.

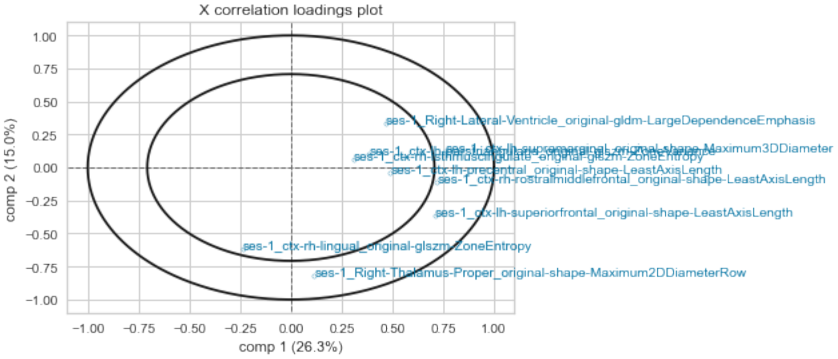

Figure A.8: PCA correlation plot on Split 2 by RENT selected features on clinical data.

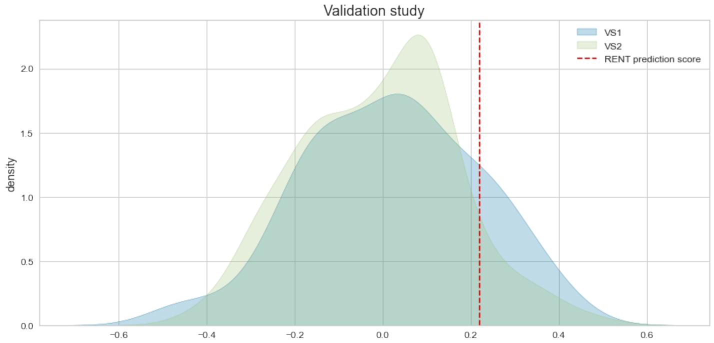

Figure A.9: Validation study on Split 12 by RENT selected features on clinical data. based on $K = 100$ models.

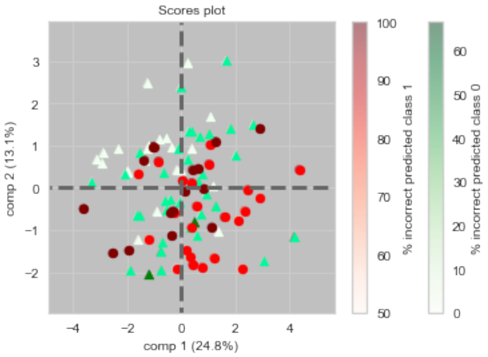

Figure A.10: Classification and misClassification on Split 12 by RENT selected features on clinical data.

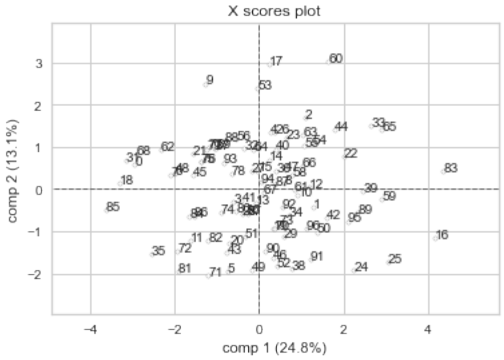

Figure A.11: PCA score plot on Split 12 by RENT selected features on clinical data.

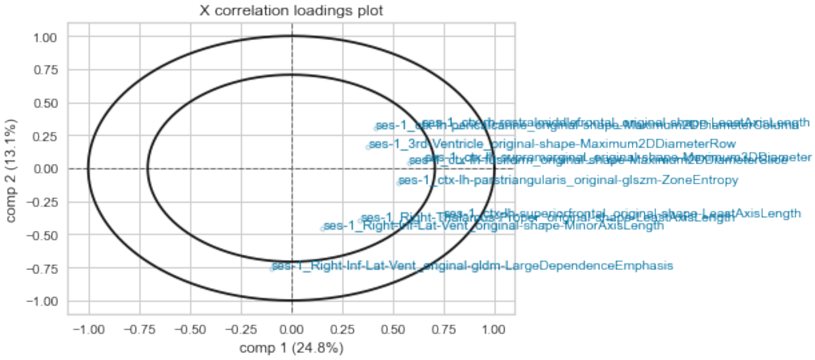

Figure A.12: PCA correlation plot on Split 2 by RENT selected features on clinical data.

