Norwegian University
of Life Sciences

Master's Thesis 2021 30 ECTS
Faculty of Science and Technology
Professor Cecilia Marie Futsæther

Diagnosing patients with Major Depressive Disorder using radiomics features extracted from MR scans of the brain

Kristin Tukun
Environmental Physics and Renewable Energy

Acknowledgements

There are many bright minds and supportive people that have helped me produce this thesis. First and foremost, I want to thank my main supervisor, Professor Cecilia Marie Futsæther, for being helpful, supportive, and always cheerful throughout the process. And a big thanks to Ass. Prof. Oliver Tomic, who assisted with valuable input and good ideas.

I want to thank Prof. Atle Bjørnerud, Dr Inge Groote, and Jon E Nesvold for providing the dataset, answering many questions, and giving me valuable input. Especially Jon who has assisted me and answered questions at all hours throughout this period.

I also want to thank all my peers at NMBU; our study groups, collaborations, and social gatherings have helped me immensely through all my courses at NMBU.

To my oldest friends Ingvild and Iselin, I want to thank you for motivating me; from the first day at NMBU to the last day before delivering my thesis, you have both brought so much joy into my life.

At last, I want to thank my boyfriend, who has listened to endless ramblings and contributed with valuable input without one single notion of what this thesis is all about.

Abstract

The main aim of this study was to diagnose patients with major depressive disorder (MDD) using structural T1 weighted images of the brain. The images originate from the DELHI study conducted in the Netherlands from July 2005 to February 2007. 21 images of patients diagnosed with MDD and 22 images of healthy controls were received. Patients and controls were scanned at study entry referenced to as t0. The patients were administered the antidepressant paroxetine and scanned again at 6 and 12 weeks, referenced as timesteps t1 and t2. Further, the images were sent to an application programming interface (API) called RadiomPipe that segmented the images into masks of brain regions and extracted radiomics features. The output from RadiomPipe was a high dimensional dataset that consisted of 10165 radiomics features.

This study mainly focused on the dataset containing radiomics features of patients and controls at study entry, t0. The samples were split into a training and test set three times stratified by whether the individual belonged to the patient class or the control class. The Repeated Elastic Net Technique (RENT) algorithm was applied to each of these splits to reduce the number of features in the dataset by training an ensemble of 100 elastic net regularized models and selecting features by evaluating the weight distributions of features across the models. The first split selected eight features while the two other splits selected seven features. Further, the ensemble of models predicted the diagnosis of patients with high accuracy over all three splits. The high accuracies indicated that the RENT model was a robust model that potentially can perform well on new data.

A Principal Component Analysis (PCA) was conducted with the features selected by RENT for each split. The PCA showed that it was possible to separate the patient and control class using only the features selected by RENT for each split.

In total RENT selected 14 features across the splits. Four of these features were selected by every split. A PCA was applied to the dataset containing these four features, which showed that the patient and control classes could be separated. These four features corresponded to the brain region right medial orbital gyrus. Given that all three splits found the right medial orbital gyrus useful to predict MDD diagnosed patients, it may be considered a possible neural biomarker.

Contents

List of Abbreviations

ACC	Accuracy.
API	Application Programming Interface.
DSM	Diagnostic and Statistical Manual of Mental Disorder.
fMRI	Functional Magnetic Resonance Imaging.
FN	False Negative.
FP	False Positive.
GLCM	Grey level Co-occurrence Matrix.
GLDM	Grey Level Dependence Matrix.
GLM	Generalized linear model.
GLRLM	Grey Level Run Length Matrix.
GLSZM	Grey Level Size Zone Matrix.
HDRS17	17-item Hamilton Depression Rating Scale.
ICD	International Classification of Diseases.
MDD	Major Depressive Disorder.
MR	Magnetic Resonance.
MRI	Magnetic Resonance Imaging.
NGTDM	Neighbouring Grey Tone Difference Matrix.
PC	principal component.
PCA	Principle Component Analysis.
RENT	Repeated Elastic Net Technique.
REST	Representational State Transfer.
RF	radio frequency.
ROI	Region of Interest.
TE	Time Echo.
TN	True Negative.

TR	Repetition Time.
WHO	World Health Organization.

1 Introduction

Depressive disorder affect our society as it is one of the most widespread mental disorders [1]. In 2015 The World Health Organization recorded that about 264 million people suffered from depressive disorders [2]. Depression emerges from a combination of social, psychological, and biological factors [1]. Major depressive disorder (MDD) is a psychiatric disorder characterized by different symptoms, such as sadness, little self-worth, poor appetite and reduced sleep [3]. Antidepressants are the primary type of treatment for moderate to severe depressive episodes, and six decades of efforts have yet to improve their efficiency [1]. When treating depression, a "trial-and-error" approach of prescribing antidepressants is often applied [4]. The expected time before the antidepressants takes effect, can range from 2 to 8 weeks. However, if there is no effect, a new antidepressant is prescribed. This method for treating depressive disorders leads to a prolonged treatment course, especially since four trials of different antidepressants yield a cumulative remission rate of 67% [4]. Patients may go through eight months of treatment, and the chance of treatment effect would be only 67%.

In order to shorten the treatment course of depressive patients, precision medicine has been developed [4]. Precision medicine is treatment methods that take into account the individual's variability. Instead of randomly trying treatment methods, precision medicine opens the doors for applying more suitable treatment for the patient. Precision medicine uses biomarkers for targeting treatment. Biomarkers are biological characteristics and can be molecular, anatomical, physiological, or biochemical [4]. Biomarkers can also be extracted from magnetic resonance (MR) images of the brain [4]. These radiomics features can be extracted based on the patterns in voxel intensities in MR images of the brain [5]. The MR images are then transformed into high-dimensional datasets that consists of radiomics features. In these datasets, machine learning algorithms can see patterns that the human eye may not detect. It may therefore be possible to find biomarkers for diagnosing MDD and selecting treatment strategy. However, there are certain issues with detecting biomarkers from MR images. MR images are difficult to replicate; even when using the same patient and the same MR scanner, the image can be different between visits [6]. Standardization of the images is therefore critical in order to compare the images. Multiple studies have found possible biomarkers in the frontolimbic region like the hippocampus, prefrontal cortex, anterior cingulate cortex, amygdala, and insula associated with treatment response for MDD diagnosed patients [1, 7, 8, 9]. However, their strength and association are variable [7]. Biomarkers found through

medical images also have to be validated and replicated multiple times in order to ensure their reliability [7]. Therefore, it is expected to take some time before these types of biomarkers will be used in the medical field.

The radiomics dataset analyzed in this thesis was extracted from structural T1 weighted MR images of 21 patients diagnosed with MDD and 22 healthy controls in the DELPHI study conducted in the Netherlands from July 2005 to February 2007 [10].

The primary objective of this thesis was to diagnose patients with MDD based on radiomics features extracted from MR images using the Repeated Elastic Net Technique (RENT). Furthermore, it was also of interest to investigate if the selected radiomics features were possible biomarkers for depression. The study used RENT to select features and predict diagnosis. Principal component analysis (PCA) was used on the RENT selected features to see if the patient and control classes could be separated.

The thesis starts by describing the theory associated with this thesis, explaining the fundamental background of key elements used in the method. It explains depression and how it is treated, Magnetic Resonance Imaging (MRI), radiomics, RENT, Pearson's correlation coefficient, accuracy metrics, and PCA. Chapter 3 describes the method applied in this thesis and explains how the dataset was preprocessed and how RENT and PCA were applied. The results are presented in chapter 4 and discussed further in chapter 5. Lastly, the conclusion can be read in chapter 7.

2 Theory

2.1 Depression

A patient suffering from major depressive disorder (MDD) can experience sadness, loss of interest or pleasure, irritable mood, along with somatic and cognitive changes [3]. Depressive disorders develop by interactions between social, psychological, and biological factors. The world health organization (WHO) recorded in 2017 that about 264 million people suffered from depressive disorder [2].

Depression is diagnosed based on symptoms by the use of diagnostic manuals such as the Diagnostic and Statistical Manual of Mental Disorder (DSM) or International Classification of Diseases (ICD) [4]. The 17-item Hamilton Depression Rating Scale (HDRS17) is the most widely used scale for assessing the severity of depressive symptoms [11]. Like many other diagnostic scales for depression, the Hamilton scale measures depression symptoms on a continuous scale. Strategies for treating depression are often pharmacotherapy, psychotherapy and physical therapy [1]. The common procedure for treating depressive disorder is to use the least intrusive interventions first and proceed with further treatment if the treatment outcome is not satisfactory.

Pharmacotherapy often uses a "trial-and-error" method of prescribing psychiatric medication [4]. After the first treatment, approximately 30% to 50% will experience full remission. After 4 trials of different antidepressants the cumulative remission rate is 67%. A drawback of these antidepressants are the time it takes for the antidepressant to take effect, which often take from 2 to 8 weeks [1].

Personalized treatment can take into account the variability in the population and potentially shorten the course of treatment [4]. Personalized medicine, also called precision medicine is prevention and treatment strategies that take into account individual variability. The idea is to look at the individual patients biomarkers that takes into account the individual's environment, genes and lifestyle. A biomarker is defined by the Health Research Directorate of the EU as a biological characteristic which can be molecular, anatomical, physiological or biochemical and can be evaluated objectively. An algorithm can use these biomarkers to make a decision for the patient's treatment course and provide the physician with data to make an individual assessment of the patient. Cancer treatment has been revolutionized by personalized treatment, where treatments are based off the tumor's genomic

profile [4]. An application of personalized treatment in psychiatry is challenging as there is no medical test to assert psychological diagnoses [4].

Biomarkers can be extracted from medical images of the brain [4]. The brain can be imaged on a molecular level using PET, single-photon emission computed tomography or the technique phar-macologic MR imaging ([4]. Physiological characteristics can be obtained by using functional MR imaging (fMR imaging) and perfusion imaging and biochemical properties can be extracted by using MR spectroscopy [4].

Several studies have identified fronto-limbic regions, in particular the hippocampus, prefrontal, anterior cingulate cortex, amygdala and insula as frequently predictive for treatment response for MDD diagnosed patients [1]. Fonseka et. Al (2018) [7] found multiple possible biomarkers for treatment response from structural and functional neuroimaging modalities from 95 studies. The biomarkers were mostly found in fronto-limbic regions, including the prefrontal cortex, anterior cingulate cortex, hippocampus, amygdala, and insula. Although the strength and direction of association varied. Konarski et al (2008) [8] reviewed 140 magnetic neuroimaging investigations in either bipolar disorder or MDD and found similar results. Several studies reported a reduction in volumetric changes in prefrontal cortical areas, especially in the cingulate and orbifrontal regions. There were also consistent evidence of a reduction in the hippocampal volumes in MDD diagnosed patients. A smaller volumes of striatal and amygdala volumes were also reported for MDD diagnosed patients. Lacerda et al. (2004) [9] observed a smaller volume of gray matter in the lateral and medial orbitofrontal cortex (OFC) in the MDD patients.

Although there are several possible biomarkers to extract from medical images there are several challenges connected to extracting useful biomarkers for psychiatric disorders [4]. The first main challenge is the different protocols for psychiatric diagnostics, which complicates the validation of the biomarker. The second main challenge is that the features for a psychiatric disease can be difficult to observe and can only appear under special conditions or under a specific cognitive load [4]. The neural imaging field is often focused on producing new results instead of replicating and validating biomarkers, which explains the scarcity of replicated findings. There is also a practical challenge involving cost of scans. In the United States a MR-scan cost approximately 600$ per hour for academic centers [4].

2.2 MRI

Magnetic Resonance Imaging (MRI) gives us high resolution images without the use of ionizing radiation [12]. It is based on the behavior of magnetic dipoles in the nuclei of atoms in the human body when a magnetic field is applied. Generally MRI uses the density of the hydrogen atoms mainly in water and fat to differentiate between tissue/structures in the body [13].

Let us consider an isolated proton, with a charge $+e$ and a spin angular momentum I. The protons charge can be viewed as evenly distributed and rotates around a central axis through the proton because of the angular momentum. This constructs an magnetic field and a dipole moment μ parallel (for a proton) to the angular momentum vector, and normal to the plane of charge circulation. An representation of this can be viewed in figure 1.

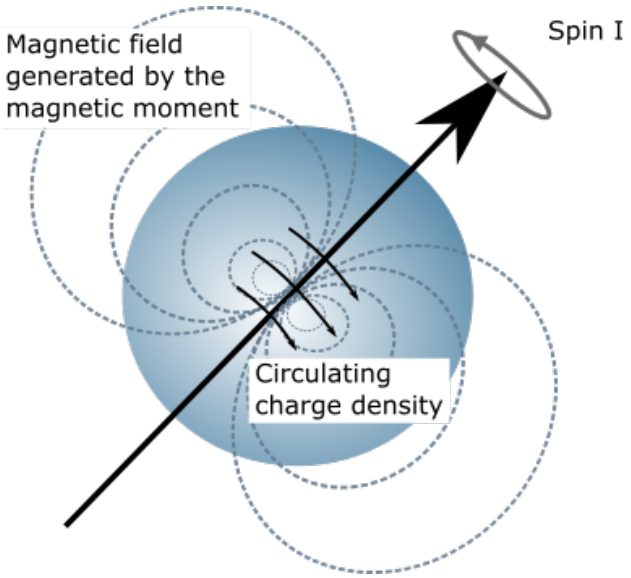

Figure 1: Representation of a nucleus with spin angular momentum I. The circulating charge density generates a magnetic moment which constructs a dipolar magnetic field. This figure was adapted from Fig 7.1 in Flower [13].

The magnetization M is the average magnetic moment per unit volume [12]. As a magnetic field B_0 is applied in the z-direction, the magnetization will increase until all the magnetic moments are aligned and the net equilibrium magnetization M_0 is aligned with B_0 [13]. By applying a pulse of a weaker magnetic field, B_1 in the xy-plane, the net magnetization, M will

experience a torque. The net magnetization is then rotated by an angle of α, due to the pulse exciting the nuclei. As the weaker field stops, the protons will again realign with the direction of B_0. This realignment is the source of low-energy radio frequency photons, referred to as RF signals [13]. These radio frequency (RF) signals are recorded by an RF-coil and interpreted into a medical image. The realignment is referred to as relaxation, the length of it varies depending on the matter that is being studied [12]. Different tissues have different relaxation times, which are easily detected in an image.

MRI can obtain tissue contrasts from T1 and T2 relaxation times [14]. The T1 relaxation time is the time it takes after the RF pulse is turned off for the protons to realign with B_0 and give up their excess energy. The T2 relaxation time is defined as the time it takes for the transverse magnetization M_{xy}, to decay. The weighting of an image contrast is accomplished by selecting the timing parameters of the RF pulse sequence [15]. The parameters repetition time (TR) and time echo (TE) are closely related to the tissue properties T1 and T2 [14]. TR and TE can be adjusted by the operator, while T1 and T2 are fixed tissue properties. TR is the time between RF-pulses. TE is the time it takes from the RF signal is delivered to the measurement is conducted. TR primarily controls the amount of T1 weighting, while TE controls the amount of T2 weighting [15]. In T1 weighted images of the brain, the white matter will have higher intensity values than gray matter. In T2 weighted images, the intensity values for white matter are lower than for gray matter.

Functional magnetic resonance imaging (fMRI) enables us to study structure and function at the same time [12]. fMRI exploits inhomogeneities in the magnetic fields due to the difference in magnetic properties between oxygenated and deoxygenated hemoglobin. There is no need for an external agent as oxygenated hemoglobin is less paramagnetic than deoxyhemoglobin [12]. A fMRI will look different before and after blood has flowed to a tissue mainly because the blood is oxygenized (change in blood oxygenation). If a brain region is active, one can usually see an increase in blood flow in this region. By using this non-invasive method, fMRI can provide the same functional information as PET, without the use of radionuclides [12].

An MRI image can be viewed as a 3D matrix of size $i \times j \times k$ containing voxels with grayscale intensity values. The $i \times j$ dimension can be viewed as one slice, as seen in figure 2, the slices make up the MR image in the k dimension.

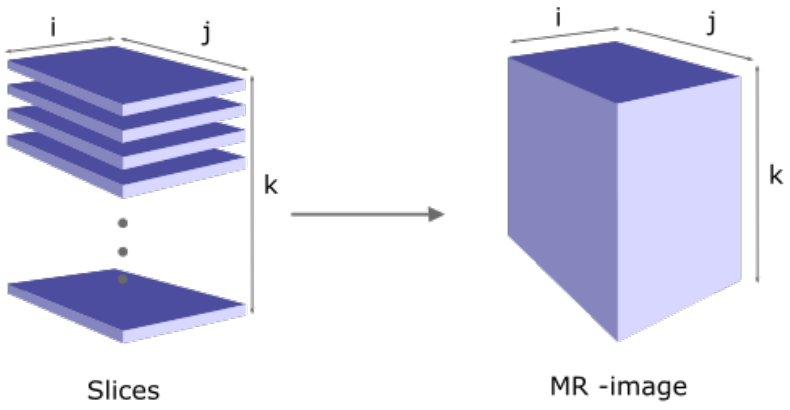

Figure 2: An MR- image can consists of k slices with dimensions, $i \times j$. Each slice contains $i \times j$ voxels, and all the slices together constructs the 3-dimensional MR-image.

2.3 Radiomics

Radiomics is not strictly defined, but essentially aims to extract quantitative and reproducible information that is based on patterns in diagnostic images, that is difficult to see for the human eye [5]. It can be used to observe tissue and lesion changes over time or treatment by extracting properties like shape and heterogeneity. Or it can be used in explorative data analysis as the datasets are often large. This enables the possibility to observe new biomarkers and patterns for disease evolution and treatment response. Extracted radiomics features are often divided into three types of features, first order statistics features, shape features and texture features.

First order statistics features describe the region of interest by using common statistical measurements that are based on the occurrence of voxel intensity values [5]. Examples are mean, variance, maximum, minimum and percentiles. To reflect the shape of the intensity distributions measurements like skewness and kurtosis are included.

Shape features describe the shape of the region of interest by properties like volume, maximum diameter along different orthogonal directions and maximum surface [16].

Texture features are calculated based on the statistical relationship between the neighboring voxels, and provides information about the spatial arrangement of the voxel intensities [16]. There are multiple subcategories of texture features. This thesis presents five of these subcategories.