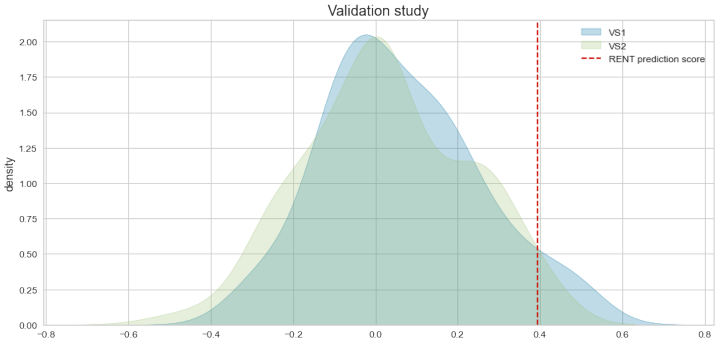

Figure A.13: Validation study on Split 11 by RENT selected features based on $K = 100$ models on clinical data.

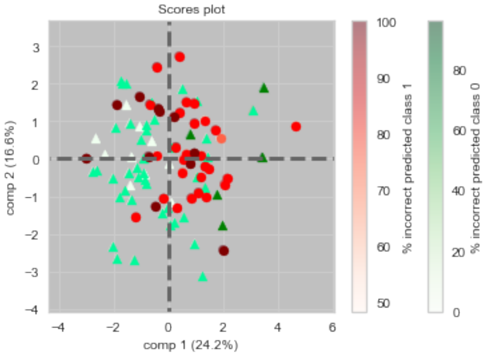

Figure A.14: Classification and misClassification on Split 11 by RENT selected features on clinical data.

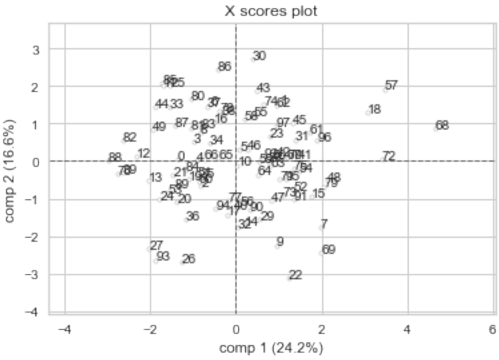

Figure A.15: PCA score plot on Split 11 by RENT selected features on clinical data.

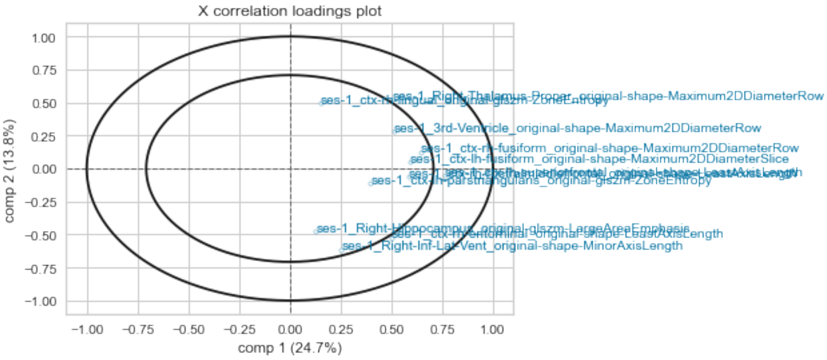

Figure A.16: PCA correlation plot on Split 11 by RENT selected features on clinical data.

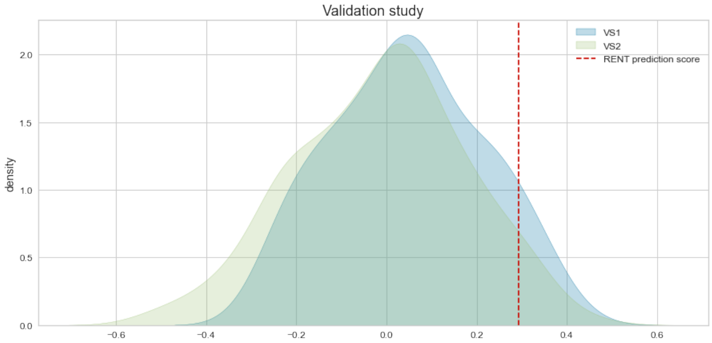

Figure A.17: Validation study on Split 1 by RENT selected features based on $K = 100$ models on anatomical data.

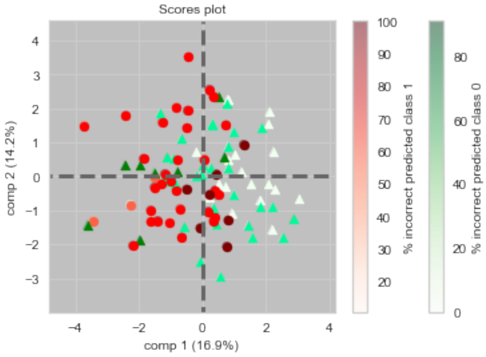

Figure A.18: Classification and misClassification on Split 1 by RENT selected features on anatomical data.

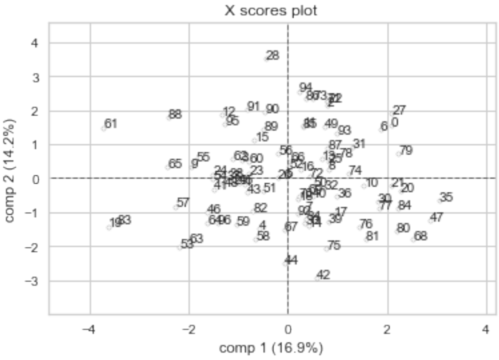

Figure A.19: PCA score plot on Split 1 by RENT selected features on anatomical data.

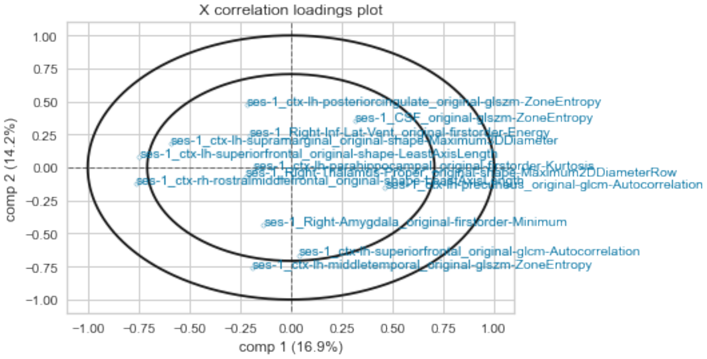

Figure A.20: PCA correlation plot on Split 1 by RENT selected features on anatomical data.

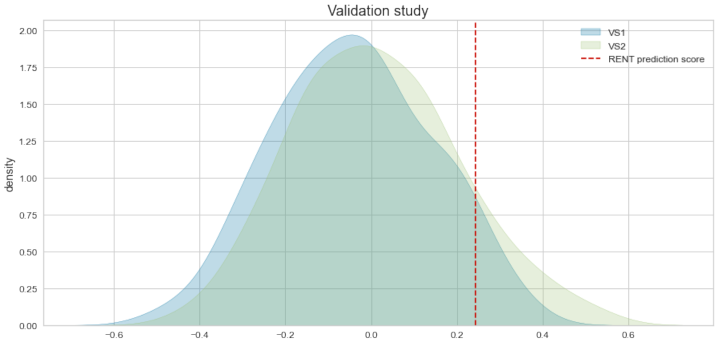

Figure A.21: Validation study on Split 2 by RENT selected features based on $K = 100$ models on anatomical data.

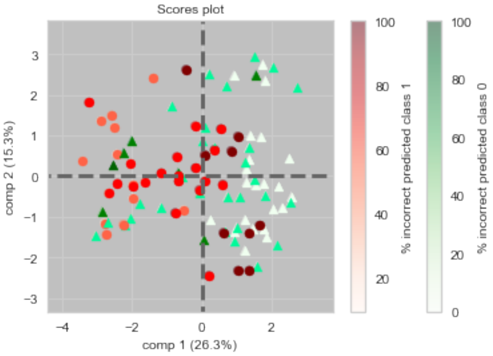

Figure A.22: Classification and misClassification on Split 2 by RENT selected features on anatomical data.