The Gray level Co-occurrence Matrix (GLCM) captures the relationship between pairs of voxels [17]. The $(i, j)^{th}$ position in the matrix represents how many times a combination of intensity values i and j occurs with a predefined distance δ along an angle θ. An example of how the GLC matrix is calculated for a two dimensional image with four intensity values is shown in figure 3. The size of the GLC matrix will be $n \times n$, where n is the number of discrete intensity values present in the image.

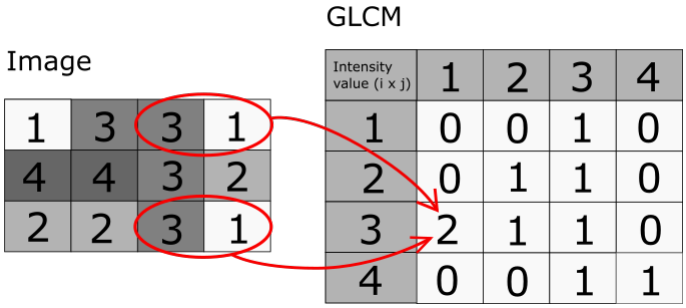

Figure 3: An example of how the GLCM is calculated with $\delta = 1$ and $\theta = 0$ (horizontally). Intensity values 3 and 1 occur two times in that order, element (3,1) in the GLC matrix is therefore equal two. As there are four intensity values, the size of the matrix is 4×4.

The Gray Level Size Zone Matrix (GLSZM) counts the number of zones with voxels of the same gray level intensity value [5]. Two voxels are considered connected if the distance is 1 [17]. The GLSZM is independent of rotation. Only one matrix is calculated for all rotations in the ROI. The GLSZM is calculated as shown in the example in figure 4. Position (i, j) represents the number of intensity zones where i is the intensity value and j is the size of intensity zone. The size of the GLSZ matrix is $i \times j$.

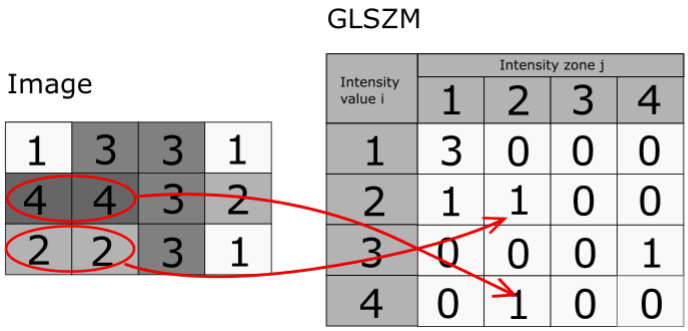

Figure 4: An example of how the GLSZM is calculated. The intensity values (i) 2 and 4 have a run length j=2 that occurs one time in the image. The positions (2,2) and (4,2) are therefore both equal to one.

The Gray Level Run Length Matrix (GLRLM) quantifies the number of voxels with the same intensity value in row at a given angle θ. The position (i, j) in figure 5 represents the number of times a gray level intensity value i occurs consecutively in a run length j at the angle $\theta = 0$ (horizontally) in the image. The intensity values $i = 2$, 3 and 4 all have one run of length $j = 2$. The size of the GLRL matrix is $i \times j$.

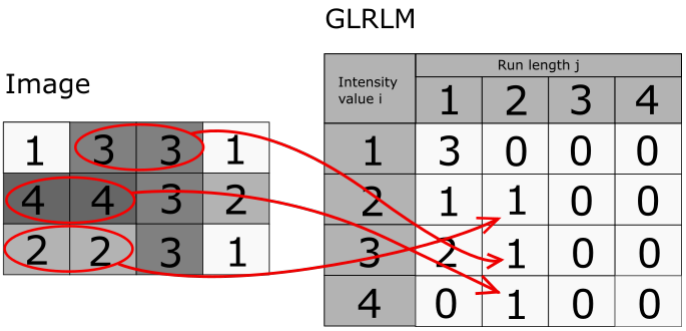

Figure 5: An example of how the GLRLM is calculated. The intensity values (i) 2, 3 and 4 have a run length j=2 that occurs one time in the image at the angle θ=0 (horizontally), the positions (2,2), (3,2) and (4,2) are therefore all equal to one. The size of this GLRL matrix is 4×4.

The Neighboring Gray Tone Difference Matrix (NGTDM) assesses the difference between the center voxel intensity value and the mean of the neighboring pixels with a distance δ. Features extracted from this type of matrix includes coarseness, busyness and complexity [5].

As seen in figure 6, the NGTDM consists of intensity values i and n_i voxels with intensity value i. p_i is the probability of a voxel having intensity value i. s_i corresponds to the sum of the absolute difference between intensity value i and the mean intensity value of the neighboring voxels. The size of the NGTD matrix is $i \times 3$, where each column corresponds to n_i, p_i and s_i.

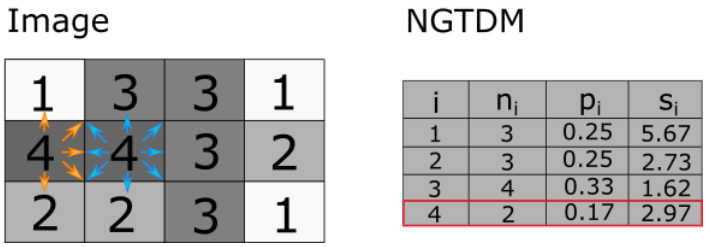

Figure 6: An example of how the NGTDM is calculated. n_i is the number of voxels with intensity value i. p_i is the probability of a voxel having intensity value i. s_i is the sum of absolute differences between the intensity value i and the mean intensity values of it's neighboring voxles. The example figure illustrates how NGTDM is calculated for intensity value $i = 4$ with a distance $\delta = 1$. Two voxels contains the intensity value 4, therefore $n_4 = 2$. The probability p_4 is calculated by $\frac{n_4}{Total\ number\ of\ voxels} = \frac{2}{12} = 0.17$. $s_4 = \left|4 - \frac{1+3+4+2+2}{5}\right| + \left|4 - \frac{4+1+3+3+3+3+2+2+}{8}\right| = 2.97$. The size of this NGTD matrix is 4×3.

The Gray Level Dependence Matrix (GLDM) is also calculated based on the relationship between a center voxel and its neighboring voxels [5]. A neighboring voxel with intensity value j is dependent on the center voxel with intensity value i if $|i - j| \leq \alpha$. The $(i, j)^{th}$ element in the GLDM is how many times a voxel with intensity value i with j dependent voxels in it's neighborhood appears in the image. In figure 7 there is an example of a GLDM that has been calculated with an $\alpha = 0$ and a distance $\delta = 1$ between voxels. The intensity value $i = 3$ has only one occurrence of a neighborhood with three dependent voxels. The size of the GLD matrix is $i \times j$.

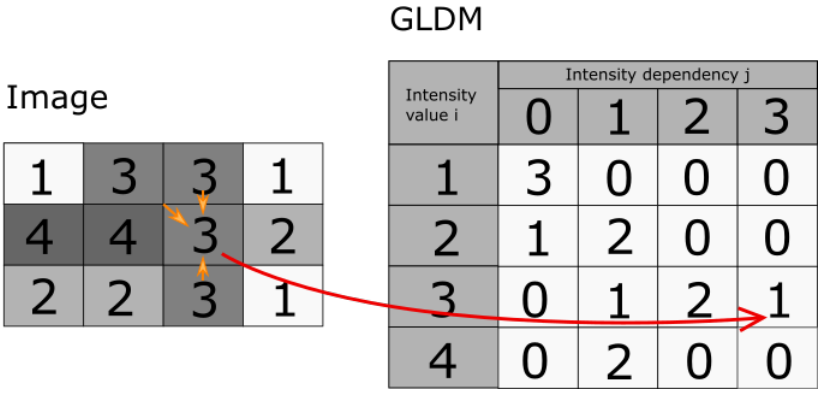

Figure 7: An example of how the GLDM is calculated. A neighbor voxel with distance $\delta = 1$, is considered dependent if $|i - j| \leq \alpha$, where i is the intensity value of the center voxel and j is the intensity value of the neighboring voxels, in this case $\alpha = 0$. The position $(3, 4)$ in the GLDM corresponds to how many times an intensity value $i = 3$ occurs on a neighborhood with three dependent voxels, $j = 3$. The total size of the GLD matrix is 4×4.

2.4 Repeated Elastic Net Technique for Feature Selection

Radiomics features often yield a large number of columns, while the sample size is often quite low. A feature selection technique is therefore essential to extract a smaller subset of the dataset. A successful feature selection yields a dataset with variation that is useful and contains less bias. The Repeated Elastic Net Technique (RENT) [18] was applied in this thesis for feature selection.

RENT utilizes that regularization in predictive model building can be useful for feature selection [18]. It uses the elastic net regularization for linear or generalized linear model (GLM) as a starting point in order to develop this feature selector approach. Let us consider a GLM,

$$g(y) = \chi\, \beta + \epsilon, \tag{1}$$

Where y is a target variable an χ is the design matrix associated with a dataset X. β is a regression parameter vector (weights). ϵ is the error which is an i.i.d. Gaussian distributed random variable with mean zero. g is known as the link-function. If the link-function is set to the identity mapping then a special case of linear regression model is obtained. The logistic regression for binary classification is a well known version of this model. The link-function,

$g(y) = \log \frac{y}{1-y}$, $y \in [0,1]$, transforms the $[0,1]$-valued target variable onto the real line $\mathbb{R}$. The corresponding inverse logistic function transforms these values into class probabilities in $[0,1]$.

Regularization is achieved by adding a penalty term to the minimized target function during training. There exists different types of penalty terms. Lasso regularization adds a L1 penalty term, which can truncate a part of the parameter to zero [18]. Ridge regression handles multicollinearities by pulling the L2-norm from the parameter vector β towards zero [18]. Elastic net regularization utilizes both the L1 and L2 penalties. This is especially useful for feature reduction in high dimensional datasets.

Elastic net can handle a large amount of correlated features and can simultaneously truncate parts of the parameters to zero [18]. Formally the elastic net method consists of a L1 term, $\lambda_1(\beta) = abs(\beta)$ and a L2 regularization term, $\lambda_2(\beta) = \|\beta\|_2$. The regularization term for the elastic net is formulated as,

$$\lambda_{enet}(\beta) = \gamma \left[\alpha\lambda_1(\beta) + (1-\alpha)\lambda_2(\beta)\right], \tag{2}$$

where $\alpha \in [0,1]$ works as a mixing parameter and γ decides the strength of the regularization. These parameters are adjusted by the user by using the input parameters c and $l1ratio$. $l1ratio$ is the mixing parameter, α, which mixes between the L1 and L2 regularization. The input parameter c is the inverse values of γ.

Given a training set, $\{X_{train} = x_i : I = 1, \ldots, I_{train}\}$, where x_i denoted an element from the N-dimensional feature space. RENT is built on the concept of training models on every $k = 1, \ldots, K$ randomized subsets of $X^k_{train} \subset X_{train}$ that is extracted independently and without replication. While the training is done on a unique subset, X^k_{train}, of the original training data X_{train}, the evaluation of every model M_k is done on the remaining samples. In other words the validation set $X^k_{val} = X_{train} \backslash X^k_{train}$, here $\backslash$ denotes the difference operator. To further increase robustness, RENT also enables the user to vary the size of the training sample X^k_{train} in proportion to the size of the entire training set X_{train}.

For every feature f_n in the total feature set F, $n = 1, \ldots, N$ of X_{train}, the trained models will give evidence for the distribution of the associated parameter values, the estimated parameters, β_{nk} from the model $M_k, k = 1, \ldots, K$. These estimates can be collected into a parameter matrix B with dimensions

$N \times K$, where every row represents the estimated parameter distribution for a feature f_n over all K trained models. RENT selects features by extracting the parameter estimations from B for every feature f_n, denoted by β_n. Since the models are trained on different samples of the training data, X_{train}, the parameter estimates β_n for feature f_n vary across the K-models, where some of the parameter estimates can have been set to zero due to the L1-regularization term as a part of the elastic net regularization [18].

A simple way to measure feature relevance is to count the number of times a features is selected across all K models through the counter $c(\beta_n)$, which in other words can be described as the ratio of non-zero parameter estimates for feature f_n,

$$c(\beta_n) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{[\beta_{n,k} \neq 0]}. \tag{3}$$

Two empirical summary statistics can be observed in the feature parameter estimate distributions, β_n. These are the feature specific mean $\mu(\beta_n)$, and the variance $\sigma^2(\beta_n)$ is defined as:

$$\mu(\beta_n) = \frac{1}{K} \sum_{k=1}^{K} \beta_{n,k} \tag{4}$$

$$\sigma^2(\beta_n) = \frac{1}{K_1} \sum_{k=1}^{K} (\beta_{n,k} - \mu(\beta_n)^2 \tag{5}$$

To RENT a feature f_n, is generally viewed as a candidate for feature selection if these criteria's are met:

- The feature f_n, has a high score $c(\beta_n)$ (Eq. 4), in other words the feature has been selected by the elastic net in a large proportion of the K models.
- The feature f_n, is stable. The parameter estimates do no alternate between positive and negative signs throughout the K models.
- The feature f_n, has consistently high non-zero model parameter estimates with low variance across the K models.

These three criteria's can be transformed into mathematical expressions in order to produce three quality metrics, $\tau_i(\beta_n)$, $i = 1, \ldots, 3$, in order to assess the feature, f_n.

$$\tau_1(\beta_n) = c(\beta_n) \tag{6}$$

$$\tau_2(\beta_n) = \frac{1}{K}\left|\sum_{k=1}^{K} sign(\beta_{n,k})\right| \tag{7}$$

$$\tau_3(\beta_n) = p_{K-1}\left(\frac{\mu(\beta_n)}{\sqrt{\frac{\sigma^2\beta_n}{K}}}\right) \tag{8}$$

Here p_{K-1} is the cumulative density function of students t-distribution witk $K-1$ degrees of freedom.

The optimal case for τ_2 is when all the parameter estimates have the same sign (positive or negative), but unfortunately this is not true for most cases, with exceptions of very small K. τ_2 decides therefore the ratio of parameter estimates that has the same sign.

$\tau_3(\beta_n)$ identifies consistently high parameter estimates, and is chosen such that it corresponds with the statistical t-test with a rejection of the null hypothesis.

$$H_0 : \mu(\beta_n) = 0 \tag{9}$$

If the null hypothesis holds, the test statistics

$$T = \frac{\mu(\beta_n)}{\sqrt{\frac{\sigma^2(\beta_n)}{K}}}, \tag{10}$$

will follow a students t-test distribution with $K-1$ degrees of freedom. This term evaluates the probability of test statistics under the H_0 distribution and yields a threshold for the chosen significance level.

In order to define a criterium for feature selection from these quality metrics, $\tau_1(\beta_n)$, $\tau_2(\beta_n)$ and $\tau_3(\beta_n)$, RENT introduces the corresponding threshold values $t_1, t_2, t_3 \in \mathbb{R}^+$. A feature $f_n \in F$ is put in the selected features $\mathcal{F}$ if it satisfies all the criteria's $\tau_i \geq t_i, \forall i \in \{1, 2, 3\}$. These quality metrics can therefore be viewed as hyperparameters in the RENT method, which gives the user the opportunity to tune feature selection by changing the values of the hyperparameters t_1, t_2, t_3 . The number of features decreases as any of these thresholds are increased.

$\tau_1(\beta_n)$, $\tau_2(\beta_n)$ and $\tau_3(\beta_n)$ are constrained within the interval $[0, 1]$. $\tau_3(\beta_n)$ represents a t-test which means that a threshold $t_3 = 0.975$ yield a significance level equal to 5% in the t-test [19].

2.5 Principal Component Analysis

Principal Component Analysis (PCA) is used in this thesis to detect clusters, trends and outliers in the dataset. PCA is an unsupervised dimensionality reduction technique that is widely used across different fields [20]. The method reduces dimensionality by finding the direction with maximum variance in a high-dimensional dataset and projecting the data onto a new feature space with fewer or equal number of dimensions. PCA will construct new axes, called principal components (PC) that lie along the direction of greatest variation but orthogonally to the other principal components. The principal components will then make up the new subspace.

PCA is preformed by constructing a $d \times x$-dimensional transformation matrix W. This transformation matrix enables us to map a vector x with d-dimensional feature space onto a new k dimensional feature space, where typically $k << d$. The vector x can be expressed as,

$$x = [x_1, x_2, ..., x_d], \;\; x \in \mathbb{R}^d. \tag{11}$$

The transformation matrix,

$$W \in \mathbb{R}^{d \times k}, \tag{12}$$

is constructed from the covariance matrix of x. The covariance matrix is decomposed into eigenvalues and eigenvectors. The eigenvectors are then sorted by decreasing order and the corresponding eigenvectors are chosen based on the k largest eigenvalues, where k is the dimensionality of the new feature space.

The transformation matrix W is used to transform x,

$$x\, W = z, \tag{13}$$

to an output vector z where,

$$z = [z_1, z_2, ..., z_k], \;\; z \in \mathbb{R}^k. \tag{14}$$

The elements of the output z are often referred to as scores, while the elements of W is known as loadings. These can be meaningful to plot in order to detect outliers, clusters or which features are affects the observations.

All components will have the largest variance possible given that the principal components are uncorrelated to each other (orthogonal).

2.6 Pearson's correlation coefficient

This thesis uses correlation matrices to investigate whether patients and controls correlate. Correlation matrices enables us to summarize linear relationships between variables [20]. This can be achieved by using the Pearson's correlation coefficient [20]. The Pearson product-moment correlation coefficient also referred to as Pearson's r measures the linear dependence between two features. It is defined as the covariance of the two features, σ_{xy}, divided by the product of the standard deviations of the two features, σ_x, σ_y. The Pearson's r is a number that ranges between 1 and -1, and is a measure of how strong the linear dependency is. If $r = 1$, there is a perfect positive correlation between the two features, if $r = 0$ there is no correlation between the two features and if the $r = -1$, there is a perfect negative correlation between the features. The Pearson's correlation coefficient can be calculated for features x and y with length n,

$$r = \frac{\sum_{i=1}^{n} \left[\left(x^{(i)} - \mu_x \right) \left(y^{(i)} - \mu_y \right) \right]}{\sqrt{\sum_{i=1}^{n} \left(x^{(i)} - \mu_x \right)^2} \sqrt{\sum_{i=1}^{n} \left(y^{(i)} - \mu_y \right)^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{15}$$

Here, μ_x and μ_y are the means of the respective features, σ_x and σ_y are the standard deviations of the features x and y, and σ_{xy} is the covariance between the two features.