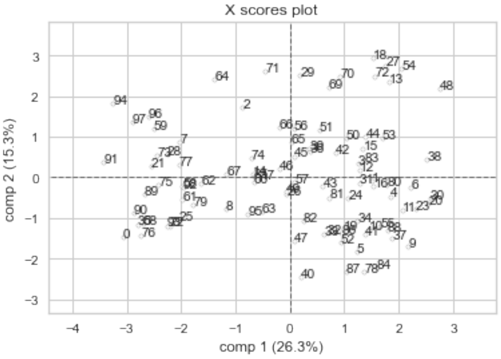

Figure A.23: PCA score plot on Split 2 by RENT selected features on anatomical data.

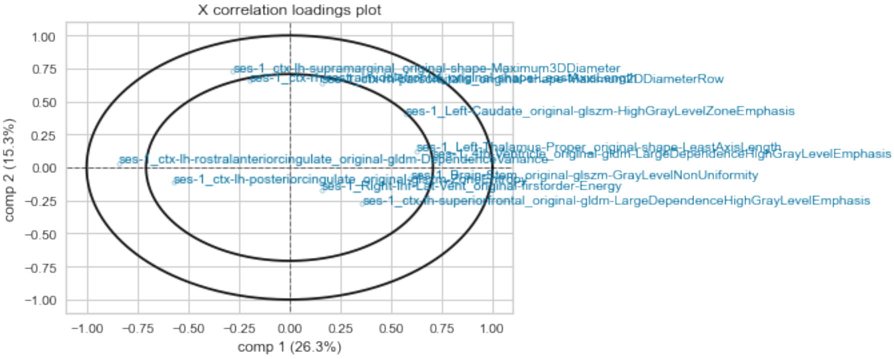

Figure A.24: PCA correlation plot on Split 2 by RENT selected features on anatomical data.

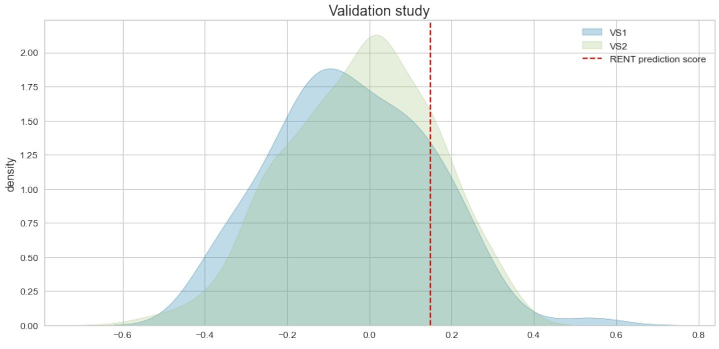

Figure A.25: Validation study on Split 11 by RENT selected features on anatomical data. based on $K = 100$ models.

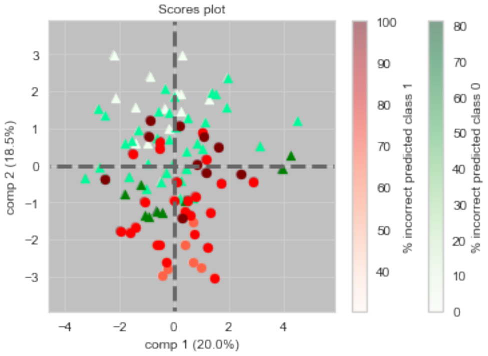

Figure A.26: Classification and misClassification on Split 11 by RENT selected features on anatomical data.

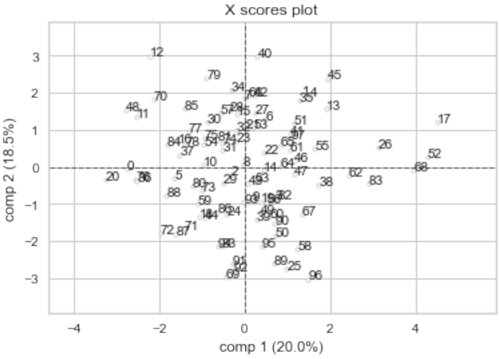

Figure A.27: PCA score plot on Split 11 by RENT selected features on anatomical data.