2.7 Accuracy metrics

Three accuracy metrics were used in thesis in order to calculate the performance of the model, the F1 score, the accuracy and Matthews correlation coefficient.
The accuracy is intuitively how many times the model predicts correctly with respect to how many samples it predicted [21]. This metric does not take into account the ratio of false positive predictions and false negative predictions given by the model. Whenever the model predict a depressed patient as being depressed the prediction is a true positive (TP). When the model predicts a control as depressed, the prediction is a false positive (FP). Similarly, a control predicted to be healthy would be a true negative (TN), while a patient predicted as healthy would be a false negative (FN).

The accuracy would be calculated as,

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

Precision is another accuracy metric that measures how many of the depressed patients were predicted correctly, calculated as,

$$P = \frac{TP}{TP + FP}. \tag{17}$$

Recall measures the ratio of how many patients that were depressed compared to the number of individuals that were predicted as depressed. Recall is calculated by,

$$R = \frac{TP}{TP + FN}. \tag{18}$$

Precision and recall enable us to expose models that seems to have high accuracies but are in fact failing to predict most depressed patients. If an unbalanced dataset had contained a small fraction of patients, then the model could predict all individuals as controls and still achieve a high accuracy. Therefore it is good practice to always include other accuracy metrics.

The F1 metric is the weighted average of the precision and recall scores, calculated by,

$$F1 = \frac{2PR}{P + R}. \tag{19}$$

Another accuracy metric the Matthews correlation coefficient (MCC), which ranges between -1 and 1, where 1 is a perfect prediction and -1 is a totally incorrectly predicted [22]. In precision, recall and F1, the TN is not a part of the equations, meaning that the TN can be any number and would not affect these accuracy metrics [23]. MCC takes TN into account and treats the classes symmetrically, as can be seen in equation 20.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{20}$$

3 Method and materials

3.1 The dataset

This thesis examined T1 weighted Magnetic Resonance images of the brain. The images originated from the DELPHI study which was conducted the Netherlands from July 2005 to February 2007 [10]. The study consisted of 22 patients (male and female aged 25-55) diagnosed with DSM (fifth edition) defined MDD. The patients scored > 18 on the 17-item Hamilton Depression Rating Scale (HDRS17) [11]. The patients were scanned at study entry and after 6 and 12 weeks of paroxetine treatment. Each patient was matched with a control of the corresponding sex and age ($\pm$ 2.5 y). The controls were scanned only at study entry.

The patients received open-labeled 20mg/d paroxetine for 6 weeks. For the next 6 weeks the non-responders was randomly (stratified by age) assigned to either a true dose escalation (paroxetine $30 - 50mg/d$) or a placebo escalation added to paroxetine $10mg/d$. Non-responders were categorized as the patients that had less than 50% decrease in HDRS17 score.

Structural T1 weighted MR images of the 22 patients and 21 controls from this study were received from the Oslo University Hospital. Each image consisted of 256×256 pixels that were stacked in 182 slices, where each slice is a cross section of the brain. Four patients were removed from further study because they did not complete the entire treatment, 18 patients were further studied. Mainly, this thesis focused on the patients and controls at study entry. The two other timesteps were not investigated due to time constraints and because our main focus was diagnosing patients, not predicting treatment response. A flowchart of the method implemented in this thesis can be seen in figure 8, and is further described in the oncoming sections.

Figure 8: This figure displays a flowchart of the method applied in this thesis. The MR images were sent to RadiomPipe, which segmented the images. The image segments were further normalized and discretized, and radiomics features were calculated from these segments, resulting in a dataset. The dataset was split into a training and test set three times called split 1, split 2, and split3. RENT was applied to each one of these splits, where it selected a subset of the radiomics features. A PCA was conducted on the selected features from each of these splits. A PCA was also conducted on the features that were selected from all three splits. Further, a PCA was also conducted on the entire dataset and selected features corresponding to the brain regions, hippocampus, and anterior cingulate.

3.2 RadiomPipe

The T1 weighted images were sent to an Application Programming Interface (API) called RadiomPipe [24] developed to segment the brain into brain regions, and to normalize and discretize the images. These images were then used to extract radiomics features. RadiomPipe is a Representational State Transfer (REST) API that is designed to make feature extraction easier and more standardized [24]. The MRI scans are sent to RadiomPipe as a dictionary, which is passed on to another REST API called BrainSeg. BrainSeg segments the brain into approximately 97 structures where each mask represents an anatomical section of the brain, that follows the FreeSurfer (Martinos Center for Biomedical Imaging, Harvard-MIT, Boston USA) standards [25]. Mask images are then passed back to RadiomPipe. The images are normalized and discretized, and then the masks are used to calculate 110 different radiomics features per mask.

3.3 Normalization

The intensity values of a MRI image vary between protocols, scanners, patients and visits [6]. This poses problems in regards to image segmentation and extraction of radiomics features. In order to compare MR scans from different patients it is essential to standardize over all patients and visits. The basic idea behind standardization is to change the intensity values in the target image to new intensity values by a transformation function [26]. The easiest way to do this is by correcting each intensity value by an offset value. As Collewet et al. (2004) [27] displayed this can be done by normalizing using the same maximum gray level or same mean for all images. These techniques are multiplicable and therefore keep the relative variations in gray levels.

Another common approach is the z-score method, which does not keep the relative variation between gray levels. The z-score method is calculated by subtracting the mean intensity value (μ_{ROI}) of the region of interest (ROI) from each voxel intensity ($I(x)$) and dividing this result by the standard deviation of the ROI (σ_{ROI})[28]:

$$I_{Zscore}(x) = \frac{I(x) - \mu_{ROI}}{\sigma_{ROI}} \tag{21}$$

This is computed for each ROI and every patient, at each visit. A z-score represents how many standard deviations the voxel intensity differs from the mean intensity of the ROI. The calculated z-scores therefore have zero mean and unit standard deviation.

Carré et al. (2020) [28] investigated the z-score method as well as Nyul's harmonizations method, which is a piecewise linear histogram matching that maps the intensity values of each image to a standard histogram, and a white stripe method which utilizes the z-score method based on normal appearing white matter. Nyul's harmonization method led to a high number of robust first-order features. However, it has been shown that this linear histogram matching affects the texture in the image. The white stripe method is dependent on the quality of the white matter segmentation, which may affect the quality of the normalization. Carré et al. (2020) recommended the z-score method as it is easy to implement, computationally efficient and is more robust as it considers all the voxels inside the ROI. The z-score method was applied in this thesis based on Carré et al. (2020) [28] and Collewet et al. (2004) [27] recommendations.

3.4 Discretization

Texture features were extracted based on co-occurrence matrices and other derived matrices calculated using the gray-level intensity values of the ROI [29]. Large numbers of intensity values yield large co-occurrence matrices, which are computationally heavy and result in texture features that are difficult to reproduce. To handle this problem discretization can be implemented. Discretization clusters similar intensity values and will reduce the number of individual intensity values. The two most common methods for discretization are, relative discretization which clusters the intensity values to a predefined number of bins, and absolute discretization which uses a fixed bin width [29]. When a relative discretization is applied, the ROI's intensity range will affect the bin sizes, which will in turn affect the extracted radiomics features. Bin sizes in the absolute discretization are independent of the intensity range and may result in lower variability in the extracted radiomics features between patients.

Goya-Outi et al. (2018) [6] recommended constant bin width with relative bounds as it is a simple method that does not require setting absolute limits for upper and lower bounds. Based on Goya et al. (2018) [6] and Duron et al. (2019) [29] recommendations, this thesis implemented constant bin width equal to 5 with relative bounds. The T1 weighted MR images were scaled to $0 - 255$ gray scale values and then grouped into bin widths equal to 5. Yielding about 51 gray scale values. Discretization is preformed after the normalization is conducted.

3.5 Dataset structure

The patients' images were sent to RadiomPipe which returned a dataset structured as shown in figure 9 after four patients were removed due to not completing the treatment. The patient column specify the patient id. Each patient was scanned at three different timesteps. Timestep t0 corresponds to the scan that was preformed at study entry. Timesteps t1 and t2 corresponded to 6 and 12 weeks after study entry, respectively. The column timestep indicates what timestep this information corresponds to. There was a total of 54 images in the patient dataset as each of the 18 patients were scanned three times. Each image was separated into 95 brain regions, the column Mask indicates what brain region the current information corresponds to. There were calculated 107 radiomics features columns for each patient at every timestep and for all the 95 brain regions. Resulting in a dataset with $18\,patients \times 3\,timesteps \times 95\,brainregions = 5130\,rows$, and 110 columns.

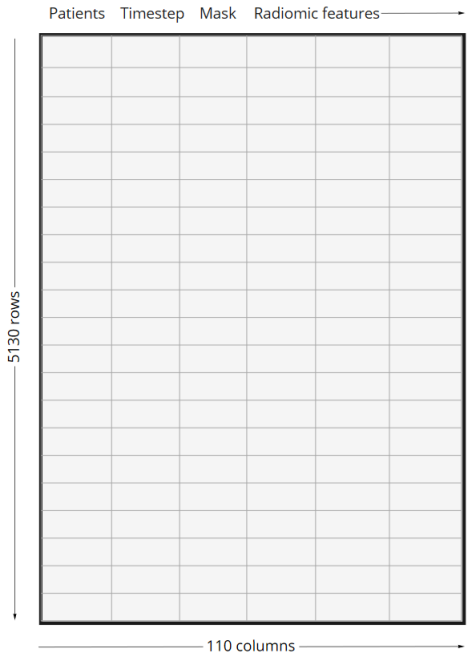

Figure 9: An overview of the dataset received from RadiomPipe after four patients was removed due to not finishing the treatment. The patients column corresponds to the patients id, the timestep column corresponds to the timestep at which the patient was scanned. Mask indicates what brain region the current information corresponds to.

To sort the data, three new datasets were constructed for each timestep as seen in figure 10. Here each patient will occur in 95 rows, once for each brain region. In order to make the dataset readable for algorithms like RENT [18] and PCA [20], it is favorable to have one row per patient. This was done by making new names for the columns corresponding to certain masks. Instead of having one patient occurring 95 times because of each mask, each patient will have 107 radiomics features per mask each beside each other.

The control dataset was structured and processed the same way, but the controls were only scanned at t0. In this study, the main focus was on the t0 timestep for the dataset. The patients dataset at timestep t0 and the controls dataset were concatenated to create one large dataset with 39 rows containing both patients and controls at t0. Some analyses were done on timesteps t1 and t2; the dataset of patients at these timesteps, respectively, were concatenated with the control dataset in the same manner as at timestep t0.

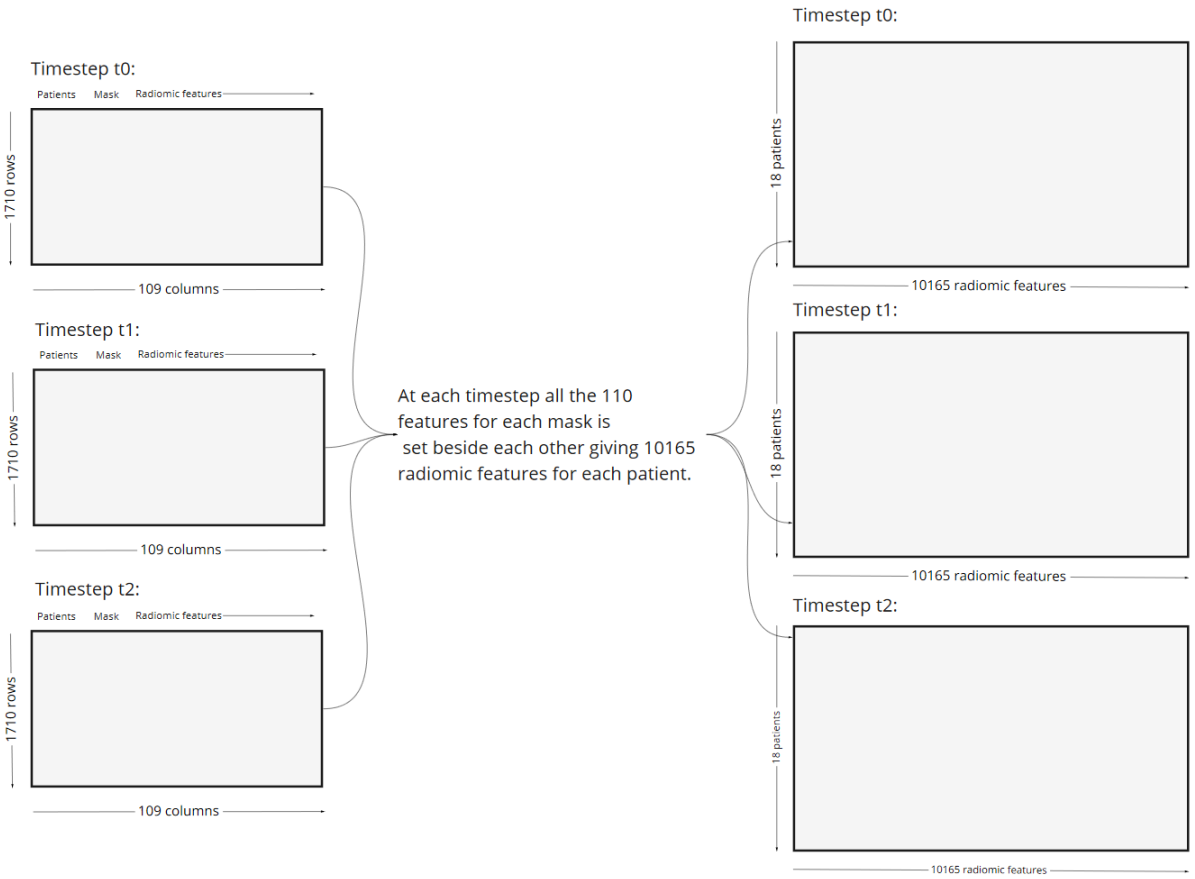

Figure 10: This figure shows how the dataset received from RadiomPipe was split into three datasets, one for each time step. Further it was rearranged such that the 107 radiomics features for each mask was set beside each other, creating one row for each patient with 107 columns per mask.

Some analyses were conducted on a subset of the dataset with patients and controls at timestep t0. This subset contained a selection of masks that corresponded to the brain regions, hippocampus, and anterior cingulate. These brain regions were chosen based on them being predictive for diagnosis and treatment response for MDD, as described in section 2.1. Therefore, this subset only contained features with one of the mask labels 17, 53, 1002, 1026, 2002, and 2026. These mask labels correspond to the FreeSurfer (Martinos Center for Biomedical Imaging, Harvard-MIT, Boston USA) standards [25] for the brain regions, hippocampus, and anterior cingulate.

3.6 Correlation matrices

Correlation matrices were used in this thesis to get a general view of how much the patients and controls correlate. The correlation matrices were calculated by using the Pandas function `DataFrame.corr()` [30] which yields a table of Pearson's correlation coefficients. The Pearson's correlation coefficient is further explained in section 2.6. The table of correlations were displayed using the Seaborn function `heatmap()` [31]. The correlation function in pandas calculates the correlation between the columns. The dataset were therefore transposed and patients and controls were transformed into columns, before calculating the correlation.

3.7 Repeated Elastic Net Technique

This study utilized the Repeated Elastic Net Technique (RENT) algorithm in order to reduce the feature space. RENT is especially effective for short wide datasets, meaning that the dataset has a small amount of samples and a large number of columns, which often is the case when extracting radiomics features. This thesis utilised RENT version 0.0.1 in python version 3.8.8. RENT is as package that has a number of useful functions that help us analyse the selected features and ensemble of models, as well as giving insight into the dataset. This thesis used RENT to train an ensemble of 100 models to predict what features are relevant to classify patients diagnosed with major depressive disorder. The response vector was arranged such that the patient corresponded to a response equal to one while the controls had a response equal to zero.

3.7.1 Splitting the data

A challenge with a short wide dataset is that the individuals that are placed in the test set can have a large impact on the model's prediction. In order to resolve this problem, ensuring robustness and reliable results, the data was split into a training set and a test set, three times, stratified by if the individual belonged to the patient or control class. This enables the feature selector approach to select features based on different individuals in the training set. An example of how this can be divided is seen in figure 11. If the approach produces similar results for all splits, then there is a higher probability that the results are reproducible. Each split is hereby referenced as split 1, split 2, and split 3. An overview of how the samples were split can be seen in Appendix A.

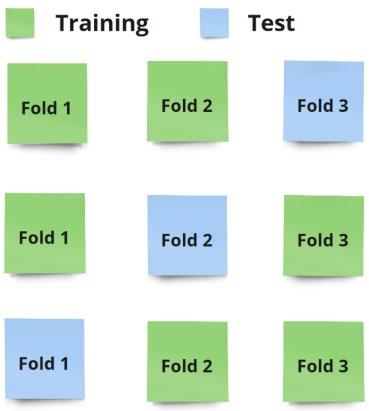

Figure 11: An overview of how the dataset was split into a training and validation set three times.

3.7.2 Determining the regularization parameters for RENT

In order to decide the combination of c and $l1$ *ratio* parameters for RENT that give the best model performance, possible c and $l1$ *ratio* parameters are stored in lists. The lists of possible c and $l1$ *ratio* parameters were, $c = [0.01, 0.1, 1, 10, 100]$ and $l1$ *ratio* $= [0, 0.1, 0.25, 0.5, 0.75, 0.9, 1]$, respectively. The RENT algorithm [18] computes the performance with a 5-fold cross validation tuning the given c and $l1$ *ratio* parameters. These computations can differ slightly if run multiple times. In order to be confident in the selection of the regularization parameters, the shuffled dataset with a 70% training set was run six times into RENT. The results can be viewed in appendix B.

RENT produces three matrices for the results of the cross-validations of the combinations of parameters. Dataframe 1 shows the average predictive performance. The highest score yields the parameter combination with the highest predictive performance. Dataframe 2 shows the average percentage of how many feature weights were set to zero, in other words how strong the feature selection was with the corresponding parameter combination. Dataframe 3 shows the harmonic mean between the first two dataframes. The highest score in Dataframe 3 yields the best parameter combination. An example of an output of these three dataframes can be seen in figure 12. Although there might be parameter combinations that yield high performance, they might have a weak feature selection. Then it can be advantageous to reduce the performance and rather use a parameter combination that has a stronger feature selection. For all the six runs of RENT, the decision of

parameter combination was inspected by comparing these three dataframes. From this analysis the parameter combination was determined to be $c = 0.1$ and $l1\,ratio = 0.5$ and this parameter combination was used for all the splits.

Dataframe 1: Average predictive preformance

l \ *c*	0,01	0,1	1	10	100
0	0,7387	0,7387	0,7387	0,7387	0,7387
0,1	0,6395	0,8748	0,8162	0,7387	0,7387
0,25	NaN	0,8748	0,8748	0,7387	0,7387
0,5	NaN	0,9333	0,8748	0,8162	0,7387
0,75	NaN	0,8162	0,8748	0,8162	0,7387
0,9	NaN	0,6978	0,8748	0,8162	0,7387
1	NaN	0,6395	0,8748	0,8162	0,7387

Max performance: 0,9333

Dataframe 2: Average precentage of feature weigths set to zero.

l \ *c*	0,01	0,1	1	10	100
0	0,0356	0,0356	0,0356	0,0356	0,0356
0,1	0,9997	0,8648	0,3724	0,0686	0,0358
0,25	NaN	0,9493	0,6152	0,1344	0,0381
0,5	NaN	0,9829	0,7642	0,2245	0,0472
0,75	NaN	0,9958	0,8312	0,3019	0,0576
0,9	NaN	0,9987	0,8559	0,3446	0,0646
1	NaN	0,9997	0,8687	0,3732	0,0686

Average percentage: 0,9997

Dataframe 3: Harmonic means between Dataframe 1 and Dataframe 2.

l \ *c*	0,01	0,1	1	10	100
0	0,0000	0,0000	0,0000	0,0000	0,0000
0,1	0,0000	0,8293	0,4419	0,0621	0,0004
0,25	NaN	0,8680	0,6867	0,1572	0,0051
0,5	NaN	0,9913	0,7775	0,2956	0,0233
0,75	NaN	0,7498	0,8128	0,3785	0,0429
0,9	NaN	0,3311	0,8250	0,4182	0,0552
1	NaN	0,0000	0,8312	0,4426	0,0621

Harmonic mean: 0,9913

Figure 12: An example of the three dataframes yielded from the 5-fold cross validation parameter tuning method in the RENT algorithm.

3.7.3 Selecting features with RENT

After running RENT for all three splits, the RENT method selected features based on the cutoff values τ_1, τ_2 and τ_3 for each split as mentioned in section 2.5. The bar plot of τ_1, was plotted using the RENT function `plot_selection_frequency()` [19]. This bar plot shows the percentage of times a feature were selected by a model, and were used to decide the cut-off values. The cut-off values were also set such that the RENT method selected as few features as possible while still maintaining a separation of the patient and control grouping in a PCA score plot, resulting in cut-off values equal to $\tau_1 = 0.7$, $\tau_2 = 0.7$ and $\tau_3 = 0.975$.

3.7.4 Describing the individuals in the dataset

To get insight into how the RENT performs on each patient/control the function `get_summary_objects()` [19] was used. The function produces a table that yields information about how many times an individual has been misclassified out of the number of times the individual was a part of the validation set. This table yields interesting information about what individuals may be outliers and possibly adding bias to the model.

3.7.5 Validating the performance of the ensemble of models in RENT

In order to validate the performance of RENT, a function called `plot_elementary_models()` [19] were used, which yields a plot with Matthews correlation coefficient metrics which shows the accuracy of each model. The Matthews correlation coefficient is an accuracy metric and were explained in section 3.8. The plot also shows how high share of features were set to zero by each of the models. Yielding an impression of how strong the feature selection is.

3.7.6 Validation studies

To ensure that the RENT model performed better than a random model, two validation studies were conducted. An example of the output of such a validation study can be seen in figure 13. Validation study 1 marked in blue, will randomly choose the same amount of features as RENT selected from the entire feature space. It will do this many times and calculate the mean model performance for the randomly selected features. Naturally we should expect that the randomly selected feature model should have a lower performance than the model with the RENT selected features. One would

also expect that the number of times the performance by the model with the random features was higher than the model with the RENT selected features should be quite low. If the number of times the performance with random selected features was larger than the performance with the RENT selected features is 4 out of 100 times, then the p-value would become $4/100 = 0.04$. A significance level is set by the user and a one-sided student's t-test is conducted. The t-test says if the null hypothesis can be rejected, which tells us if a model with RENT selected features is significantly better than a model trained on randomly selected features.

Validation study 2 is also built on a one sided student's t-test, where the response of the test data is permutated. By permutating the response several times and collecting the performances RENT calculates the mean performance. It would also here be natural that the model performance using the permuted labels would be lower than when using the correct labels and that the number of performances that was higher than the correct labels performances would be quite low. A p-value would then be calculated as in validation study 1, and a significance level is set by the user. The t-test is conducted and the null hypothesis can be rejected if the p-value is lower than the significance level. If the null hypothesis is rejected then the RENT model performs significantly better on the real labels opposed to the randomly permutated labels.

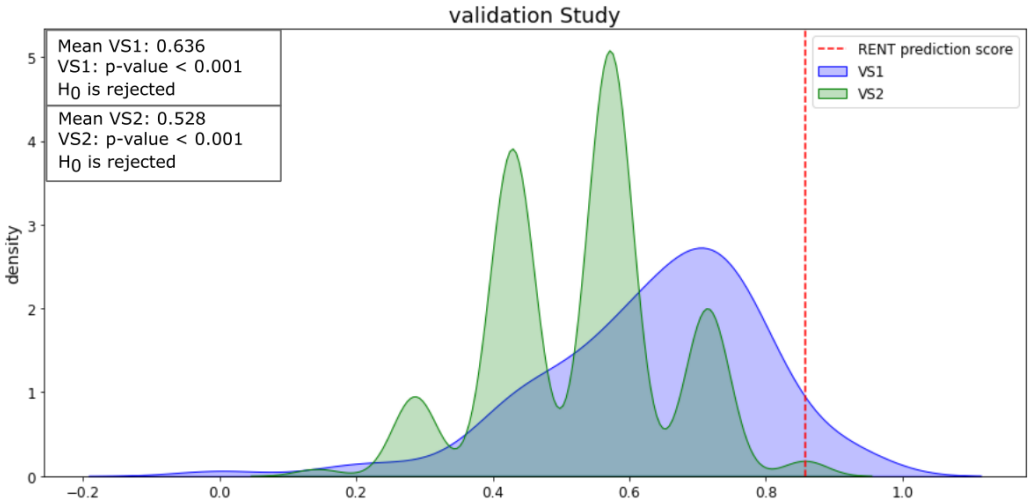

Figure 13: An example of the results from a validation study conducted through RENT. The blue and green graph is the empirical distribution of MCC scores collected from 100 runs in two validation studies, respectively. Validation study 1 (VS1) draws as many random features as RENT selected. Validation study 2 (VS2) randomly permutates the target labels, but keeps the sample features. The red line is the RENT model's MCC score.

3.8 Evaluating the test data

In order to ensure that the RENT algorithm's predictions is not faulty an external logistic regression method was used on the test data for every split. The logistic regression model was applied without regularization and with $c = 1.0$. This model was only given the test samples that were excluded from the training of the RENT model. The model used the features that were selected by RENT in each split. Three accuracy metrics were used in order to calculate the performance of the model, the F1 score, the accuracy and Matthews correlation coefficient. These accuracy metrics are described in section 3.8. Both the F1: 1 and the F1:0 metrics were calculated. These are both calculated as F1 score described in section 3.8. The difference lies in wich class is defined as positive and negative, respectivly. In F1:1, the response for the patients is denoted as one, while the responses of the controls are denoted as zero. In F1:0, the responses are flipped so that patients have the response zero. Using three accuracy metrics enables us be more confident in the estimated performance of the model.

3.9 Principle component analysis

Several principle component analyses were conducted in this thesis, to investigate in what degree the patients and controls classes were separable and what patients/controls deviated from the others. The PCA was conducted using the Hoggorm package vesion 0.13.3 [32]. There was conducted a PCA for every split with the dataset consisting of RENT selected features in the respective splits. A PCA was also conducted using all features in the dataset, to explore how the dataset behaved. To further investigate how the dataset behaved with a smaller selection of features a PCA was conducted for for the brain regions, hippocampus, and anterior cingulate. These are brain regions that are often associated with depression, and is further explained in section 2.1. The selected features from RENT was further investigated by running PCA for features that were selected by all the three splits.

4 Results

The aim of this thesis was to separate depressed patients from healthy controls by using only a subset of the total corresponding radiomics features. Additionally to investigate what radiomics features and corresponding brain regions were predictive for separating the two classes. A correlation matrix was calculated to display how well the individual patients and controls correlate. The dataset was split into training and test set three times. The Repeated Elastic Net (RENT) algorithm was applied to each of these splits to reduce the number of features in the dataset by training an ensemble of 100 elastic net regularized models, and select features based on weight distributions of features across the models. The frequency of how many times the features were selected were investigated. And an analysis of the ensemble models were performed, yielding the average accuracy over all models. Each individual was looked into in order to see how many times it was wrongly classified. Then, a logistic regression was used on the test set to validate the performance of the RENT model. A validation study was also performed to check if the RENT model performed better than a random model. At last the selected features for each split were used in a PCA, where we could indentify possible the clusters of patients and controls and possible outliers. A PCA was also conducted for the whole brain, and a selection of the brain. A PCA was also conducted on the four features that all splits selected.

4.1 Correlation between patients and controls

The correlation matrix of the dataset containing all masks for patients and controls at timestep t0 can be seen in figure 14. There was a large overlap of information among patients and controls. Patient 08 and patient 15 were somewhat less correlated than the rest of the patients and controls.

The correlation matrix calculated from the dataset containing patients and controls at timestep t0, with a selection of masks corresponding to the brain regions, hippocampus and anterior cingulate can be seen in figure 15. In this case, there was a clearer distinction between patients and controls compared to figure 14, where all brain regions where considered. The controls were to an extent more correlated. Still there were some controls like control 027 that had relatively low correlation to the other controls.

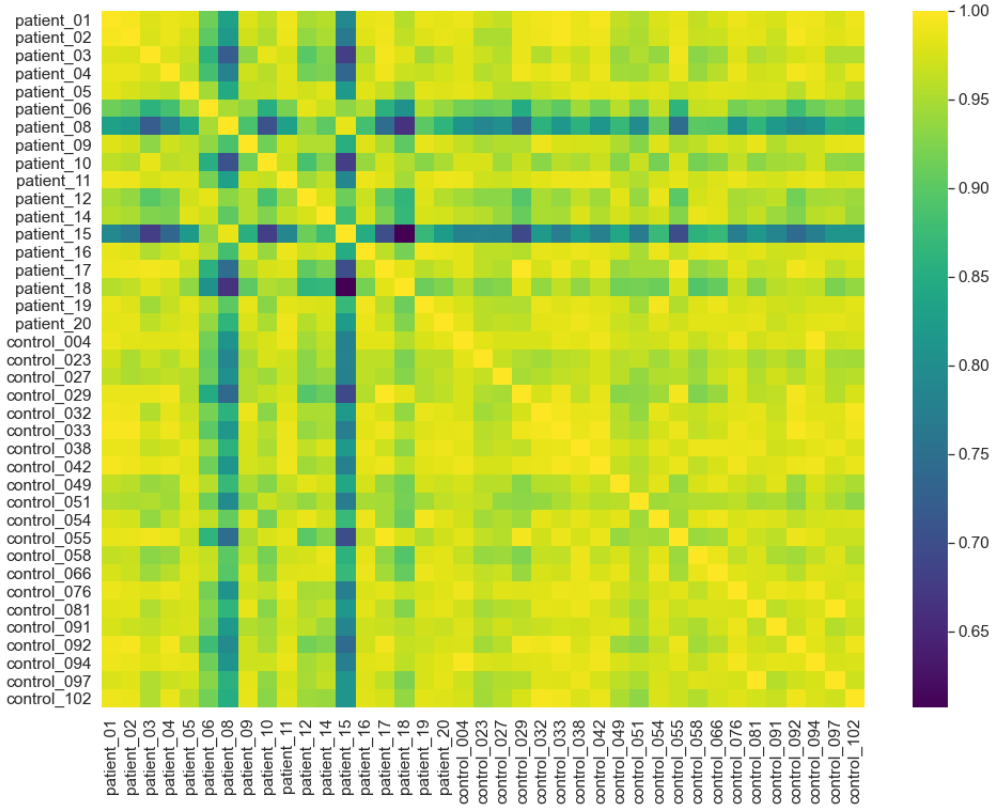

Figure 14: A correlation matrix calculated from the dataset containing features from all masks of patients and controls at timestep t0. The figure displays how closely correlated patients and controls were. The color bar to the right gives the Pearson correlation coefficient.

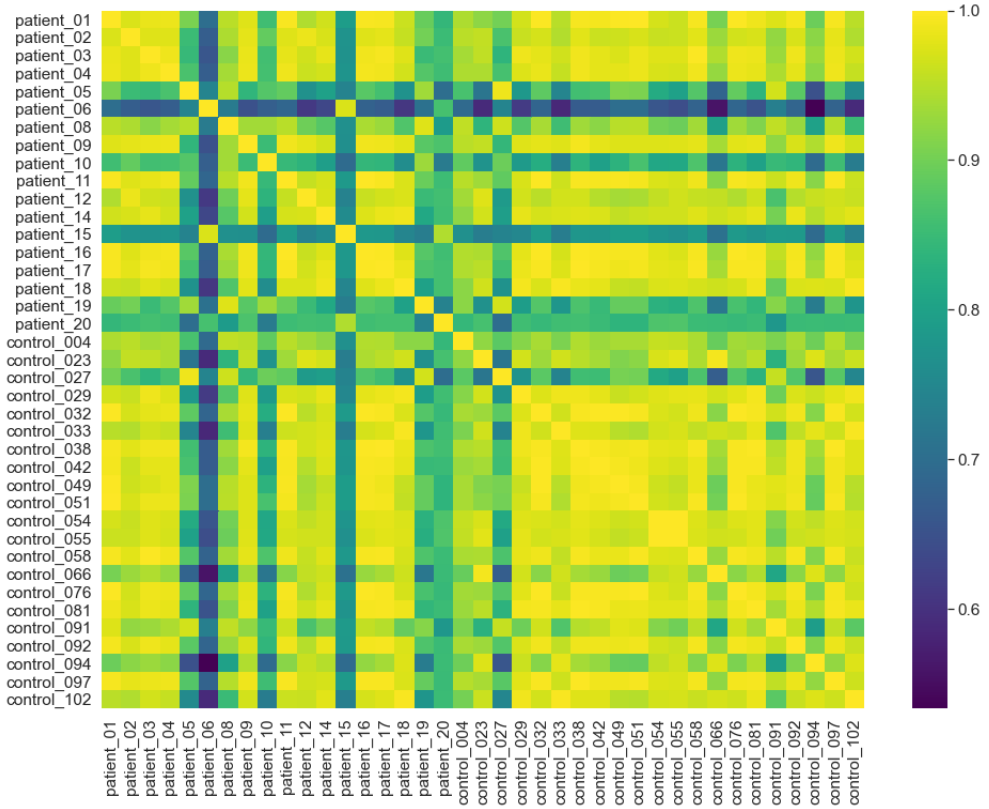

Figure 15: A correlation matrix calculated from the dataset with patients and controls at timestep t0 with a selection of masks corresponding to the brain regions, hippocampus and anterior cingulate. The figure displays how closely correlated the patients and controls were. The color bar to the right gives the Pearson correlation coefficient.

4.2 Feature selection

RENT was applied on the dataset with patients and controls at timestep t0. Figure 16, displays the fraction of the ensemble models in RENT that selected each feature for every split. For all the splits it is quite clear that the majority of features were selected close to 0% of the time. And almost all features were selected in less than 40% of the models. A handful of features were selected in 70% of all models or more.

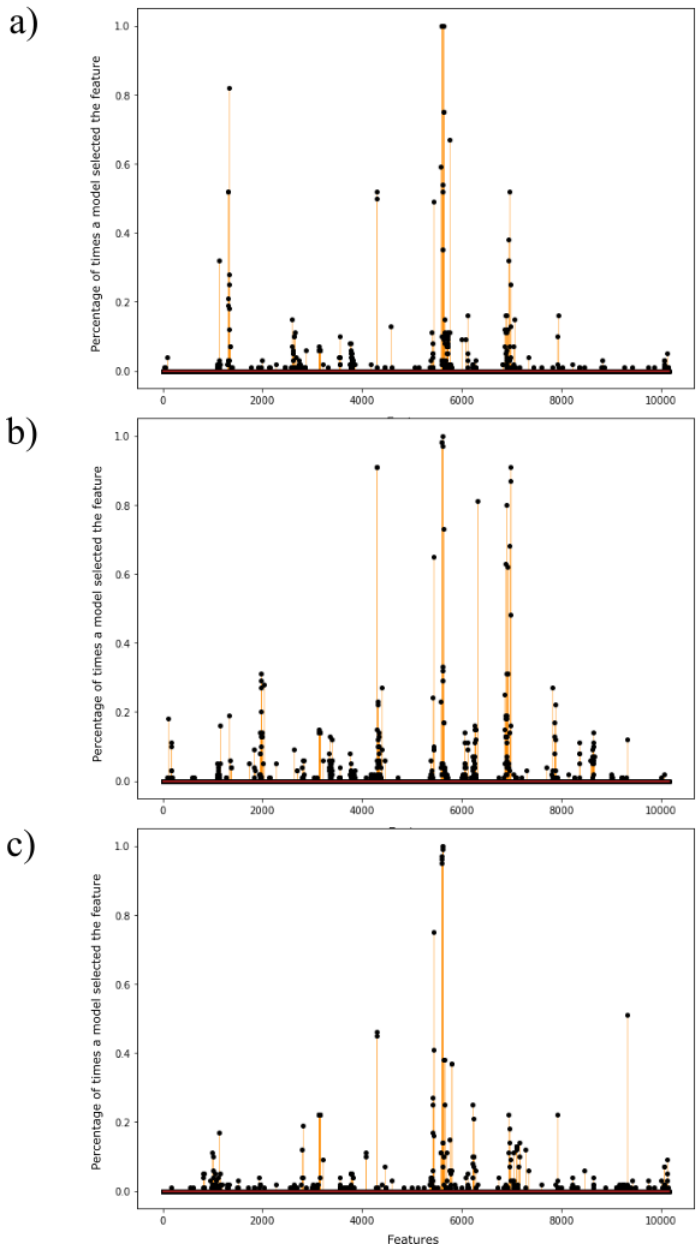

Figure 16: This figure shows the fraction of models that selected a given feature, for split 1 (a), split 2 (b) and split 3 (c).

The number of features that the RENT method selected is based of the cut-off values τ_1, τ_2 and τ_3, previously discussed in section 2.4. The lowest amount of features were selected based on figure 16 and while still maintaining a separation of the patient and control grouping in a PCA score plot. The cut-off values were set to $\tau_1 = 0.7$, $\tau_2 = 0.7$ and $\tau_3 = 0.975$, resulting

in eight features selected in the first split, while the two other splits selected seven features.The features selected by RENT in each split can be seen in figure 17.

	Split 1	Split 2	Split 3
1	glcm Imc1 label 2014	glcm Imc1 label 2014	glcm Imc1 label 2014
2	glcm Imc2 label 2014	glcm Imc2 label 2014	glcm Imc2 label 2014
3	glcm ClusterProminence label 2014	glcm ClusterProminence label 2014	glcm ClusterProminence label 2014
4	glcm ClusterShade label 2014	glcm ClusterShade label 2014	glcm ClusterShade label 2014
5	glrlm GrayLevelNonUniformityNormalized label 2014	firstorder Range label 2	glcm Correlation label 2014
6	glrlm GrayLevelVariance label 2014	firstorder_Maximum_label_2	glcm MCC label 2014
7	glrlm RunEntropy label 2014	firstorder MeanAbsoluteDeviation label 2027	glrlm ShortRunHighGrayLevelEmphasis label 2012
8	glcm Imc2 label 1014		

Figure 17: An overview of the features selected by RENT in each split, respectively.

The frequency of how many times a feature was selected by RENT in one of the splits can be seen in figure 18. The majority of the features were only selected by one split, but four of the features were selected by all three splits. Figure 19 displays the distribution of radiomics feature types. 78.6% of the total selected features from all splits were texture features, while the remaining features consisted of first-order features. No shape features were selected by RENT. The right chart in figure 19 displays the distribution of the type of texture features selected by RENT. Two types of texture features were selected by RENT. 63.6% of all selected texture features were GLCM features, while the rest consisted of GLRLM features. Figure 20 is a pie chart that illustrates the distribution of labels corresponding to brain regions in the 14 features selected by RENT.

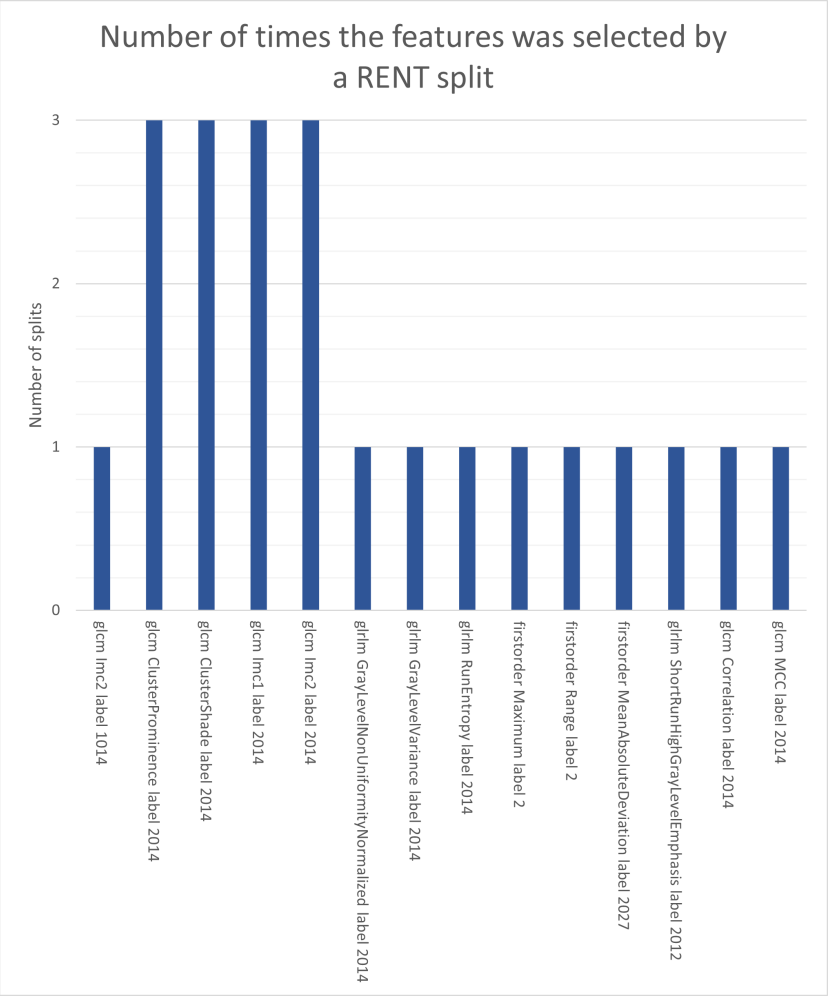

Figure 18: This chart shows how many times each feature was selected by a RENT split. Four features were selected by all three splits

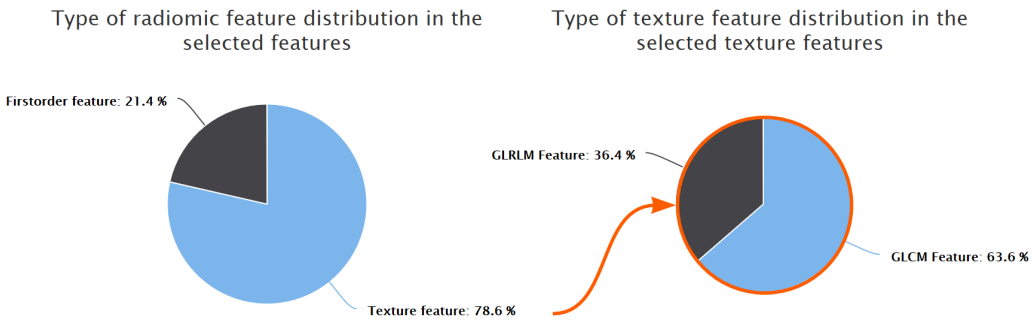

Figure 19: The left pie chart illustrates the distribution of the different types of radiomics features in the features selected by RENT. The right pie chart illustrates the type of texture feature distribution within the texture feature.

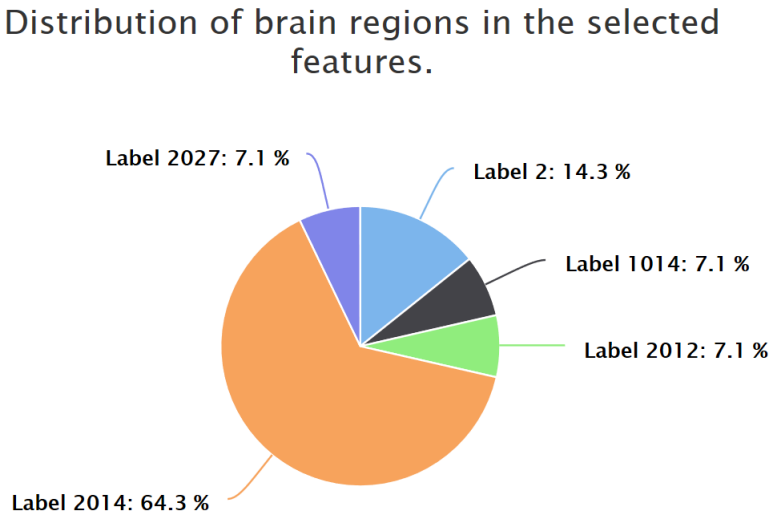

Figure 20: This pie chart illustrates the brain region label distribution in the features selected by RENT.

Figure 21 gives us an overview of what brain regions the labels in the RENT selected features corresponds to. Four of the labels are tied to the orbitofrontal cortex in the brain, while the last one is the left cerebral white matter.

Label 2014	Gray matter of right medial orbital gyrus
Label 2012	Gray matter of right lateral orbital gyrus
Label 2027	Gray matter of anterior part of right middle frontal gyrus
Label 2	White matter of left cerebral hemisphere
Label 1014	Gray matter of left medial orbital gyrus

Figure 21: Overview of labels of the RENT selected features and the corresponding brain regions.

4.3 Performance across models

The MCC accuracies was calculated for the 100 elastic net regularized models in the RENT model for each split. The scores was plotted together with the percentage of weights set to zero, and can be viewed in figure 22. In all three splits the model performance ranged between 0.6 and 1, although for some models the performance was as low as zero.

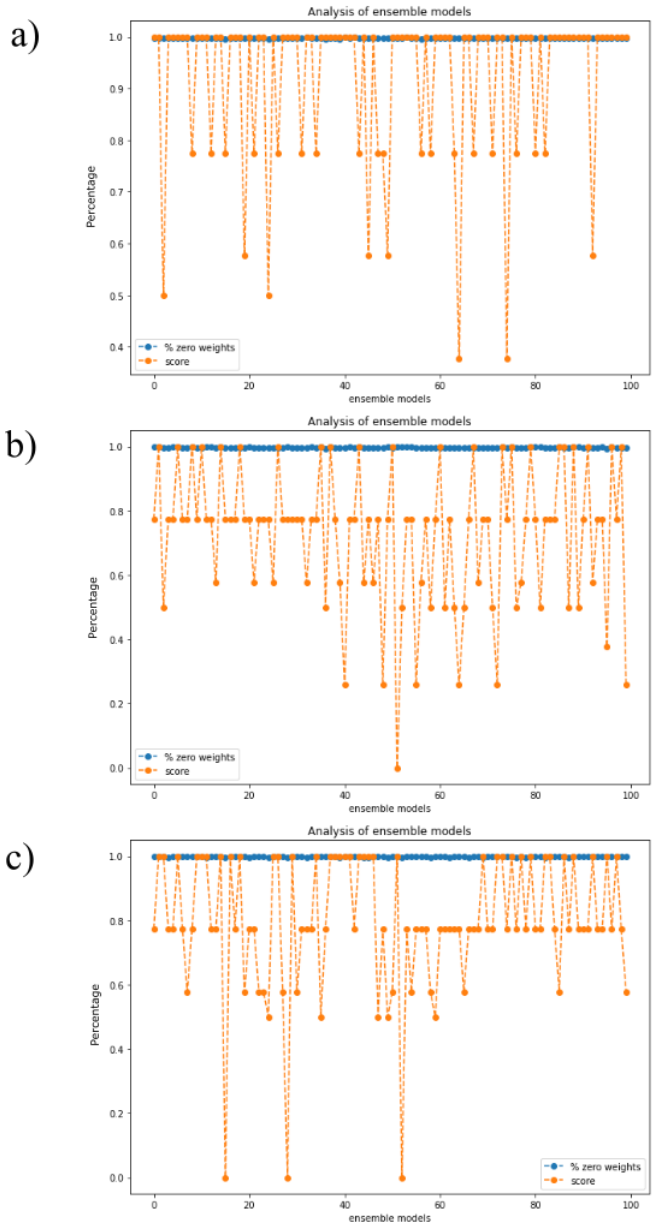

Figure 22: This analysis of ensemble models gives insight into the performance of each model and percentage of weights that were set to zero, in split 1 (a), split 2 (b) and split 3 (c).

4.4 Summary of the individuals

Tables 2−4 displays how often the model predicted each individual correctly, for each split, respectively. They tell us how many times an individual was in the test set, what class label the individual belongs to and how many times the individual was wrongly classified. The percentage of incorrect classifications displays how many times the individual was classified wrongly in relation to how many times the individual was in the test set. A large proportion of the individuals in split 1 was correctly predicted. As seen in table 2, the percentage of wrongly classified individuals was mostly below 14%, except for control 054 which was predicted incorrectly in 50% of the cases.

	# test	class	# incorrect	% incorrect
patient_03	32	1	2	6.25
patient_06	37	1	0	0.00
patient_08	36	1	5	13.89
patient_09	30	1	0	0.00
patient_10	44	1	0	0.00
patient_12	32	1	0	0.00
patient_14	35	1	5	14.29
patient_16	35	1	1	2.86
patient_17	29	1	2	6.90
patient_18	24	1	0	0.00
patient_19	33	1	0	0.00
patient_20	33	1	0	0.00
control_004	28	0	1	3.57
control_023	27	0	3	11.11
control_027	22	0	1	4.55
control_032	36	0	0	0.00
control_042	28	0	0	0.00
control_049	37	0	0	0.00
control_051	30	0	0	0.00
control_054	30	0	15	50.00
control_055	33	0	0	0.00
control_076	27	0	0	0.00
control_081	27	0	0	0.00
control_091	20	0	0	0.00
control_094	29	0	1	3.45
control_102	26	0	0	0.00

Table 2: This table summarizes how well the RENT model predicts the class of patient/control in the test set for split 1. #test indicates how many times the individual was in the test set. Class corresponds to the true class label of the individual. #incorrect gives how many times the individual was classified incorrectly and %incorrect yields the percentage of incorrect predictions in reference to the total number of times the individual was in the test set.

There was generally a larger percentage of incorrectly classified individuals in split 2 than in split 1, as can be seen in table 3. Patient 01 and control 066 were predicted incorrectly in about 80% of the times they were in the test set, while control 058 was predicted incorrectly every time it was in the test set. Half of the time patient 14 was in the test set it was predicted incorrectly.

	# test	class	# incorrect	% incorrect
patient_01	32	1	26	81.25
patient_02	37	1	0	0.00
patient_04	36	1	0	0.00
patient_05	30	1	0	0.00
patient_06	44	1	1	2.27
patient_08	32	1	2	6.25
patient_09	35	1	1	2.86
patient_11	35	1	7	20.00
patient_14	29	1	16	55.17
patient_15	24	1	0	0.00
patient_16	33	1	2	6.06
patient_19	33	1	0	0.00
control_029	28	0	1	3.57
control_032	27	0	0	0.00
control_033	22	0	0	0.00
control_038	36	0	5	13.89
control_051	28	0	0	0.00
control_055	37	0	0	0.00
control_058	30	0	30	100.00
control_066	30	0	26	86.67
control_076	33	0	0	0.00
control_081	27	0	0	0.00
control_091	27	0	0	0.00
control_092	20	0	0	0.00
control_097	29	0	0	0.00
control_102	26	0	1	3.85

Table 3: This table summarizes how well the RENT model predicts the class of patient/control in the test set for split 2. #test indicates how many times the individual was in the test set. Class corresponds to the true class label of the individual. #incorrect gives how many times the individual was classified incorrectly and %incorrect yields the percentage of incorrect predictions in reference to the total number of times the individual was in the test set.

In split 3 there was also a generally low percentage of incorrectly classified individuals, as seen in table 4. Similarly as split 2, split 3 also had difficulties with classifying patient 01 and control 058, incorrectly classifying them in about 50% of the times. Control 027 was incorrectly classified in 86% of the times, although only being classified incorrectly 4.6% of the times in split 1.

	# test	class	# incorrect	% incorrect
patient_01	32	1	15	46.88
patient_02	37	1	0	0.00
patient_03	36	1	0	0.00
patient_04	30	1	4	13.33
patient_05	44	1	2	4.55
patient_10	32	1	0	0.00
patient_11	35	1	5	14.29
patient_12	35	1	0	0.00
patient_15	29	1	0	0.00
patient_17	24	1	2	8.33
patient_18	33	1	1	3.03
patient_20	33	1	3	9.09
control_004	28	0	0	0.00
control_023	27	0	2	7.41
control_027	22	0	19	86.36
control_029	36	0	1	2.78
control_033	28	0	0	0.00
control_038	37	0	13	35.14
control_042	30	0	0	0.00
control_049	30	0	1	3.33
control_054	33	0	6	18.18
control_058	27	0	16	59.26
control_066	27	0	0	0.00
control_092	20	0	0	0.00
control_094	29	0	0	0.00
control_097	26	0	0	0.00

Table 4: This table summarizes how well the RENT model predicts the class of patient/control in the test set for split 3. #test indicates how many times the individual was in the test set. Class corresponds to the true class label of the individual. #incorrect gives how many times the individual was classified incorrectly and %incorrect yields the percentage of incorrect predictions in reference to the total number of times the individual was in the test set.

4.5 Checking performance with a logistic regression model

The performance of a logistic regression model performed on the test set with RENT selected features can be seen in table 5 for every split. For split 1 the accuracies ranged between 73% and 86% across all four metrics, while for the other two splits the accuracies was 100% across all the metrics.

Split 1

f1 1	0.86
f1 0	0.83
Accuracy	0.85
Matthews correlation coefficient	0.73

Split 2

f1 1	1.0
f1 0	1.0
Accuracy	1.0
Matthews correlation coefficient	1.0

Split 3

f1 1	1.0
f1 0	1.0
Accuracy	1.0
Matthews correlation coefficient	1.0

Table 5: This table shows an overview of the performances of a logistic regression model run on the test set for every split. Four performances metrics were calculated, the f1 1 and f1 0 metric and the accuracy and the Matthews correlation coefficient.

4.6 Validation Study

The validation study for every split can be seen in figure 23. The null hypothesis, H_0 was rejected at a 5% significance level for both the validation study 1 colored in blue and validation study 2 colored in green. For all splits, the p-value was quite low for validation study 1 and 0 for validation study 2. This means that there was almost no models with randomly selected features or permuted labels that had a better performance than the RENT model.

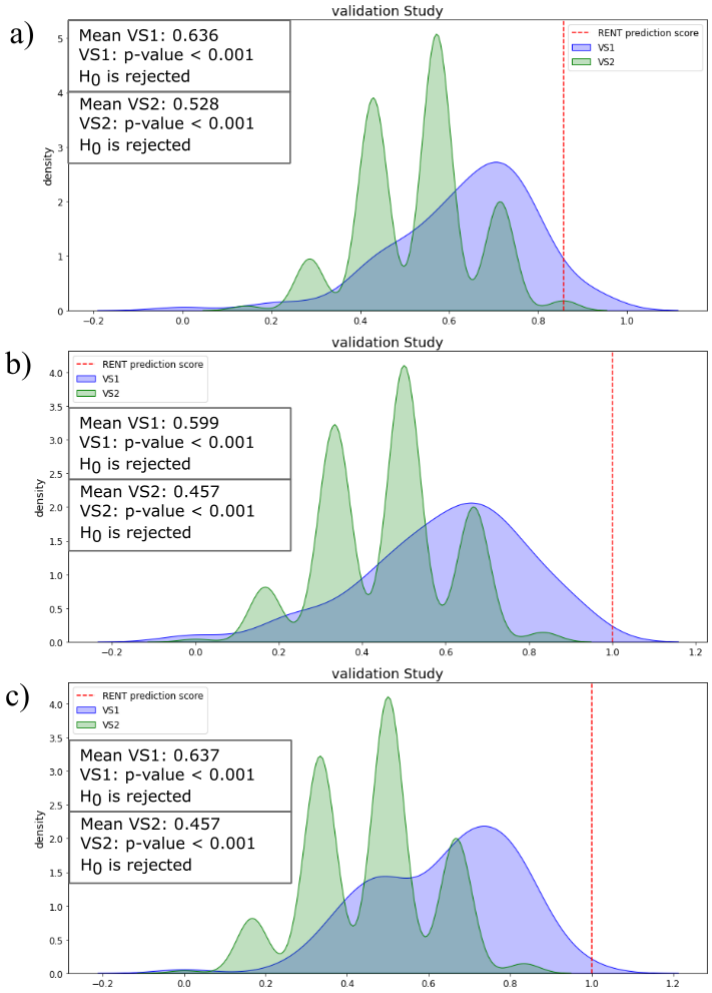

Figure 23: This figure displays two validation studies conducted in split 1 (a), split 2 (b) and split 3 (c). In validation study 1 (VS1) random features are drawn while in validation study 2 (VS2) the response target are permutated. In both these tests, 100 logistic regression models were trained and predicted on unseen test data. The MCC scores were compared with predictions based on features selected by RENT. In order to compare the MCC scores, one sided Student's t-test are conducted with a 5% significance level. The null hypothesis claims that the RENT MCC is lower than the average MCC from VS1 and VS2, respectively

4.7 PCA for every split

Figure 24 displays a PCA completed for each split. The PCA was conducted on a subset of the dataset consisting of RENT selected features from each split, respectively. The PCA for each split consists of a score plot and a loadings plot.

There is a separation between controls and patients in the score plot in every split. The patient and control clusters are denser for split 2 and 3. Generally for all splits the controls 038 and 058 are positioned closer to the patient group than the other controls. Patient 01 lies further towards the control group than the other patients. In split 1, there seems to be some tendency of responders being located higher on PC2 in the plot than the non-responders. The RENT selected features in all the splits are separable in the loadings plots on the PC1 axis. The PC1 in all the splits explains most of the variance in the dataset. The patients have high values of the texture features GLCM Imc1 and GLCM ClusterShade for label 2014 in all splits. While controls have high values of the texture features GLCM Imc2 and GLCM Cluster Prominence for label 2014 in all three splits. There is a marked variation in PC1 for all splits.

The GLCM Imc2 feature in label 1014 is responsible for most of the variation in PC2 for split 1. PC1 and PC2 explain 97.1% of the variance in the dataset in split 1, as seen on the axes of the scores and loadings plot.

Control 049 is positioned outide of the control cluster in the score plot of split 2, and may be an outlier. The first order feature Mean Absolute Deviation of label 2027 caused the variation in PC2 as seen in the loadings plot of split 2. Control 049 has a high value of this particular feature. The two first principal components explain 94% of the variance in the dataset in split 2.

The patients 08 and 014 have a lower value of PC2 than the other individuals, as seen in the score plot of split 3. The GLCM MCC and GLCM Correlation for label 2014 are positively correlated according to the loadings plot in split 3. The two outliers mentioned above have high values for these features.The two first principal components explain 93.3% of the total variance in the dataset in split 3.

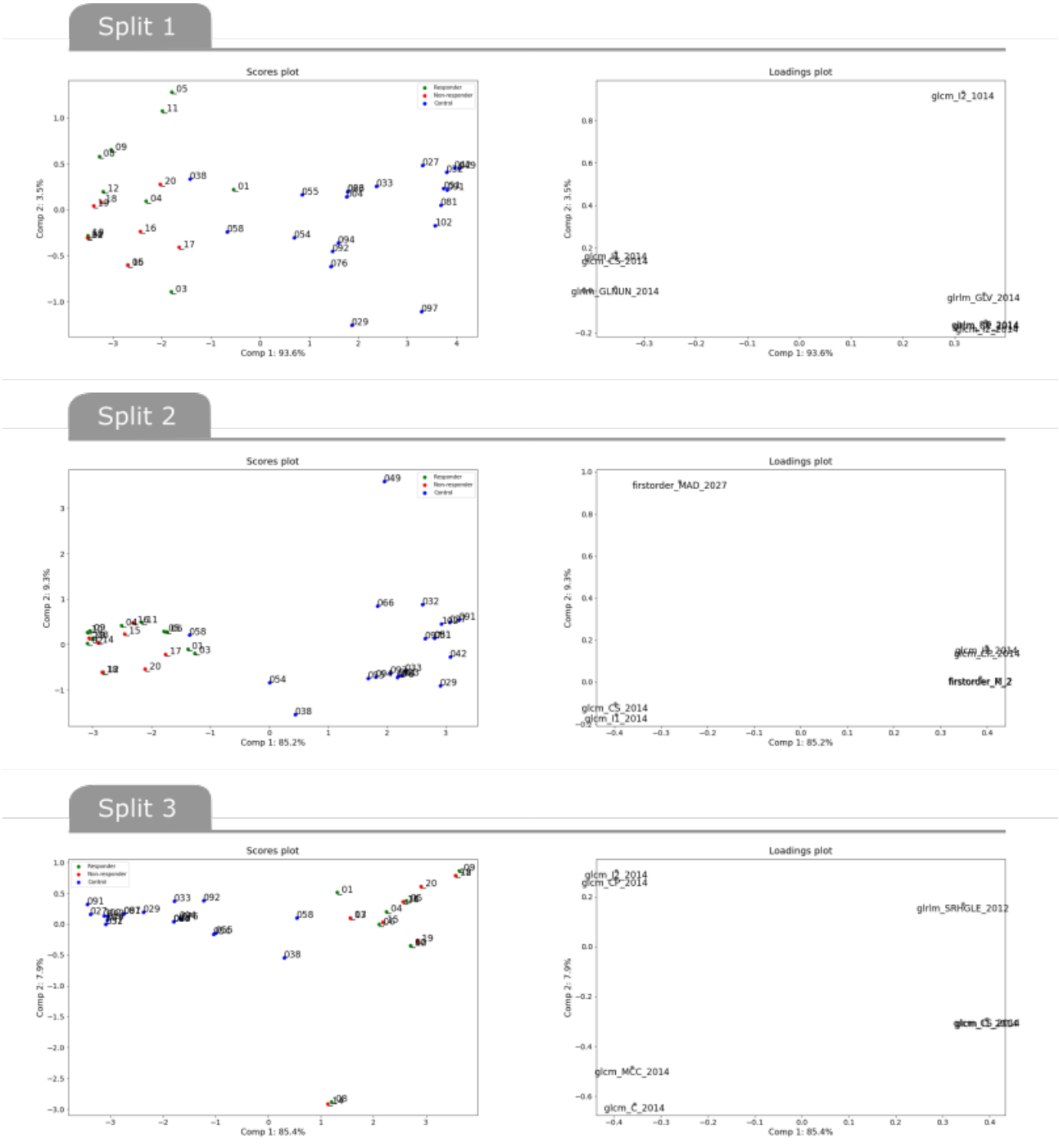

Figure 24: Score plot and loadings plot of the responders, non-responders and controls at timestep t0 from a PCA conducted on RENT selected features from split 1, split 2 and split 3.

4.8 PCA on predefined brain region

Two PCA analyses were conducted on the patients and controls at timestep t0; one was for all masks corresponding to the entire brain and the other was for masks corresponding to the brain regions hippocampus and anterior cingulate. Figure 25 show the scores and cumulative explained variance plot for the PCA conducted on the entire brain dataset. The clustering of the control and patient group was not as distinct as in the PCA conducted on the RENT selected features for every split. Although there is definitely a separation between the two classes. Controls 058 and 054 are positioned further into the patient class, compared to the other controls. The two first principal components explains only $17,2\%$ of the dataset and as seen in the explained variance plot, five principal components yields below 50% explained variance.

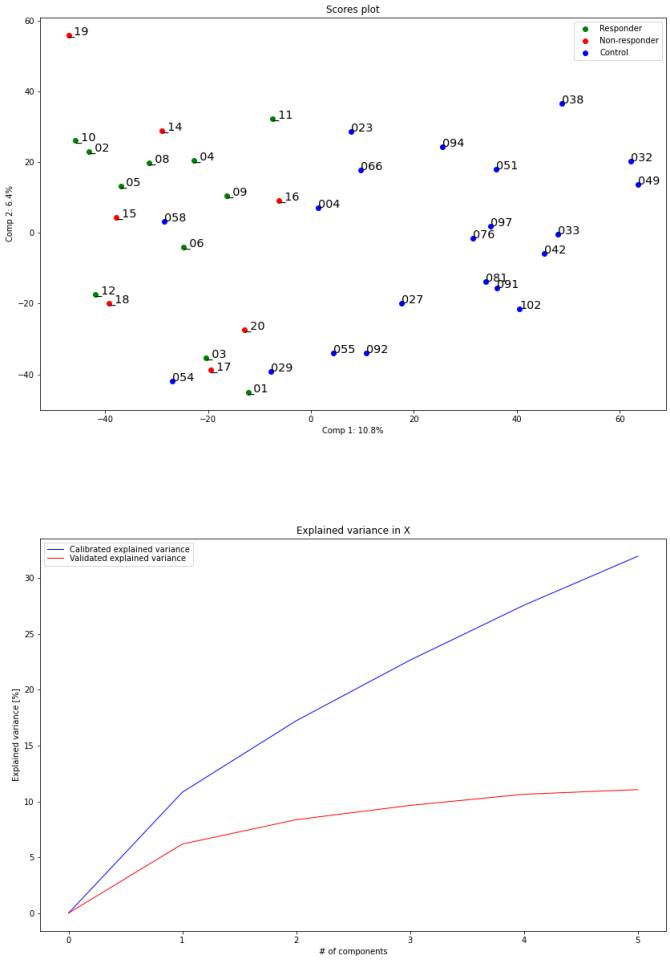

Figure 25: Score plot and cumulative explained variance plot of the responders, non-responders and controls at timestep t0 from a PCA conducted on the entire dataset corresponding to the entire brain.

Figure 26 show the scores and cumulative explained variance plot for the PCA conducted on the dataset of the two brain regions hippocampus and anterior cingulate; the two classes (patients and controls) can be separated nicely. Patient 01 is more similar to the control group compared to the other patients. Patients 15 and 06 are similar but are outliers from the rest of the individuals. The two first principal components explain 30.8% of the variance in the dataset and increases to just above 50% if five principal components are added.

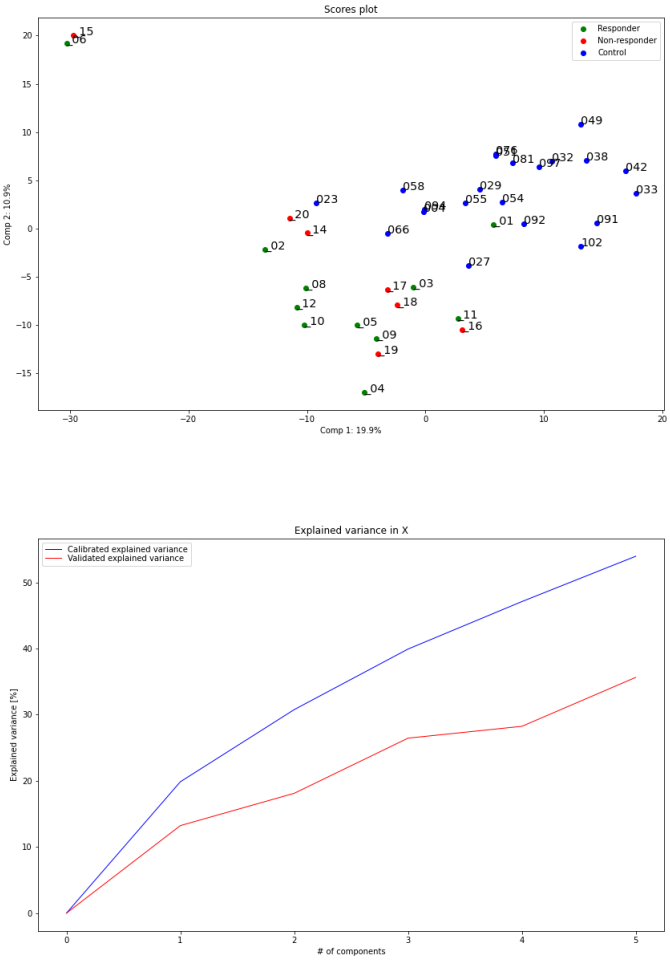

Figure 26: Score plot and cumulative explained variance plot of the responders, non-responders and controls at timestep t0 from a PCA conducted on a selection of features corresponding to the brain regions hippocampus and anterior cingulate.

4.9 PCA with RENT selected features

The three PCA's shown in figure 27, was conducted on the dataset with controls and patients at each timestep containing only the four features selected by RENT in all splits. The score plots shows a clear separation between the classes in timesteps t0 and t1, but not in t2. Controls 058 and 038 seem to lie closer to the patients than the other controls in the scoreplots from timesteps t0 and t1. In the loadings plots for timesteps t0 and t1 the controls have a high values of the texture features GLCM Imc2 and GLCM Cluster Prominence for label 2014, while the patients have high values for the texture features GLCM Imc1 and GLCM ClusterShade for label 2014. The first two principal components explain around 99.5% of the variance in the dataset for both timesteps t0 and t1.

In timestep t2, the loadings plot is similar to the two other timesteps, although there is a no clear seperation of the classes in the score plot. Therefore there is not similar connections between the features and the classes as in timesteps t0 and t1.

There was also conducted a PCA on the dataset for patients and controls at each timestep containing all features RENT selected in the three splits. The result can be seen in appendix C.

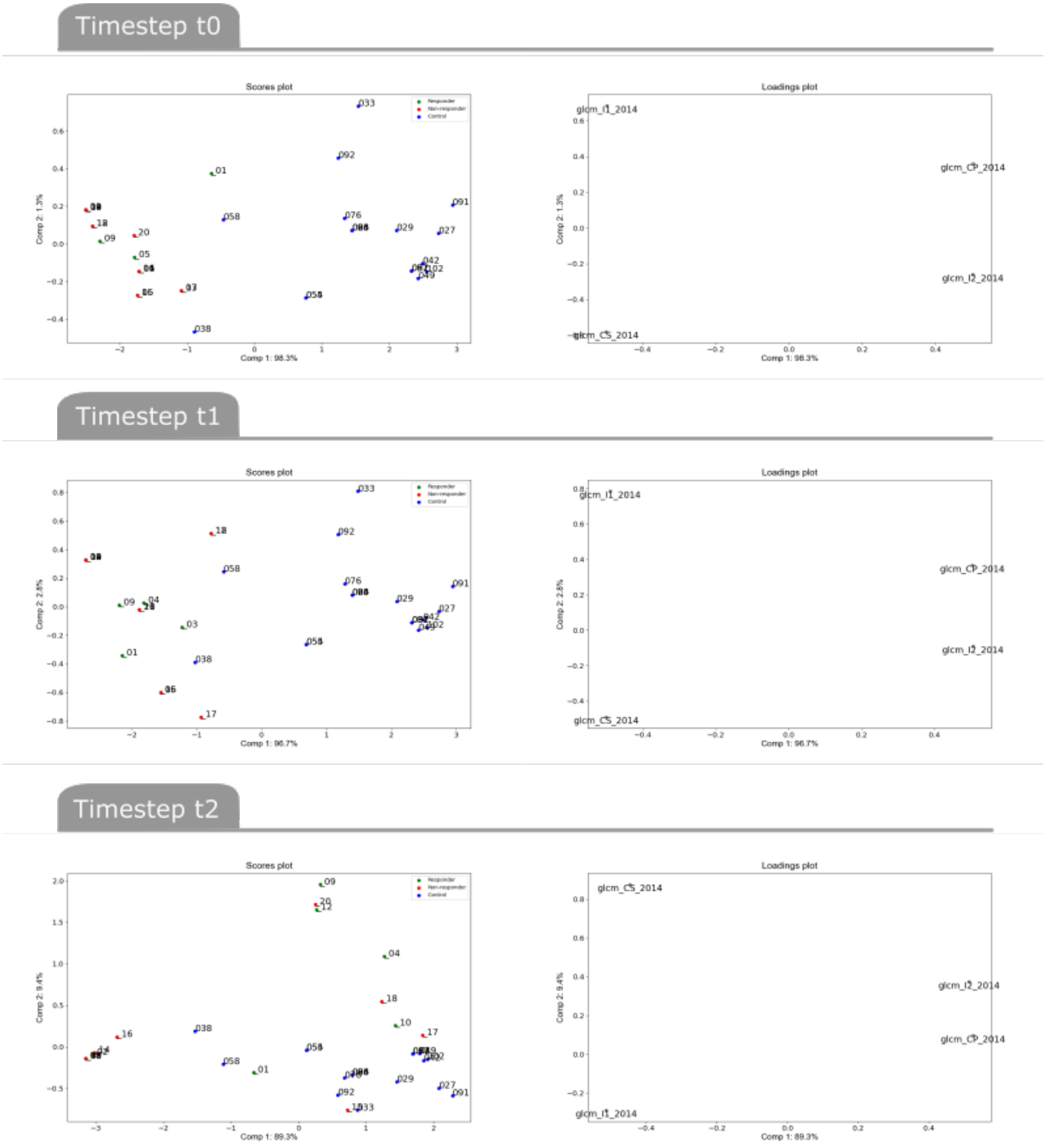

Figure 27: This figure shows the scores and loadings plot from a PCA conducted on the dataset consisting of patients and controls at timesteps, t0, t1, and t2. The dataset contained features that the RENT model selected in every split, in total 4 features.

5 Discussion

The main objective for this thesis was to detect differences between patients diagnosed with Major Depressive Disorder (MDD) and healthy controls using radiomics features extracted from structural MR images. In addition, the study searched for reliable biomarkers that can potentially be used to diagnose patients with MDD. The RENT algorithm separated patients and controls with high accuracy in every split of the dataset and selected in total 14 features. Four of these features were selected in every split and may be considered possible neural biomarkers for predicting a depression diagnosis.

5.1 Evaluating the selected features

The RENT algorithm selected eight features in split 1 and seven features in the two other splits. The features selected in each split can be viewed in figure 17. In total, 14 features were selected across all splits, seen in figure 18. Four of these features were selected by all splits, suggesting that these four features were consistently important for RENT to separate the classes. Figure 19 displays the proportion of types of radiomics features selected by RENT. 78.6% of the 14 selected features were texture features, while the remaining features were first-order features. No shape features were selected by RENT, indicating that variations in brain regions' shape were not as useful for RENT to separate the classes.

The 14 selected features corresponded to five mask labels which again corresponded to five brain regions. An overview of mask labels and brain regions can be viewed in figure 21. The brain regions labeled 2012, 2014, and 1014 correspond to the orbitofrontal gyrus, which is located in the orbitofrontal cortex (OFC) [33]. Label 2027 corresponds to the right middle frontal gyrus, and label 2 corresponds to white matter in the left cerebral hemisphere. The distribution of labels in the selected features can be seen in figure 20, in total, 78.5% of the selected features corresponded to regions in the orbitofrontal gyrus (figure 21). The selected features indicate that the orbitofrontal gyrus may be an essential brain region to predict a depression diagnosis. Four features were selected in all splits, which can be seen in figure 18. These four features are texture features that corresponded to the mask label 2014. 64% of the 14 selected features also correspond to this brain region. Label 2014 is the brain region, right medial orbital gyrus located in the OFC [33]. All three splits found these four features important for predicting the depression diagnosis, indicating that these features in the medial orbitofrontal regions may be considered possible biomarkers.

Several studies have found connections between the OFC and depression. Fonseka et al. (2018) reviewed 95 articles on studies examining predictors of treatment response from structural and functional neuroimaging modalities [7]. They found multiple possible biomarkers in frontolimbic regions, including the prefrontal cortex, anterior cingulate cortex, hippocampus, amygdala, and insula, most frequently influenced response outcome although the strength and direction of the biomarker's association with clinical response varied, likely due to study differences. Lacerda et al. (2004) [9] studied 31 patients diagnosed with major depressive disorder and 34 controls subjects. The study observed a smaller volume of gray matter in the lateral and medial OFC in the MDD patients. Konarski et al. (2008) [8] reviewed 140 magnetic neuroimaging investigations with either bipolar disorder or MDD diagnosed patients. Several studies reported a reduction in OFC grey matter volumes in MDD patients. However, some of the 140 studies did not find these volumetric changes in gray matter volume in the OFC. Frontolimbic regions like the prefrontal cortex, anterior cingulate cortex and the insula are interesting regions in the discovery of new neural biomarkers [4]. The results of this thesis may contribute to validating the medial orbitofrontal gyrus as a possible neural biomarker. However, in order for neural biomarkers to be used in the medical field, the biomarkers have to be replicated and validated many times in large independent sets of samples [7].

5.2 Evaluating the model performance

The performance of the RENT model was evaluated by reviewing the performances of the 100 ensemble models. A validation study was also applied to the RENT model in order to detect if the model performed better than a random model. Further, a logistic regression model was applied to the test data of every split using RENT selected features to test how relevant and reliable the features were for predicting patients diagnosed with MDD.

RENT created an ensemble of models to predict whether an individual belonged to the patient group or the control group. The performance of the models over the three splits is shown in figure 22. The performance is generally high as most model performances range from 0.6 to 1.0 over all three splits. The consistency of model performances in each split is an indication that the model performed well on unseen data [20]. The performances were also consistent over all three splits, which again reassures that the model predicted well independently of which individuals were in the training set. It appear that many features weights were set to zero, but since the dataset is

large with 10165 features, setting 99% of the weights to zero would still yield 102 features. In all three splits, the share of features set to zero was quite high, which means that the strength of the feature selection is consistent.

Two validation studies were performed to ensure that the features selected by RENT were significant for the high model performance. In validation study 1 (VS1), random features were drawn, while in validation study 2 (VS2), the response target is permutated [18]. In both these tests, RENT trained 100 logistic regression models and predicted on unseen validation data. Then RENT compared the MCC scores of these tests with predictions based on features selected by RENT. In order to compare the MCC scores, a one-sided Student's t-test was conducted. The null hypothesis claimed that the RENT MCC was lower than the average MCC from VS1 and VS2, respectively [18]. The null hypothesis was rejected for all three splits (figure 23), meaning that RENT selected relevant and important features for predicting whether a patient has depression. There was generally a high MCC score in all the validation studies for each split, marked by the red line. The red line was consistently further to the right than most of the VS1 and VS2 distributions, indicating that RENT performed well independently of the training set.

A logistic regression model using only features selected by RENT was applied to the test set for every split. The performance metrics were quite high as they range between 73% and 86% in split 1 and 100% for all metrics in split 2 and 3 (figure 5). The high metrics are promising for using these features on new data, indicating that the RENT selected features may be possible biomarkers for predicting a depression diagnosis. Metrics as high as 100% can look suspicious; keep in mind that the test size is quite small, only testing 13 samples. One correct prediction has a high impact on the metrics. It can also be a sign of overfitting that the features selected by RENT make it simple to separate the two groups, especially in this dataset. When applied to another dataset, the model might not yield such a high performance.

5.3 Separation of the classes

The thesis used PCA analyses to see if the control and patient classes were separable at different subsets of the dataset and to detect possible outliers. PCA was conducted on the dataset for patients and controls at timestep t0, with RENT selected features for each split (figure 24). The separation between the classes was evident in the score plots for all three splits. This separation shows that it is possible to separate the two classes with just the features selected by RENT for every split. Therefore, it is possible to see a

difference in controls and patients only using the features selected by RENT. In the score plot for split 1, the responders tend to lie higher along PC2 than the non-responders, which may indicate that the RENT selected features can be predictive for treatment response.

The dataset for patients and controls at timestep t0 is a short-wide dataset with 10165 columns and only 39 samples. The size of this dataset yields enormous information that corresponds to each individual (patient/control). Correlation matrices were calculated for the patients and controls at timestep t0 to give insight into how much information overlaps between the patients and controls. Two correlation matrices were calculated—one of the entire dataset and one containing masks corresponding to the brain regions hippocampus and anterior cingulate. Changes in the hippocampus and anterior cingulate are often associated with MDD [4]. Figure 14 displays the correlation matrix conducted on the entire dataset. The correlation between the individuals was generally high. There was no clear distinction between the patients and the controls, which means that the patients were not necessarily more correlated than with the controls and vice versa. There was no indication of separating the patient and control group by investigating the correlation matrix for the entire dataset. The patients/controls did not correlate as highly in the correlation matrix conducted on the brain regions hippocampus and anterior cingulate (fig. 15) as on the entire dataset. The yellow box in the lower-left corner indicates that the controls were more similar to each other. A separation of the classes was more evident in the hippocampus and anterior cingulate regions than in the entire brain, indicating that every brain is alike in its entirety. However, if the dataset has fewer columns, it is easier to see differences between the patients and controls.

A PCA was conducted on the entire dataset with patients and controls at timestep t0, with masks corresponding to the whole brain. The score plot (figure 25) showed that the patient and control classes do not cluster together tightly, although it was easy to see a clear separation between classes. The two first principal components only explained 17% of the variance in the dataset. Much of the variation in the dataset was therefore not accounted for. Another PCA analysis was conducted on the dataset with the patients and controls at timestep t0, with columns corresponding to the brain regions hippocampus and anterior cingulate (figure 26). The score plot showed that the patients and controls were generally more clustered than the score plot with the entire brain. There was also a separation between the classes. The two first principal components explained 31% of the dataset, which is higher than the analysis for the entire brain, meaning there was still a separation

between the classes. However, the dataset is smaller, and principal components can explain more of the variation in the dataset. This separation may indicate that the hippocampus and the anterior cingulate may be predictive for a depression diagnosis. Although there was a separation, there were still many features, and the principal components explained little of the variation in the dataset. The features selected by RENT can be a solution to reduce the feature space even further and still be able to separate the two classes.

The score plot and loadings plot for the PCA analysis conducted on the patients and controls at timesteps t0, t1 and t2, with just the four features selected by all splits is displayed in figure 27. All the splits consistently selected these features. Patients and controls can be separated in a score plot for timestep t0, using only these four features, suggesting that these features may be reliable as biomarkers for diagnosing depression. The score plot for timestep t1 shows a clear separation between the classes. Timestep t1 corresponds to 6 weeks into the treatment, which means that the features were useful to separate the classes even though the patients received medication for depression. In the score plot at timestep t2 the two classes did not cluster together as much as the two other timesteps, although the controls stayed in the same region. Timestep t2 corresponds to 12 weeks into the treatment. The lack of clustering in the patient group may indicate that the patients were not as similar after 12 weeks. However, the controls scanned at t0 and patients at timesteps t1 and t2 might not be comparable, as there can be differences between MR-images between visits [6]. The images were normalized to reduce these effects.

5.4 Outliers

Outliers deviate from the other samples in the study and can be the source of bias in the model. This thesis calculated correlation matrices for the dataset of patients and controls at timestep t0 to detect how much the patients and controls correlated. Two correlation matrices were calculated—one of the entire dataset and one containing masks corresponding to the brain regions hippocampus and anterior cingulate, (figures 14 and 15). In the correlation matrix calculated on the entire dataset, patients 08 and 15 were less correlated with the other patients/controls and should be investigated further as they may be considered outliers. In the correlation matrix conducted on the brain regions hippocampus and anterior cingulate (figure 15), control 027 was not as correlated to the other controls and may be an outlier.

The tables 2−4 summarizes the patients and controls for every split and can be used to further investigate outliers. These tables inform how many times an individual was in the test set, its true class label and how many times it was predicted wrong. It also yields the percentage of incorrect predictions. Generally, there was a low percentage of incorrect predictions for most individuals. Patient 01 and control 058 had a high percentage of incorrect predictions in split 2 and 3, which may indicate that these individuals were outliers. Patient 14 and controls 054 and 066 had a high percentage of incorrect predictions in one of the splits but low incorrect predictions in another split. These might be outliers but are dependent on what samples were in the training set. None of the patients or controls that were uncorrelated in the correlation matrices were difficult to predict for the RENT model. The information that made the patients/controls uncorrelated was not predictive for the response. Therefore RENT had no issues classifying them.

The PCA analysis is also a good method to detect and shed light on deviations in the dataset. A PCA analysis was conducted on the dataset for patients and controls at timestep t0, with RENT selected features for each split, (figure 25). In the scores plots over all three splits, patient 01 and controls 38, 54, and 58 were closer to the opposite class than their own class. Patient 01 and control 058 were also mispredicted in two splits according to the summary of individuals mentioned above. Therefore, these two samples can be viewed as outliers in the feature space constructed by the features selected by RENT. Control 54 may also be an outlier as it was difficult to predict in one of the splits and was difficult to separate it in the score plot in the PCA analysis. Patient 14 and control 066 are placed in their classes in the PCA analysis, although the RENT model has difficulty classifying them, indicating that they may be outliers. On the other hand, control 038 is difficult to separate in the PCA score plots, but RENT has minor issues predicting it's response.

Whether an individual is an outlier is challenging to assess. Patient 01 and control 058 acted as outliers from their class, which can be seen in the PCA analysis and how well the RENT models predict these samples. These individuals did not have features with extreme values; they are difficult to classify and therefore add bias to the RENT model.

6 Further work

The largest obstacle to face when exploring datasets of extracted radiomics features from MR images is the variation in the MR images. Methods for extracting biomarkers need to be tested on different independent datasets to validate robust and reliable biomarkers. Therefore using RENT as a feature selector should be applied to independent larger sets of MR images to validate the RENT method used in this thesis. The 3-fold split could be performed several times, which would further validate the stability and robustness of the selected features.

This thesis applied RENT to the images taken at study entry; it would be interesting to apply the method to the images taken at 6 weeks (t1) and 12 weeks (t2) after study entry. Further, it should be investigated if RENT may be able to select features that are predictive for treatment response. It may be helpful to apply this method to only the patients dataset at each timestep or the differences between the patients datasets at the different timesteps. Analyses for the patients and controls may also be investigated further at t1 and t2 to detect if the responders become more similar to the controls.

RENT could also be performed on just one type of radiomics feature. Several studies have shown that volume differences in several frontolimbic regions could predict MDD and treatment response [1]. Therefore running RENT on just the shape features may also contribute to finding other biomarkers and see if these correspond to similar findings.

The preprocessing of MR images could also be investigated as there is no protocol for normalization and discretization. There are different methods for normalization like Nyul's harmonizations method and the white stripe method; it could be interesting to try some of these methods to see if the results change. It would also be interesting to try different bin widths in the discretization. Comparing the results from different bin widths can show how much discretization affects the extraction of radiomics features and how this affects RENT.

7 Conclusion

Radiomics features were extracted from T1 weighted MR images from 21 patients diagnosed with MDD and 22 healthy controls. The feature space of radiomics features were reduced by using RENT as a feature selector as well as predicting the diagnosis of the patients with MDD.

In this thesis, we succeeded in reducing the feature space to four features with RENT while maintaining a separation of patients diagnosed with MDD and healthy controls. Furthermore, the three dataset splits jointly selected four features that corresponded to the brain region's right medial orbital gyrus. Certain characteristics of the right medial orbital gyrus may therefore be a possible biomarker for diagnosing patients with MDD. RENT should be tested further on an independent set of samples to replicate these findings and further validate image characteristics of the right medial orbital gyrus as a possible biomarker.

RENT had a high performance across all models over all three dataset splits, indicating that RENT predicted MDD diagnosed patients with high accuracy. The method was further investigated by performing two validation studies, investigating if the model performed better than a random model. The RENT method passed the validation studies in all splits, indicating that the radiomics features selected by RENT were significant for predicting the MDD diagnosis. A logistic regression model was applied to the test set with only the features selected by RENT in the specific split. The accuracies ranged from 73% and 100%, which again supported that the features RENT selected were of importance. Although the high accuracies over all three splits point to a robust model, it should be mentioned that with accuracies this high, overfitting may be an issue.

RENT is a useful tool to reduce feature spaces, even for datasets with a much greater number of features compared to samples, as often is the case when extracting radiomics features.

References

[1] Fang Y. Depressive Disorders: Mechanisms, Measurement and Management. vol. 1180. Springer; 2019.

[2] Abate KH, Abebe Z, Abil OZ, Afshin A, Ahmed MB, Alahdab F, et al.. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. England: Elsevier Ltd; 2018.

[3] Kim YK. Understanding Depression: Volume 1. Biomedical and Neurobiological Background. Singapore: Springer Singapore Pte. Limited; 2018.

[4] Schrantee A, Ruhé HG, Reneman L. Psychoradiological Biomarkers for Psychopharmaceutical Effects. Neuroimaging Clinics of North America. 2020;30(1):53–63. Psychoradiology. Available from: `https://www.sciencedirect.com/science/article/pii/S1052514919300851`.

[5] Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypinski P, Gibbs P, et al. Introduction to Radiomics. Journal of Nuclear Medicine. 2020. Available from: `https://jnm.snmjournals.org/content/early/2020/02/13/jnumed.118.222893`.

[6] Goya-Outi J, Orlhac F, Calmon R, Alentorn A, Nioche C, Philippe C, 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 and Biology. 2018 04;63.

[7] Fonseka TM, MacQueen GM, Kennedy SH. Neuroimaging biomarkers as predictors of treatment outcome in Major Depressive Disorder. Journal of Affective Disorders. 2018;233:21–35. Are there Biomarkers for Mood Disorders? Available from: `https://www.sciencedirect.com/science/article/pii/S0165032717310431`.

[8] Konarski JZ, McIntyre RS, Kennedy SH, Rafi-Tari S, Soczynska JK, Ketter TA. Volumetric neuroimaging investigations in mood disorders: bipolar disorder versus major depressive disorder. Bipolar disorders. 2008;10(1):1–37.

[9] Lacerda ALT, Keshavan MS, Hardan AY, Yorbik O, Brambilla P, Sassi RB, et al. Anatomic evaluation of the orbitofrontal cortex in major

depressive disorder. Biological Psychiatry. 2004;55(4):353–358. Available from: https://www.sciencedirect.com/science/article/pii/S0006322303009491.

[10] Ruhé H, Booij J, Veltman D, Michel M, Schene A. Successful Pharmacologic Treatment of Major Depressive Disorder Attenuates Amygdala Activation to Negative Facial Expressions: A Functional Magnetic Resonance Imaging Study. The Journal of clinical psychiatry. 2011 08;73:451–9.

[11] Bobo WV, Angleró GC, Jenkins G, Hall-Flavin DK, Weinshilboum R, Biernacka JM. Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to Clinical Global Impression scale ratings: analysis of Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS) data. Human psychopharmacology. 2016;31(3):185–192.

[12] Hobbie RK. Intermediate physics for medicine and biology. New York: Springer; 2007.

[13] Flower MA. Webb's Physics of Medical Imaging, Second Edition. Series in Medical Physics and Biomedical Engineering. Taylor & Francis; 2012. Available from: https://books.google.no/books?id=qkFljemf7y0C.

[14] Ray H H, Christopher J L, William B. MRI: The Basics.. vol. Fourth edition. Wolters Kluwer Health; 2018. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=2013143&site=ehost-live.

[15] Val M R, Wolfgang R N, Johannes T H. The Physics of Clinical MR Taught Through Images.. vol. Fourth edition. Thieme; 2018. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=1804450&site=ehost-live.

[16] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. European radiology experimental. 2018;2(1):36–36.

[17] van Griethuysen J, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 2017 11;77:e104–e107.

[18] Jenul A, Schrunner S, Liland KH, Indahl UG, Futsaether CM, Tomic O. RENT – Repeated Elastic Net Technique for Feature Selection; 2021.

[19] Jenul A. RENT Release 0.0.1; 2021. Accessed: 20.06.2021. https://rent.readthedocs.io/en/stable/.

[20] Raschka S, Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition. Packt Publishing; 2019. Available from: https://books.google.no/books?id=sKXIDwAAQBAJ.

[21] Hackeling G. Mastering machine learning with scikit-learn : apply effective learning algorithms to real-world problems usung scikit-learn. 1st ed. Community experience distilled. Birmingham: Packt Publishing; 2014.

[22] Idris I. Python Data Analysis Cookbook. Packt Publishing; 2016.

[23] Shmueli B. Matthews Correlation Coefficient Is The Best Classification Metric You've Never Heard Of. towardsdatascience com. 2019.

[24] Nesvold JE. Radiom Pipe - Radiomic feature extraction made simple. 2021 Jan.

[25] Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage. 2006;31(3):968 – 980. Available from: http://www.sciencedirect.com/science/article/B6WNP-4JFHF4P-1/2/0ec667d4c17eafb0a7c52fa3fd5aef1c.

[26] Schindler S, Schreiber J, Bazin PL, Trampel R, Anwander A, Geyer S, et al. Intensity standardisation of 7T MR images for intensity-based segmentation of the human hypothalamus. PloS one. 2017;12(3):e0173344–e0173344.

[27] Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magnetic resonance imaging. 2004;22(1):81–91.

[28] Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Scientific reports. 2020;10(1):12340–12340.

[29] Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik JC, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PloS one. 2019;14(3):e0213459–e0213459.

[30] McKinney W, et al. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. vol. 445. Austin, TX; 2010. p. 51–56.

[31] Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021. Available from: https://doi.org/10.21105/joss.03021.

[32] Tomic O, Graff T, Liland KH, Næs T. hoggorm: a python library for explorative multivariate statistics. The Journal of Open Source Software. 2019;4(39). Available from: http://joss.theoj.org/papers/10.21105/joss.00980.

[33] Neil M Borden M, Scott E Forseen M, Cristian Stefan M. Imaging Anatomy of the Human Brain : A Comprehensive Atlas Including Adjacent Structures. Demos Medical; 2016. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=1081584&site=ehost-live.

A Overview of how the samples were divided into three splits.

Split 1	Split 2	Split 3
patient_03	patient_01	patient_01
patient_06	patient_02	patient_02
patient_08	patient_04	patient_03
patient_09	patient_05	patient_04
patient_10	patient_06	patient_05
patient_12	patient_08	patient_10
patient_14	patient_09	patient_11
patient_16	patient_11	patient_12
patient_17	patient_14	patient_15
patient_18	patient_15	patient_17
patient_19	patient_16	patient_18
patient_20	patient_19	patient_20
control_004	control_029	control_004
control_023	control_032	control_023
control_027	control_033	control_027
control_032	control_038	control_029
control_042	control_051	control_033
control_049	control_055	control_038
control_051	control_058	control_042
control_054	control_066	control_049
control_055	control_076	control_054
control_076	control_081	control_058
control_081	control_091	control_066
control_091	control_092	control_092
control_094	control_097	control_094
control_102	control_102	control_097
patient_01	patient_03	patient_06
patient_02	patient_10	patient_08
patient_04	patient_12	patient_09
patient_05	patient_17	patient_14
patient_11	patient_18	patient_16
patient_15	patient_20	patient_19
control_029	control_004	control_032
control_033	control_023	control_051
control_038	control_027	control_055
control_058	control_042	control_076
control_066	control_049	control_081
control_092	control_054	control_091
control_097	control_094	control_102

Figure B1: The figure shows how the patients and controls were split three times into a training set colored in green and a test set colored in blue. These splits were stratified so that there was an even distribution of patients and controls in the training and test set.

B Determining regularization parameters for the RENT algorithm

dataFrame_1: average scores for predictive performance.
The higher the score, the better the parameter combination.

		c				
		0,01	0,1	1	10	100
l	0	0,7387	0,7387	0,7387	0,7387	0,7387
	0,1	0,6395	0,8748	0,8162	0,7387	0,7387
	0,25	NaN	0,8748	0,8748	0,7387	0,7387
	0,5	NaN	0,9333	0,8748	0,8162	0,7387
	0,75	NaN	0,8162	0,8748	0,8162	0,7387
	0,9	NaN	0,6978	0,8748	0,8162	0,7387
	1	NaN	0,6395	0,8748	0,8162	0,7387

Max performance: 0,9333

		c				
		0,01	0,1	1	10	100
l	0	0,6870	0,6000	0,5972	0,6231	0,7044
	0,1	0,6597	0,8894	0,8150	0,7564	0,4936
	0,25	NaN	0,9414	0,7081	0,7495	0,7414
	0,5	NaN	0,8558	0,9414	0,6867	0,6667
	0,75	NaN	0,8639	0,9225	0,6667	0,7014
	0,9	NaN	0,7533	0,9414	0,7414	0,6748
	1	NaN	0,3536	0,9414	0,7642	0,4639

Max performance: 0,9414

		c				
		0,01	0,1	1	10	100
l	0	0,6228	0,6894	0,6748	0,4453	0,3822
	0,1	0,6124	0,9333	0,8558	0,4561	0,6350
	0,25	NaN	0,9225	0,8081	0,5939	0,5972
	0,5	NaN	0,8894	0,8894	0,6081	0,6867
	0,75	NaN	0,6414	0,8309	0,8748	0,7561
	0,9	NaN	0,7387	0,9414	0,7344	0,7453
	1	NaN	0,6597	0,9414	0,9414	0,6894

Max performance: 0,9414

		c				
		0,01	0,1	1	10	100
l	0	0,8748	0,6053	0,4081	0,4789	0,6639
	0,1	0,6440	0,9333	0,6870	0,4228	0,6611
	0,25	NaN	0,9225	0,7783	0,5667	0,4081
	0,5	NaN	0,9333	0,8894	0,7414	0,6347
	0,75	NaN	0,8053	0,9225	0,6817	0,6309
	0,9	NaN	0,8162	0,9225	0,8558	0,6558
	1	NaN	0,4167	0,8162	0,7455	0,4825

Max performance: 0,9333

		c				
		0,01	0,1	1	10	100
l	0	0,7642	0,6558	0,4527	0,5867	0,6561
	0,1	0,6869	0,9333	0,8053	0,7561	0,5642
	0,25	NaN	0,8000	0,7455	0,6053	0,6936
	0,5	NaN	0,9225	0,9333	0,6680	0,7483
	0,75	NaN	0,7225	0,9414	0,7150	0,6456
	0,9	NaN	0,6936	0,8817	0,7711	0,8228
	1	NaN	0,5577	0,8894	0,6558	0,6309

Max performance: 0,9414

		c				
		0,01	0,1	1	10	100
l	0	0,7495	0,5533	0,7642	0,6228	0,5455
	0,1	0,3895	0,9333	0,7972	0,6044	0,5748
	0,25	NaN	0,9414	0,7972	0,7122	0,7645
	0,5	NaN	0,9414	0,9414	0,8231	0,5939
	0,75	NaN	0,7225	0,8639	0,7228	0,6162
	0,9	NaN	0,6350	0,9414	0,8894	0,8748
	1	NaN	0,3062	0,9414	0,7864	0,7455

Max performance: 0,9414

DataFrame_2: average percentage of how many feature weights were set to zero. The higher the average percentage, the stronger the feature selection with the corresponding paramter combination.

l \ *c*	0,01	0,1	1	10	100
0	0,0356	0,0356	0,0356	0,0356	0,0356
0,1	0,9997	0,8648	0,3724	0,0686	0,0358
0,25	NaN	0,9493	0,6152	0,1344	0,0381
0,5	NaN	0,9829	0,7642	0,2245	0,0472
0,75	NaN	0,9958	0,8312	0,3019	0,0576
0,9	NaN	0,9987	0,8559	0,3446	0,0646
1	NaN	0,9997	0,8687	0,3732	0,0686

Average percentage: 0,9997

l \ *c*	0,01	0,1	1	10	100
0	0,0380	0,0356	0,0365	0,0357	0,0363
0,1	0,9998	0,8678	0,3843	0,0683	0,0367
0,25	NaN	0,9508	0,6163	0,1370	0,0386
0,5	NaN	0,9834	0,7685	0,2283	0,0482
0,75	NaN	0,9954	0,8337	0,3141	0,0603
0,9	NaN	0,9990	0,8570	0,3525	0,0674
1	NaN	0,9999	0,8723	0,3793	0,0725

Average percentage: 0,9999

l \ *c*	0,01	0,1	1	10	100
0	0,0365	0,0365	0,0356	0,0365	0,0356
0,1	0,9998	0,8671	0,3777	0,0735	0,0359
0,25	NaN	0,9507	0,6161	0,1362	0,0390
0,5	NaN	0,9824	0,7656	0,2263	0,0491
0,75	NaN	0,9949	0,8275	0,3107	0,0607
0,9	NaN	0,9988	0,8564	0,3521	0,0653
1	NaN	0,9998	0,8682	0,3812	0,0719

Average percentage: 0,999803

		c				
		0,01	0,1	1	10	100
l	0	0,0364	0,0356	0,0370	0,0365	0,0364
	0,1	0,9995	0,8664	0,3794	0,0669	0,0361
	0,25	NaN	0,9507	0,6072	0,1376	0,0389
	0,5	NaN	0,9830	0,7657	0,2287	0,0498
	0,75	NaN	0,9956	0,8334	0,3139	0,0601
	0,9	NaN	0,9987	0,8562	0,3533	0,0638
	1	NaN	0,9998	0,8695	0,3772	0,0740

Average percentage: 0,9998

		c				
		0,01	0,1	1	10	100
l	0	0,0364	0,0366	0,0356	0,0357	0,0372
	0,1	0,9997	0,8689	0,3841	0,0688	0,0368
	0,25	NaN	0,9528	0,6168	0,1362	0,0400
	0,5	NaN	0,9833	0,7613	0,2270	0,0465
	0,75	NaN	0,9955	0,8286	0,3099	0,0599
	0,9	NaN	0,9989	0,8563	0,3543	0,0644
	1	NaN	0,9993	0,8709	0,3859	0,0753

Average percentage: 0,9997

		c				
		0,01	0,1	1	10	100
l	0	0,0356	0,0364	0,0364	0,0371	0,0364
	0,1	0,9998	0,8673	0,3905	0,0713	0,0358
	0,25	NaN	0,9489	0,6130	0,1343	0,0377
	0,5	NaN	0,9844	0,7668	0,2267	0,0487
	0,75	NaN	0,9959	0,8325	0,3140	0,0604
	0,9	NaN	0,9985	0,8547	0,3532	0,0657
	1	NaN	0,9993	0,8694	0,3815	0,0662

Average percentage: 0,9998

DataFrame_3: harmonic means between dataFrame_1 and dataFrame_2. The parameter combination with the highest

		c				
		0,01	0,1	1	10	100
l	0	0,0000	0,0000	0,0000	0,0000	0,0000
	0,1	0,0000	0,8293	0,4419	0,0621	0,0004
	0,25	NaN	0,8680	0,6867	0,1572	0,0051
	0,5	NaN	0,9913	0,7775	0,2956	0,0233
	0,75	NaN	0,7498	0,8128	0,3785	0,0429
	0,9	NaN	0,3311	0,8250	0,4182	0,0552
	1	NaN	0,0000	0,8312	0,4426	0,0621

Harmonic mean: 0,9913

		c				
		0,01	0,1	1	10	100
l	0	0,0048	0,0000	0,0018	0,0002	0,0015
	0,1	0,6849	0,8867	0,4951	0,0646	0,0022
	0,25	NaN	0,9739	0,6027	0,1820	0,0063
	0,5	NaN	0,9141	0,8637	0,2954	0,0254
	0,75	NaN	0,9274	0,8922	0,3746	0,0490
	0,9	NaN	0,8093	0,9200	0,4387	0,0622
	1	NaN	0,0000	0,9292	0,4720	0,0635

Harmonic mean: 0,9739

		c				
		0,01	0,1	1	10	100
l	0	0,0018	0,0020	0,0000	0,0019	0,0000
	0,1	0,5831	0,9198	0,5001	0,0606	0,0006
	0,25	NaN	0,9575	0,6725	0,1636	0,0070
	0,5	NaN	0,9430	0,8253	0,2655	0,0273
	0,75	NaN	0,6324	0,8117	0,4310	0,0502
	0,9	NaN	0,7782	0,9197	0,4315	0,0589
	1	NaN	0,6633	0,9267	0,5277	0,0704

Harmonic mean: 0,9575

		c				
		0,01	0,1	1	10	100
l	0	0,0016	0,0000	0,0000	0,0019	0,0016
	0,1	0,6197	0,9257	0,4266	0,0300	0,0010
	0,25	NaN	0,9640	0,6440	0,1567	0,0000
	0,5	NaN	0,9912	0,8293	0,3045	0,0285
	0,75	NaN	0,8596	0,8970	0,3714	0,0479
	0,9	NaN	0,8740	0,9107	0,4753	0,0551
	1	NaN	0,0321	0,8186	0,4568	0,0621

Harmonic mean: 0,9912

		c				
		0,01	0,1	1	10	100
l	0	0,002	0,002	0,000	0,000	0,003
	0,1	0,648	0,920	0,482	0,065	0,003
	0,25	NaN	0,814	0,601	0,157	0,009
	0,5	NaN	0,972	0,853	0,274	0,022
	0,75	NaN	0,710	0,903	0,372	0,048
	0,9	NaN	0,660	0,864	0,439	0,058
	1	NaN	0,354	0,880	0,388	0,074

Harmonic mean: 0,972

		c				
		0,01	0,1	1	10	100
l	0	0,0000	0,0015	0,0017	0,0032	0,0015
	0,1	0,2319	0,9207	0,4987	0,0686	0,0004
	0,25	NaN	0,9729	0,6749	0,1764	0,0042
	0,5	NaN	0,9920	0,8626	0,3187	0,0263
	0,75	NaN	0,7905	0,8515	0,4010	0,0489
	0,9	NaN	0,6818	0,9187	0,4848	0,0602
	1	NaN	0,0000	0,9275	0,4866	0,0606

Harmonic mean: 0,9920

C PCA analysis conducted on all features selected by any split

PCA analysis conducted on the dataset consisting of patients and controls at timestep t0, t1 and t2. The dataset contained features selected by any one of the RENT models performed on the three splits.

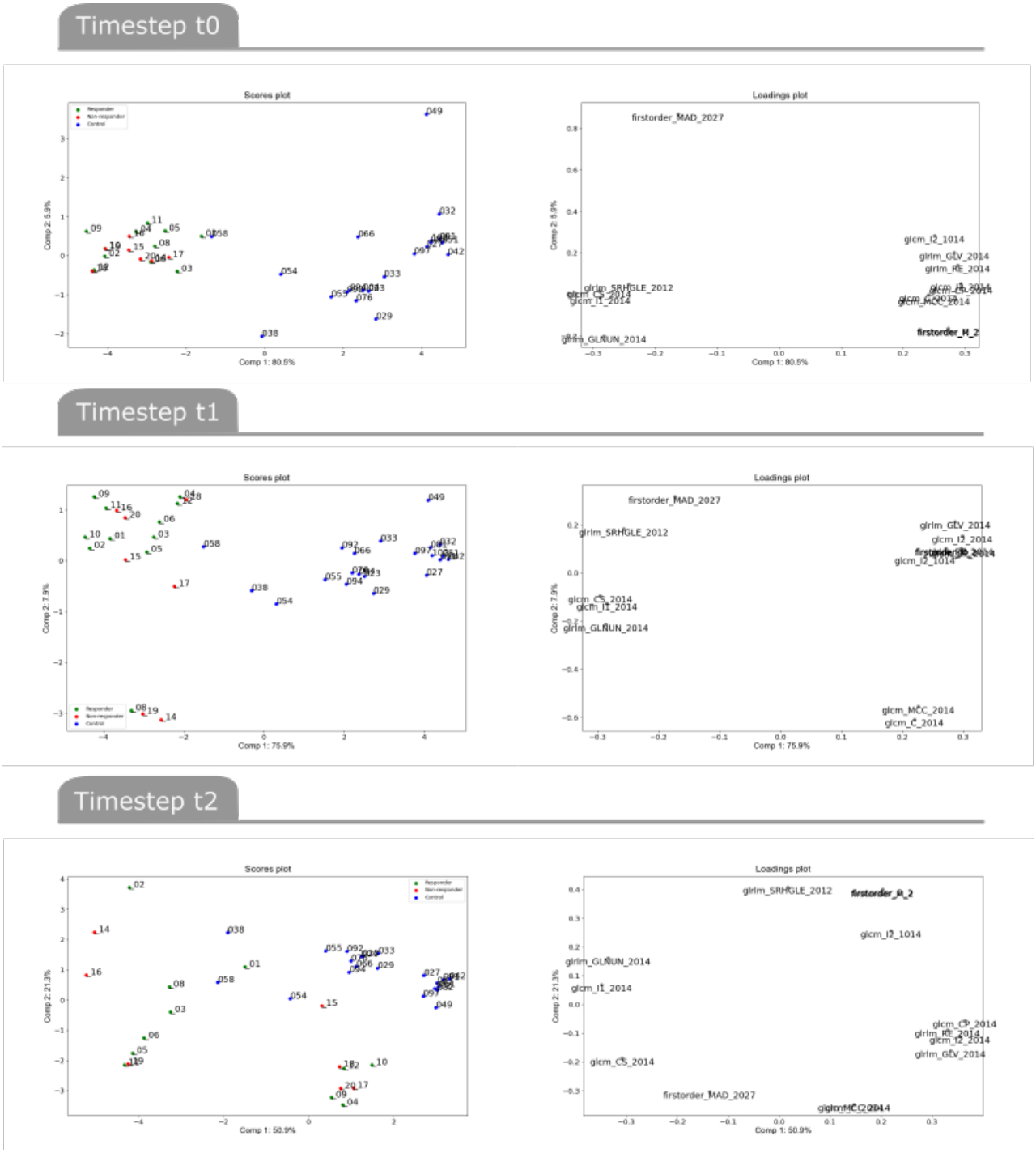

Figure C1: This figure shows the scores and loadings plot from a PCA conducted on the dataset consisting of patients and controls at timesteps, t0, t1, and t2. The dataset contained features that the RENT model selected in any split, in total 14 features.

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

Postboks 5003
NO-1432 Ås
Norway