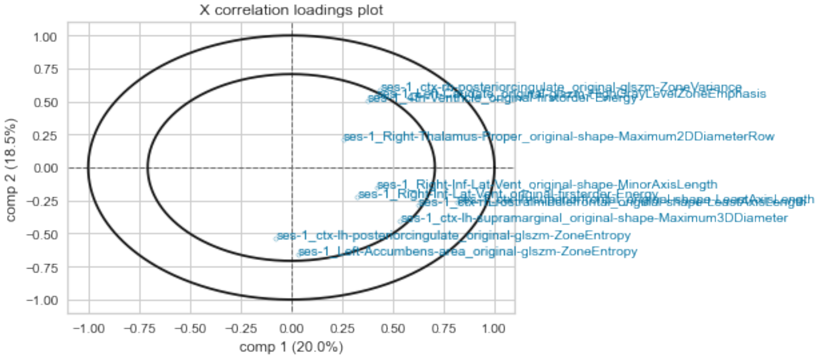

Figure A.28: PCA correlation plot on Split 11 by RENT selected features on anatomical data.

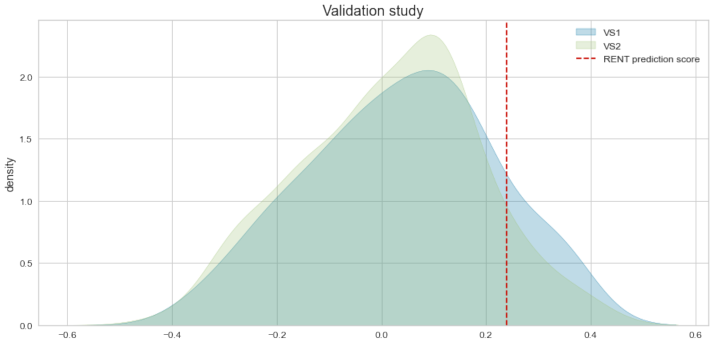

Figure A.29: Validation study on Split 12 by RENT selected features based on $K = 100$ models on anatomical data.

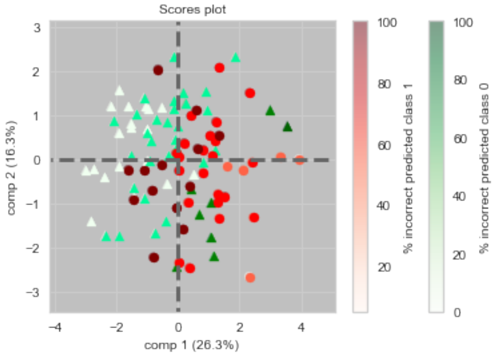

Figure A.30: Classification and misClassification on Split 12 by RENT selected features on anatomical data.

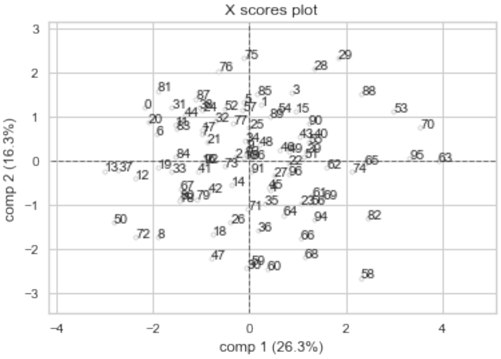

Figure A.31: PCA score plot on Split 12 by RENT selected features on anatomical data.

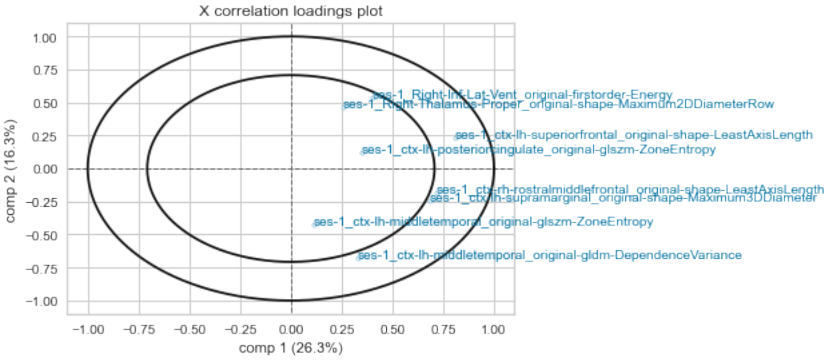

Figure A.32: PCA correlation plot on Split 2 by RENT selected features on anatomical data.

Thank you.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway