A comparison of machine learning techniques for combining heterogeneous neuroimaging data sources in the context of classifying individual clinical status

Hovius, Luuk

Faculteit der Sociale Wetenschappen

A comparison of machine learning techniques for combining heterogeneous neuroimaging data sources in the context of classifying individual clinical status

Master's Thesis

Luuk Hovius

Master's Thesis Methodology and Statistics Master

Methodology and Statistics Unit, Institute of Psychology,

Faculty of Social and Behavioral Sciences, Leiden University

Date: June 2021

Student number: 1542338

Supervisor: Dr. Tom F. Wilderjans & Dr. Jenny Ceccarini (KU Leuven)

Abstract

The ability to correctly classify an individual's clinical status could help pave the way for early treatment programs for disorders and illnesses of mental- and physical nature alike. Neuroimaging data could serve as a basis in reaching this goal of accurate classification. The use of said data, however, does come with challenges. A prominent one of which is the fact that neuroimaging data is highly dimensional, meaning that the amount of features largely exceeds the number of subjects within the data set. Furthermore, research has indicated that the use of heterogeneous, but complimentary, data derived from multiple modalities can be an asset to a model used in a classification setting (Zhang et al., 2010; Schouten et al., 2016). The challenge arises in how to combine the different modalities within a single model. A solution could be the use of machine learning algorithms which search for patterns in the data to draw conclusions. Current literature is, however, lacking meaningful comparisons between different machine learning techniques.

Within this project, three different algorithms (support vector machines, Gaussian process classification and multiple kernel learning) have been selected to get insight into (1) whether or not machine learning is able to cope with challenges in the use of neuroimaging data, (2) the difference in performance between these methods and (3) whether or not the use of multiple modalities leads to better results in classification. To this end, 16 alcohol-dependent respondents have been selected along with 32 age-matched healthy controls and have been subjected to both MRI and PET. Models have been trained on data from both separate modalities and on data combining the two modalities. The performances of the models have been assessed by leave-one-subject-out cross-validation and expressed in balanced accuracy and area under the curve. Results indicate that the chosen methods are effective in overcoming challenges arising in the use of neuroimaging data as a means of classification. High balanced accuracies have been found ranging from 76.56% (GPC using PET data) to 100% (GPC using MRI data). Different situations are cause for different solutions and the right choice of algorithm seems to be dependent on, for instance, the fact if either unimodal- or multimodal data is used. Also, settings/optimization of parameters within the model can make a large impact on accuracy. It is therefore advised that researchers try different algorithms and settings before selecting a technique. Different options need to be weighed in order to receive the best possible outcome.

Contents

1. Introduction

In 2017, a study published in Nature made headlines around the world (Gallaghe, 2017; Sample, 2017). In this study, infants of varying ages (6 to 24 months) were subjected to MRI scans and tests regarding their intelligence. It was hypothesized that, even at such a young age, the brain of a child whom would develop an autism spectrum disorder (ASD) would show signs of hyper expansion of the cortical surface area. This expansion is said to be temporally linked to the emergence of the defining behaviors of ASD. A machine learning algorithm (support vector machine) had been used in the classification of 106 infants with a high risk of developing ASD and 42 infants with a low risk of developing ASD. In the end, the analysis yielded a positive result with a predictive value of 81% and a sensitivity of 88% (Hazlett et al., 2017). Findings such as these and developments in the ability to predict individual clinical status could help pave the way for early treatment programs for disorders and illnesses of mental- and physical nature alike.

Even though the use of neuroimaging as a means for the prediction of individual clinical status looks to be quite promising, it does come with its challenges. First of which being high dimensionality of the data (i.e., the dataset containing many features, more than the number of cases used in the classification). When using neuroimaging data, derived from a structural MRI-scan for instance, this is going to be the case. The data, for example, gives an indication of the amount of gray matter per voxel. Since you receive data from tens of thousands of voxels, high dimensionality of the dataset will be something that needs to be taken into account. Datasets containing many features in comparison to the amount of observations face what is also known as the 'curse of dimensionality' (Bellman, 1961). A high amount of features poses two main problems with regard to classification. The first one being the risk of massively overfitting the model to the data, which is detrimental for the generalizability of the obtained results. The other problem which arises has to do with the classification of data points. Many techniques regarding classification are largely based on distances between data points, defined in, for example, Euclidian distances. If the amount of features become very large, the Euclidian distances between the points will tend to appear (nearly) equal due to the large amount of features being incorporated in the hyperspace used to make distinctions. This phenomenon makes clustering and classification based on (Euclidian) distance problematic.

A second challenge which may arise is finding a way to deal with features derived from multiple modalities. Using multiple modalities may provide the analysis with heterogeneous, but – hopefully– complimentary information (features) about the processes or structures under analysis.

Using information gathered from multiple modalities (e.g., both MRI- and fMRI data) could help improve the performance of classification of clinical status, as for example had been indicated by a study by Zhang et al. (2010). Within this study, 51 patients suffering from Alzheimer's Disease were included, as well as 99 patients with mild cognitive impairment (a prodromal stage of Alzheimer's disease, MCI) and 52 healthy controls. All respondents had been subjected to the following three different measurements of biomarkers (after inclusion): (1) structural magnetic resonance imaging (MRI), in order to get an indication of brain atrophy, (2) functional imaging in the form of FDG-PET for hyper metabolism quantification and (3) a cerebrospinal fluid sample to allow for the quantification of specific proteins. Using a multiple kernel support vector machine, the classification that uses the multiple modalities of patients who suffer from Alzheimer's disease, was compared to a support vector machine which uses just one modality. The results showed that the classification that uses multiple modalities outperforms any of the ones that use the single biomarker measurements.

In another study, also focusing on the classification of patients suffering from Alzheimer's disease versus healthy controls, similar results were found (Schouten et al., 2016). In this study, elastic net classification was used to build a classification model based on six different biomarker modalities, all measured by means of magnetic resonance imaging; (1) grey matter density (GMD), (2) white matter density (WMD), (3) fractional anisotropy (FA), (4) mean diffusivity (MD), (5) full correlations between ICA components (FC) and (6) regularized partial correlations between ICA components (PC). Just like in the previously mentioned study, the model using multiple modalities scores best across all indicators of the models' performance. These results indicate that multiple modalities provide each other with complementary information, hence increasing the models' classification performance. However, an appropriate means of combining the different modalities must be devised which is where the challenge arises.

Machine learning techniques could be a way of tackling these previously mentioned challenges which (could) arise in the classification of individual clinical status on the basis of neuroimaging data. These techniques use patterns (i.e., combinations of voxels) in the data to draw eventual conclusions. The goal of supervised learning (in which the class labels of the training data are known beforehand), is to build a concise model of the distribution of class labels in terms of predictor features (i.e., neuroimaging data). The resulting classifier is then used to assign class labels to the testing cases of which the values of the predictor variables are known, but the corresponding class label is unknown (Kotsiantis et al., 2007).

Machine learning is most suited for situations where no 'clear rule of classification' can be defined or coded. This usually happens in situations where a high number of features/predictors are present within the dataset. In these situations, as is the case when using neuroimaging data, where patterns of values in features tend to overlap (especially when multiple data sources are included), machine learning can effectively tackle this problem. Also, the performance of machine learning algorithms tend to increase as the program is being fed more information (i.e., more features and cases). Since neuroimaging data contains a lot of variables (one value per voxel), machine learning is a viable option in the current situation. Incorporating multiple modalities within the same model only (highly) increases the amount of information used within the analysis.

A disadvantage of these same techniques is the interpretation of the 'decision making process' behind the classification. A machine learning model creates a black box of sorts between inputs and predictions. The algorithm can assign a case to a certain class that it is most likely to belong to (based on the rendered calculations by the algorithm), but it does not explain how it came to this conclusion (i.e., based on which specific predictors). This fact will, however, not be a problem in the goal of predicting/classifying individual clinical status as the goal is to open up possibilities for (early) treatment programs. In light of this goal, The underlying rules and patterns of voxels at which these classifications are based on are of lesser importance.

In recent years, research has been conducted into the predictive power of various machine learning techniques. Three different algorithms, of which the effectiveness has been proven within said studies (which will be discussed in the methodology section), and which also support the inclusion of multiple modalities, are: (1) support vector machines (SVM), (2) Gaussian process classification (GPC) and (3) multiple kernel learning (MKL).

As mentioned before, the classification problems at hand pose two main challenges. The first of which being the inclusion of complementary, heterogeneous data (multiple modalities) into a single analysis. The second challenge stems from the fact that neuroimaging data contains more features than cases (high dimensionality). The three different algorithms as mentioned above pose possible solutions in tackling these challenges. First of all, the machines allow for the inclusion of multiple modalities within a single analysis. The machines use (possible) patterns within the data (features) in order to make distinctions. Different modalities can be added to the set of features from where the machine draws said patterns. The multiple kernel learning algorithm could especially excel in these situations as it allows for the computation of a specific kernel for each of the modalities.

As for the challenge of high dimensionality, all three machines are able to perform in these situations, which isn't the case for all forms of machine learning. This is due to the fact that they use the complete set of data to find (possible) patterns within the given set of features. Especially the support vector machine tends to perform well when using high dimensional data. The combination of the fact that these algorithms allow for the use of multiple modalities and the use of high dimensional data, is why they are chosen as a basis for comparisons within this project.

Even though the previously mentioned techniques seem promising in the classification of individual clinical status, current literature is lacking comparisons between these different approaches. The machine learning techniques have all been used in the classification between healthy controls and those suffering from (prodromal) Alzheimer's disease (Challis et al., 2015; Davitzikos et al., 2008; Ferreira et al., 2018; Youssofzadeh et al., 2017; Zhang et al., 2010). The results within these studies could give some indication of differences in the performance of the models, but due to varying designs, respondents and datasets, meaningful comparisons cannot be comprised in a straightforward way. In order to receive such insights, models using the different algorithms need to be fit to a single dataset.

Also, the impact of using heterogeneous, but complementary (multimodal) data needs to be studied in a similar manner. In order to do this, the same means of analysis needs to be put to both a unimodal dataset and a multimodal one (consisting of scans rendered from the same participants). To make this possible, the same set of respondents need to be subjected to varying types of neuroimaging data. These will be the main topics within this project. In conclusion, the research questions posed within this thesis are as follows:

- Are the different algorithms able to cope with the challenges of high dimensionality and the use of heterogeneous but complimentary data (multiple modalities)?
- To what degree do the performances of the different classification methods differ from each other?
- Does the use of heterogeneous but complimentary data (multiple modalities) lead to better results in classification compared to using a single modality, and if so, to what extent?

2. Methodology

In order to get insight into possible answers to the research questions, information has to be gathered into the workings of the different algorithms. For example: What are the advantages and disadvantages of each of the techniques? Within this section, the different machine learning techniques and how they've been used within previous research are the first thing that shall be expanded upon. Afterwards, details regarding the data and sample shall be given. Finally, the process of computing, optimizing and assessing the performance of the models will be discussed.

2.1 *Machine learning classifiers*

2.1.1 *Support vector machines*

The support vector machine (SVM) is an example of a supervised machine learning technique. In order to make classifications between different groups, the algorithm of the support vector machine fits a hyperplane in an N-dimensional space (with N being the number of features) that separates the data points (as good as possible). To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective of the algorithm is to find a plane that holds the maximum margin (i.e., the maximum distance between data points of both classes). The larger the margin distance, the easier it is to discriminate between the groups. A large margin distance provides some reinforcement so that future data points (test data) can be classified with more confidence since a slightly higher/lower value on the feature will still not have it overlap with the other group. Support vectors are data points that are closest to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, the method maximizes the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help build the SVM. In real life situations, however, it is unrealistic that data point are perfectly separable by group on each of the features. Some overlap is to be expected. The algorithm allows for some error made by overlap named the soft margin parameter. The goal of the program is to maximize the margin distance whilst minimizing the soft margin parameter.

The main advantage of the use of the support vector machine lies in its effectiveness in high dimensional spaces (i.e., when the number of features is higher than the amount of observations). Also, due to the fact that it uses only certain points of the training data (the support vectors) in the decision making, it is also computationally efficient. A major disadvantage of the SVM is that it doesn't perform as well when the data contains (a lot of) noise. This happens when the classes of the data points overlap each other to a large extent (i.e., a high value on the soft margin parameter). Also, the support vector machine does not provide estimations of probability of class memberships. In order to receive probabilistic explanations, calculations must be rendered using cross-validations which can be computationally expensive.

SVM has already been used in several studies. As mentioned before, mild cognitive impairment (MCI) is generally regarded as a transitional period between the cognitive changes of normal aging and the earlier changes related to Alzheimer's disease. In a study by Long et al. (2016), multiple variations of magnetic resonance imaging (MRI) were used as a means to discriminate between patients suffering from MCI (N = 29) and healthy controls (N = 33). A support vector machine approach was chosen as a means of classification between the two groups. After training the model, leave-one-out cross-validation had been used to evaluate the performance of the model. A classification accuracy of up to 96.77% was found. In another example regarding the classification of patients with prodromal Alzheimer's disease (MCI) versus healthy controls, (multi-modal) structural data obtained by magnetic resonance imaging (density of grey-matter, density of white matter and cerebrospinal fluid) was used to make a distinction between the two groups (Davatzikos et al., 2008). Thirty elderly individuals, fifteen of whom were suffering from MCI, were subjected to a scan as part of the Baltimore Longitudinal Study of Aging neuroimaging substudy (Resnick et al., 2000). Using a support vector machine, a model was trained. Leave one-scan-out cross validation had been used to assess the predictive power of the model and a classification accuracy of 90% was found.

Besides classifications regarding Alzheimer's disease and MCI, the support vector machine has also been proven useful in the classification of individual clinical status of other diagnoses. Impairments in executive function and language processing are characteristic of both schizophrenia and the bipolar disorder. In a study posted by Costafreda et al. (2011), 32 patients with schizophrenia in remission, 32 patients with bipolar disorder in an euthymic state and 40 healthy controls underwent functional magnetic resonance imaging (fMRI) whilst performing a phonological verbal fluency task. During this task, both patient groups (in comparison to the healthy controls) showed increased activation in the anterior cingulate, left dorsolateral prefrontal cortex and right putamen.

support vector machine classification was used and assessed by leave-one-out cross-validation. In conclusion, SVM was able to correctly classify schizophrenic patients with a sensitivity of 91% and specificity of 92%. Some of the bipolar patients (12 out of 32) were misclassified as healthy controls, resulting in a lower accuracy (79%) with a sensitivity of 56% and a specificity of 89%.

2.1.2 *Gaussian process classification*

The Gaussian process classifier (GPC) is another example of a supervised machine learning technique. The GPC classifies a case to a certain class based on a probability. This probability is related to an unconstrained latent function defining the covariance function of the data, and is estimated based on the data onto which the algorithm is trained. The relationship of a data point to this latent function is quantified using the probit transformation (Kuss et al., 2005). When new data (test cases) are introduced, their values on the latent function are again determined and likelihoods are quantified using the same transformation and their relationship to the training data. Based on these calculations, the new cases are assigned to one of the groups (i.e., modal assignments). Namely the group with the largest probability of the case belonging to said group. An advantage of the Gaussian process classifier is that since the prediction is probabilistic, empirical estimations of said probability can be computed. A disadvantage of this machine is that the model can lose efficiency in high dimensional spaces (when the number of features greatly exceeds the number of cases).

GPC has been applied successfully in previous studies. In the classification of mild cognitive impairment and Alzheimer's disease versus healthy controls, a study by Challis et al. (2015) uses Gaussian process logistic regression (classification) on resting-state data of the brain, obtained by means of functional magnetic resonance imaging (fMRI) to make distinctions between the different (patient) groups. In total, data from 116 participants had been gathered and analyzed, 39 of which were healthy controls, 50 with a diagnosis of MCI and 27 patients suffering from Alzheimer's disease. In the end, the implemented model achieved a 75% accuracy in the discrimination of MCI versus healthy controls. In the classification of patients suffering from AD versus those with MCI, a 97% level of accuracy was found.

Gaussian process classification has also been used for other means than the classification of patients with AD/MCI. Post-traumatic stress disorder (PTSD) has been found to be associated with decreased regulatory activation from the medial prefrontal cortex and increased activation of the limbic system. In a study by Andrew et al. (2018), the amplitude of low frequency fluctuation was measured by fMRI in order to map the brain's activation in a resting state of 181 participants. 81 of these respondents were diagnosed with PTSD, 49 with PTSD of the dissociative subtype and the remainder functioned as the control group (N = 51). A multiclass Gaussian process classifying algorithm had been used to make the distinctions between the three groups. An overall accuracy of 91.63% was found, based on correct classification of healthy controls (87.50%), PTSD (94.81%) and PTSD+DS (89.90%).

2.1.3 *Multiple kernel learning*

The multiple kernel machine learning technique aims at simultaneously learning a kernel and the associated (relevant) predictors in a supervised learning setting (Rakotomamonjy et al., 2008). A kernel serves two distinct functions: it defines the similarity between two examples (i.e., regions of interest), whilst defining an appropriate regularization term for the learning (classification). Combinations of kernels derived from multiple sources are used in the overall goal of increasing the models' accuracy.

The main advantage of MKL arises in situations where multimodal data (heterogeneous but complimentary) is used as input within one model. The algorithm defines a specialized kernel for each of the modalities, along with kernels for each of the ROI's as defined by an atlas (per modality). The combination of these specialized kernels could lead to an enhancement of the accuracy when compared to only using a single modality. A disadvantage of multiple kernel learning is that the technique only works on datasets containing a relatively small set of features (e.g., 50 per modality). Since the amount of features used within this project greatly exceeds that amount, an atlas, indicating regions of interest (and therefore sets of voxels), is included when using this technique. The incorporation of the atlas limits the amount of kernels that need to be specified (per modality) to the amount of ROI's as defined by the atlas. Another disadvantage of multiple kernel learning seems to be that results are less interpretable than other techniques and computationally expensive (as to evaluate the model output you need to evaluate all of the base kernels).

MKL has also been applied in the classification of patients suffering from Alzheimer's disease from those with mild cognitive impairment and healthy controls. In a study by Youssofzadeh et al. (2017), multimodal data was used in the classification of 286 participants (AD N=58, MCI N = 108, HC N = 120). Data gathered by magnetic resonance imaging (MRI) and by positron emission tomography (PET) was combined and analyzed with the multiple kernel learning algorithm. The analysis yielded positive results for each of the three classification problems. For AD vs HC, it posed a balanced accuracy of 95.7%, 95.1% in the discrimination between MCI vs HC and the balanced accuracy for AD vs MCI was set at 95.1%. In order to get insight into the (possible) effect that the combination of the modalities had, unimodal analyses were run on the PET data by means of a support vector machine. The balanced accuracies were found to be 90.07%, 81.6% and 79.59% for the three classification problems respectively, underlining the previously mentioned positive effect found by combining heterogeneous but complementary data into a single analysis.

A great challenge in the early treatment of mood disorders currently is making the differential diagnosis between a major depression disorder (MDD) and a bipolar disorder (BD), as 60% of depressed bipolar patients are initially misdiagnosed with MDD (Goodwin, 2012). In a recent study, multiple kernel learning had been used successfully in the classification of individual clinical status between three groups of respondents. 74 patients with a current MDD, 74 with bipolar disorder (type 1) and 74 healthy controls (all of similar demographics) had been included. Multiple modalities of structural neuroimaging data had been gathered and analyzed using the machine learning technique, rendering a balanced accuracy of 73.65% (Vai et al., 2020).

2.1.4 *Overall expectations for the situation at hand*

Even though the techniques have proven effective in the prediction of individual clinical status (as in the examples listed above), each come with their own advantages and disadvantages. The choice for the right algorithm might therefore vary over different situations. Within this thesis project, three different classification problems shall be addressed. First, the classifications based on MRI data, secondly discriminations based on PET data and lastly a classification problem using both modalities together. With regard to the unimodal analyses, I would expect the support vector machine to perform better than the other techniques since its main advantage is the fact that it tends to perform well in situations of high dimensionality (as is the case when using neuroimaging data).

The Gaussian process classifier on the other hand, is said to underperform in this situation. When using both modalities of data simultaneously (multimodal analyses), multiple kernel learning might be the best course of action since the algorithm specifies different kernels for each of the modalities, which might improve the overall classification performance.

2.2 *Data and sample*

16 subjects with a diagnosis of alcohol dependence (DSM-IV) have been included through the recruitment by a board-certified psychiatrist (specializing in substance dependency) at the Psychiatric Hospital 'Alexianen Tienen' and the Psychiatry department of the University Hospital Leuven. All participants have been included within their first two weeks of supervised abstinence. 32 healthy controls were recruited through local advertisement and randomly selected based on age. Each alcohol-dependent participant had two age-matched controls. Healthy controls consuming more than seven units of alcohol per week or drank more than five units in one sitting regularly, were excluded from the study. Out of the alcohol-dependent respondents, three out of sixteen are female and the mean age within the group is 46 years (SD = 8). In case of the healthy controls, fourteen are female (N =32). The age of the healthy controls has a mean of 45 with a standard deviation (SD) of 13. Each participant of both the alcohol-dependent group, and the healthy controls underwent PET-imaging and structural MRI (Leurquin-Sterk et al., 2018).

Magnetic resonance imaging (MRI) is a non-invasive imaging technology that produces three dimensional, detailed anatomical images and is often used for disease detection, diagnosis and treatment monitoring. It is based on sophisticated technology that excites and detects the change in the direction of the rotational axis of protons found in the water that makes up living tissues. Data represents the amount of grey matter per voxel. As for PET (positron emission tomography), a radioactive tracer is injected, inhaled or swallowed before the scan. The tracer collects in areas in the brain where high chemical activity is taking place, and is measured/quantified by the scan. This type of scan has both a structural and a functional use.

The PET data at hand has already been used in the classification of individual clinical status of the two groups. A support vector machine approach had been used to discriminate between the two groups and its performance was assessed by leave-one-out cross-validation. In the end, a balanced accuracy of 79.7% had been found (Devrome et al., n.d.). It will be interesting to see if optimizing the parameters of the model, using a different algorithm, or combining multiple modalities will have impact on this value of balanced accuracy (and therefore the performance of the model).

2.3 *Implementation of the analyses*

2.3.1 *PRoNTo*

The 'Pattern Recognition for Neuroimaging Toolbox (PRoNTo)' (Schrouff et. al., 2013) is a toolbox within the MATLAB environment which supports all three of the previously mentioned machine learning techniques. In PRoNTo, brain images are treated as spatial patterns and statistical learning models (machine learning algorithms) are used to identify statistical properties of the data that can be used to discriminate between groups of subjects (supervised classification). The program uses NIFTI files as input which are mostly designed to analyze structural magnetic resonance imaging (MRI) and positron emission tomography (PET) data. Also, PRoNTo includes the possibility to incorporate multiple modalities within a single model, making the program ideal in the current situation. The effectiveness of the program has been proven various times over multiple different study designs aiming to classify individual clinical status (Ferreira et al., 2018; Fernandes et al, 2020; Portugal et al., 2019; Ranlund et al., 2018). In order to reach the best possible outcome of the different models, settings on certain options within the toolbox need to be optimized. The three most influential ones shall be discussed below.

2.3.2 *Scaling*

The scaling option within PRoNTo allows for the specification of a constant value to scale each scan. The range of the data is then adjusted to a maximum of the indicated value. During the feature set preparation phase of the analysis, a vector containing constant values can be added containing this constant value. The vector must be of the same size as the number of scans within the modality. In case of modalities with a (possible) large variation in range across different subjects, such as PET, this step (adding a value for scaling the data) is required since it ensures the convergence of the machine learning algorithm. As mentioned before, PET data represents the concentration of a radioactive tracer in a certain voxel. This value can range from 0 to several hundreds and therefore cannot always be compared across voxels and subjects in a straightforward way. The scale of the MRI data, however, is more comparable across voxels and subjects since values in cells represent the percentage of grey matter per voxel (expressed on a scale from 0 to 1). Due to this fact, scaling is not a mandatory step in the use of MRI data as it is for PET data. Within this thesis project, multiple (constant) values for scaling of the PET data shall be applied within the analyses in order to find the optimal value that renders the best possible classification accuracy. In particular, the following four values are used: 0.1, 1, 50 and 100. For the analyses using MRI data, the effect of using no scaling shall also be put to the test. This won't be the case for the PET data since scaling is a mandatory step in analyses using said scans.

2.3.3 *Masks*

A mask is used to optimize the feature selection of the different datasets. The PRoNTo toolbox allows for the specification of both a first-level mask and a second-level mask. The first-level mask is used to discard uninteresting features (voxels) within the dataset, such as areas outside of the brain. It is compulsory to add a first-level mask to each modality (it can be the same for each of the modalities), in order to render a feature set. The inclusion of a second-level mask isn't mandatory but can be helpful regarding certain research questions. The second-level mask can, for example, restrict the analysis to certain areas of the brain. Within this project, the effect on performance of two different first-level masks and the inclusion of a second-level mask will be put to the test. All different masks have been provided by Dr. Ceccarini to be used within this project.

The two provided first-level masks are different versions of the same one, discarding the extra-cerebral signals (such as the skull). The difference between the two is a variation in appointed voxel size. Within this project, they will be named Mask-A and Mask-B in order to avoid confusion. Mask-A shall be included in each of the models unless otherwise specified, since the inclusion of a first-level mask is mandatory. The second-level mask is defined by the ROI's where significant decreased mGLuR5 binding has been found in alcohol-dependent respondents in comparison to healthy controls in a study that uses the same datasets as used within this project (Leurquin-sterk et al., 2018).

2.3.4 *Atlases*

A brain atlas divides an image into certain sections (combinations of features), indicating different parts of the brain (representing regions of interest, ROI's), and allows for a specific kernel to be made for each. Also, the computation of a weight map (in order to get insight into the contribution of certain voxels/regions of interest to the performance of the model) is based on the regions specified in the atlas. The inclusion of an atlas is a necessary step in order to render a model based on multiple kernel learning (MKL). The MKL-algorithm specifies a specific kernel for a set of features as specified by the atlas. Without predefined sets of features (regions of interest), the program tries to calculate a kernel for each separate feature, which isn't possible when the original data is used as it contains too many voxels. This is why the inclusion of an atlas is mandatory when using this particular machine. Two different atlases are offered within PRoNTo. Also, two different atlases have been provided by Dr. Jenny Ceccarini for this thesis project. The effect of the four atlases will be put to the test to get insight into their effect on the performance of models using multiple kernel learning. The atlases are named as follows:

1. The AAL1-atlas (included within PRoNTo)
2. The Brodmann-atlas (included within PRoNTo)
3. The AAL2-atlas (provided by Dr. Ceccarini)
4. The Hammers-atlas (provided by Dr. Ceccarini)

Note that two different versions of the AAL-atlas shall be used. Even though the two atlases define the same (number of) regions of interest, due to a difference in voxel size as defined by the atlases, the boundaries of the regions vary between the two.

2.3.5 *Assessing the models' performance and interpretation*

The performance of the different models will be quantified and compared using two different performance measures. The first of which being the balanced accuracy (BA). The balanced accuracy is calculated by taking the average of the proportion of correct classifications of each class individually. In other words, it is computed by adding the sensitivity and the specificity to each other and dividing said number by two, which gives a value between 0% and 100%. An advantage of the use of this measure is that it gives a good indication of performance, even when the amount of subjects in each class is imbalanced (i.e., the number of cases belonging to one class being –much– larger than the other), as is the case within this project.

As for the second measure, the area under the curve (AUC) is a diagnostic measure for the performance of binary classification, scaled on a range between 0 and 1. This measure calculates the area under the receiver operating characteristic curve (ROC), a graph where two parameters are plotted. The first being the true positive rate (the proportion of correctly identified alcohol-dependent respondents) and the second being the false positive rate (the proportion of healthy controls misclassified as belonging to the alcohol-dependent group). The parameters are plotted using several threshold settings. This threshold represents the boundary between the two groups given the features in the data as calculated by the algorithm. The higher the area beneath the ROC-curve (and therefore the higher the value on AUC), the better the performance of the model. Besides the balanced accuracy (BA), the value of the AUC shall be given for each of the models when displaying results. Since the ROC curve is plotted using different thresholds (i.e., variations on the boundary as calculated by the algorithm) and the summary across the different thresholds is reported in area under the curve, it is possible that the value of AUC will represent a better performance than indicated by the balanced accuracy.

In order to test the generalization ability of the performance of the different models, leave-one-subject-out cross-validation (LOSO) is employed when running the different analyses. This step is necessary in order to estimate the test error of a model. When running the model, the parameters are fit to a partition of the data (the training set), consisting of all scans except for one subject (either one scan or two, depending on either a unimodal or multimodal analysis). After the model is put to the training data, the performance is assessed by using it to classify the unseen data/scan(s). If the data is repartitioned repeatedly (i.e., leaving a different subject out at each split), it is possible to approximate the generalization error of the model by calculating the average BA or AUC across the different partitions of the data.

Permutation testing is a non-parametric procedure, aimed to obtain meaningful *P*-values for (in this case) class-specific accuracy's and the balanced accuracy (BA) of the different models. The process behind the permutation test is to redo the cross-validation procedure a certain number of times (R), herewith randomly shuffling the subjects across the different groups (alcohol-dependent and healthy controls). As such, based on the R permutation samples, an empirical sampling distribution of plausible values of the balanced accuracy under the null hypothesis is obtained; this allows for the computation of a *P*-value for this statistic indicating the probability that this statistic would be at least as extreme as we have observed, if the null hypothesis is true. The null hypothesis in this situation being that no distinction between the two groups can be found based on the data at hand. The smallest increment in *P*-value is equal to 1/R. Within this thesis project, the final models (after optimizing other parameters such as the proper scaling values) will be rendered a last time. For those models, *P*-values shall be computed using 300 permutations, giving a lowest possible *P*-value of .0033.

PRoNTo allows for the computation of weight maps which can help with the interpretation of the rendered models. An atlas (as mentioned before) is included to indicate regions of interest within the brain (ROI's). Weight maps are images which indicate which of the pre-defined regions of interest (by the atlas) contribute (most) to the performance of the model, and to what degree. Also, during this computation, tables are provided displaying percentages of the contribution of the different regions to the discriminative power of the classification model. For the final (best performing) MKL models for each of the classification problems, weight maps shall be computed, and presented/discussed in the results section. The decision to use the MKL analyses as a basis for the weight maps is due to the fact that each of these analyses incorporates an atlas, whereas the original data (with voxels instead of ROI's) will be used for the support vector machines and Gaussian process classifiers.

3. Results

As has been previously mentioned, in order to make meaningful comparisons between the performances of the different machine learning algorithms, the value of parameters and settings need to be optimized. Three different comparisons shall be made; (1) One between the algorithms using the structural MRI data, (2) one using the PET data and (3) one involving multimodal comparisons between the machines' performances. For each of these comparisons, different values for scaling the data are explored. After finding the optimal value for scaling (for each of the machines), different atlases have been tried out (a necessary step when using multiple kernel learning). Using this information, different combinations of masks have been applied to eventually find optimal settings and values for the different parameters. Using these optimal combinations, final results are rendered and compared on the performance measures and weight maps (to investigate the importance of different regions of interest) are inspected.

3.1 *Unimodal analyses using MRI data*

The first analyses that will be compared to one another are the unimodal ones using the structural MRI data. As discussed above, the first step in the process had been exploring different values for scaling. For these computations, Mask-A has been used and no second level mask has been included. The impact of using different masks shall be discussed later on.

Table 1
Unimodal scaling experiments for GPC and SVM using MRI data

Machine	Modality	Atlas	Scaling	CV-Scheme	Balanced Accuracy	AUC
GPC	MRI	-	-	LOSO	100%	1.00
SVM	MRI	-	-	LOSO	98.44%	1.00
SVM	MRI	-	1	‘’	98.44%	1.00
SVM	MRI	-	50	‘’	98.44%	1.00
SVM	MRI	-	100	‘’	98.44%	1.00

Table 1 displays the results of the scaling experiments for both the Gaussian process classifier and the support vector machine algorithm. In case of the GPC, an optimal outcome (with a 100% classification performance) had been found at the first try. As for the support vector machine, the value of scaling did not seem to result in a change in performance. A high value of accuracy had been found nonetheless, indicating promising results for the classifiers.

Table 2

Atlas- and scaling experiments for MKL using MRI data

Machine	Modality	Atlas	Scaling	CV-Scheme	Balanced Accuracy	AUC
MKL	MRI	AAL	-	LOSO	48.44%	0.62
MKL	MRI	AAL	0.1	‘’	48.44%	0.61
MKL	MRI	AAL	1	‘’	Error	Error
MKL	MRI	AAL	50	‘’	92.19%	0.96
MKL	MRI	AAL	100	‘’	92.19%	0.96
MKL	MRI	Brodmann	-	LOSO	46.88%	0.32
MKL	MRI	Brodmann	0.1	‘’	53.13%	0.68
MKL	MRI	Brodmann	1	‘’	Error	Error
MKL	MRI	Brodmann	50	‘’	96.88%	1.00
MKL	MRI	Brodmann	100	‘’	96.88%	1.00
MKL	MRI	Hammers	-	LOSO	71.88%	0.77
MKL	MRI	Hammers	0.1	‘’	42.19%	0.51
MKL	MRI	Hammers	1	‘’	71.88%	0.77
MKL	MRI	Hammers	50	‘’	98.44%	0.97
MKL	MRI	Hammers	100	‘’	98.44%	0.97
MKL	MRI	AAL2	-	LOSO	48.44%	0.70
MKL	MRI	AAL2	0.1	‘’	48.44%	0.59
MKL	MRI	AAL2	1	‘’	48.44%	0.70
MKL	MRI	AAL2	50	‘’	98.44%	0.97
MKL	MRI	AAL2	100	‘’	98.44%	0.97

Table 2 shows both the results for the scaling experiments and the impact of the different atlases for the multiple kernel learning algorithm using the MRI data. Note that for two of the rendered analyses, the value of balanced accuracy and area under the curve has not been included as PRoNTo indicated that the computation of these statistics is not possible due to improper scaling. Out of the different atlas options, the value of scaling seems to have the least impact on the performance when using the Hammers-atlas. The largest impact of scaling is found when using the Brodmann-atlas. In general, using a too low scaling value (no scaling or 1) leads to a worse classification that using a larger scaling value (50 or 100). When a large enough scaling value is chosen, recovery is very good (more than 90%) for all atlases.

After optimizing the value for scaling the data, a perfect value (of 1) on area under the curve has been found when including the Brodmann-atlas. The same values on BA and AUC, and overall best (combination of) results have been found when using either the Hammers- or the AAL2-atlas. For further analyses, it has been chosen to use the AAL2-atlas since this atlas has also been found to be the best option when using the PET- or multimodal datasets (Table 7 and Table 13).

Table 3

Mask experiments for unimodal analyses using MRI data

Machine	Modality	Atlas	Scaling	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	AUC
GPC	MRI	-	-	Mask-A	-	LOSO	100%	1.00
GPC	MRI	-	-	Mask-B	-	‘’	100%	1.00
SVM	MRI	-	-	Mask-A	-	LOSO	98.44%	1.00
SVM	MRI	-	-	Mask-B	-	‘’	98.44%	0.98
SVM	MRI	-	-	Mask-B	YES	‘’	98.44%	0.97
SVM	MRI	-	-	Mask-A	YES	‘’	98.44%	0.97
MKL	MRI	AAL2	50	Mask-A	-	LOSO	98.44%	0.97
MKL	MRI	AAL2	50	Mask-B	-	‘’	98.44%	0.97
MKL	MRI	AAL2	50	Mask-B	YES	‘’	92.19%	0.97
MKL	MRI	AAL2	50	Mask-A	YES	‘’	89.06%	0.97

After finding the optimal settings with regard to scaling and the atlas (for MKL), these values were used as a basis in order to find what mask(s) would result in the best possible outcome. The results corresponding to the models using different options for mask-selection are displayed in Table 3. First of all, the inclusion of the second-level mask seems to lower the computed performances of the different models using MKL. Across most models, the incorporation of either Mask-A or Mask-B does not seem to make a difference. Only in the use of the support vector machine does the incorporation of the Mask-B result in a slightly lower value for AUC. For this reason, Mask-A has been included into the final models for the unimodal analyses based on MRI data.

Table 4
Final outcomes of unimodal analyses of models using MRI data

Machine	Modality	Atlas	Scaling	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	Class AC HC	Class AC ALC	AUC
GPC	MRI	-	-	Mask A	-	LOSO	100% *P* = .0033	100% *P* = .3056	100% *P* = .0033	1.00
SVM	MRI	-	-	Mask A	-	LOSO	98.44% *P* = .0033	96.88% *P* = .0033	100% *P* = .0033	1.00
MKL	MRI	AAL2	50	Mask A	-	LOSO	98.44% *P* = .0033	96.88% *P* = .0033	100% *P* = .0033	0.97

Table 4 displays the final outcomes of the unimodal analyses of models trained on MRI data. Overall, the different methods have all performed really well in discriminating between healthy controls and alcohol-dependent respondents. During these final analyses, permutations (consisting of 300 repetitions) have been rendered. Due to these permutations, *P*-values are added to the table. The classifications and the corresponding statistics seem to be quite stable except for the class specific accuracy of the healthy controls in the GPC analysis. The accuracy of the classification to this class had been set at 100% but apparently this value wasn't extreme in the distribution of plausible values of this statistic (as calculated during the permutation process) under the null-hypothesis.

In the other two models, one classification error had been made. In both cases, one of the healthy controls has been misclassified as belonging to the alcohol-dependent group (which can be seen by the class accuracies in Table 4).

After the final renditions of the models based on unimodal MRI data, weight maps have been computed (images of which can be found in Appendix I). Since the multiple kernel learning method makes a specific kernel for each of the ROI's specified by the included atlas, this is the ideal model to get insight into the contribution of areas of the brain to the discriminative power of the model. Table 5 shows these contributions expressed in percentages (rounded to two decimals). The results posted in the table indicate that especially the angular gyri and hippocampus (in both hemispheres), play a large role in making the distinction between the two groups as they account for almost all of the contribution (i.e., about 98.6%) to the predictive power of the model.

Table 5

Weights of ROI's in finalized MKL-model using MRI data

Region of interest	Contribution to performance
Angular Gyrus Left	43.62%
Hippocampus Right	25.42%
Hippocampus Left	15.18%
Angular Gyrus Right	14.40%
Frontal Inferior Operculum Right	1.39%

Note. Percentages are rounded to two decimals.

3.2 *Unimodal analyses using PET data*

After finalizing the unimodal analyses based on MRI data, the same steps were taken with regard to the unimodal analyses based on PET data. The first step is the search for optimal values for scaling. Note that analyses based on PET data require the inclusion of a scaling value, so a scaling vector has been added to each of the rendered models. The results of scaling experiments for the support vector machine and Gaussian process classifier are summarized in Table 6.

Table 6
Unimodal scaling experiments for GPC and SVM using MRI data

Machine	Modality	Atlas	Scaling	CV-Scheme	Balanced Accuracy	AUC
GPC	PET	-	0.1	LOSO	67.19%	0.79
GPC	PET	-	1	‘’	48.44%	0.00
GPC	PET	-	50	‘’	73.44%	0.76
GPC	PET	-	100	‘’	73.44%	0.77
SVM	PET	-	0.1	LOSO	73.44%	0.78
SVM	PET	-	1	‘’	73.44%	0.78
SVM	PET	-	50	‘’	73.44%	0.78
SVM	PET	-	100	‘’	73.44%	0.78

When looking at the results in Table 6, the first thing to notice is that the overall performances across the different models have been set at lower values than the models trained on MRI data (Table 1). Just as with the unimodal MRI-analyses, the value for scaling does not impact the performance of the support vector machines. The same cannot be said for GPC, where a higher value of scaling seems to improve its accuracy and setting it at the same value as for the SVM. Note that a value of 0.00 on area under the curve has been found when using a scaling value of 1 in the GPC model. When using this value, all of the cases (from both groups) are classified as healthy controls, resulting in a value of 0 on AUC. The next step in the process of optimizing the models is finding the optimal value for scaling in MKL and choosing an appropriate atlas. The results of this search are summarized in Table 7.

Table 7
Atlas- and scaling experiments for MKL using PET data

Machine	Modality	Atlas	Scaling	CV-Scheme	Balanced Accuracy	AUC
MKL	PET	AAL	0.1	LOSO	71.88%	0.63
MKL	PET	AAL	1	‘’	78.13%	0.77
MKL	PET	AAL	50	‘’	60.94%	0.68
MKL	PET	AAL	100	‘’	67.19%	0.75
MKL	PET	Brodmann	0.1	LOSO	Error	Error
MKL	PET	Brodmann	1	‘’	59.38%	0.63
MKL	PET	Brodmann	50	‘’	64.06%	0.70
MKL	PET	Brodmann	100	‘’	70.31%	0.74
MKL	PET	Hammers	0.1	LOSO	71.88%	0.79
MKL	PET	Hammers	1	‘’	71.88%	0.79
MKL	PET	Hammers	50	‘’	59.38%	0.70
MKL	PET	Hammers	100	‘’	57.81%	0.69
MKL	PET	AAL2	0.1	LOSO	78.13%	0.83
MKL	PET	AAL2	1	‘’	78.13%	0.83
MKL	PET	AAL2	50	‘’	78.13%	0.83
MKL	PET	AAL2	100	‘’	75.00%	0.81

Out of the models posted in the table above, the best combination of both balanced accuracy and area under the curve has been found when setting the value of scaling at either 0.1, 1 or 50 and including the AAL2-atlas. For the next step, a scaling value of 50 has been chosen. This decision has been made in order to keep the value the same across the thee machines (the optimal value for scaling in SVM and GPC was also found at 50). In the first model using the Brodmann-atlas, no performance measures have been calculated due to bad scaling. Just like the results reported in Table 6, the performance of multiple kernel learning models based on PET data are lower than the same models based on MRI data (Table 2).

Table 8
Mask experiments for unimodal analyses using PET data

Machine	Modality	Atlas	Scaling	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	AUC
GPC	PET	-	50	Mask-A	-	LOSO	73.44%	0.77
GPC	PET	-	50	Mask-B	-	‘’	64.06%	0.79
GPC	PET	-	50	Mask-B	YES	‘’	73.44%	0.79
GPC	PET	-	50	Mask-A	YES	‘’	76.56%	0.79
SVM	PET	-	50	Mask-A	-	LOSO	73.44%	0.78
SVM	PET	-	50	Mask-B	-	‘’	73.44%	0.80
SVM	PET	-	50	Mask-B	YES	‘’	79.69%	0.83
SVM	PET	-	50	Mask-A	YES	‘’	79.69%	0.84
MKL	PET	AAL2	50	Mask-A	-	LOSO	78.13%	0.83
MKL	PET	AAL2	50	Mask-B	-	‘’	78.13%	0.84
MKL	PET	AAL2	50	Mask-B	YES	‘’	75.00%	0.75
MKL	PET	AAL2	50	Mask-A	YES	‘’	75.00%	0.78

Table 8 shows the results of different renditions using all possible combinations of first- and second-level masks. In these same experiments using (unimodal) MRI data, the best option for all models had been set at the use of Mask-A and no inclusion of a second-level mask (Table 3). The same cannot be said for the analyses displayed in Table 8 for the PET data. For the GPC, the optimal combination seems the inclusion of both the Mask-A and the second-level mask. The same combination of masks came out as the best scoring one when using the SVM algorithm. Note that the value of BA is the same as when using first-level Mask-B, but the AUC value scores (slightly) better when using Mask-A. A slight change in AUC is also what makes the final distinction between combinations of masks in the analyses based on multiple kernel learning. For this algorithm, the best combination seems to be the use of Mask-B without a second-level mask. Out of all renditions, MKL performs best in these analyses based on PET data. Final renditions for these unimodal analyses (including *P*-values for balanced- and class specific accuracies) are displayed in Table 9.

Table 9

Final outcomes of unimodal analyses of models using PET data

Machine	Modality	Atlas	Scaling	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	Class AC HC	Class AC ALC	AUC
GPC	PET	-	50	Mask-A	YES	LOSO	76.56% *P* = .0066	90.63% *P* = .9336	62.50% *P* = .0033	0.79
SVM	PET	-	50	Mask-A	YES	LOSO	79.69% *P* = .0033	90.63% *P* = .0399	68.75% *P* = .0033	0.84
MKL	PET	AAL2	50	Mask-B	-	LOSO	78.13% *P* = .0033	93.75% *P* = .0166	62.50% *P* = .0266	0.84

When looking at the results in Table 9, the first thing to notice is that the discriminative power of models trained on the PET data are set at a lower value than the models trained on the MRI-data (Table 4). Overall, the values of balanced accuracy seem to be quite stable. The same can be said for the class specific accuracy's, except for the classification accuracy of the healthy controls of the GPC-model (*P* = .9336). This rather high *P*-value indicates that the value found in the original analysis isn't extreme in the distribution of plausible values (of this statistic) under the null-hypothesis as created by the permutation samples. The null-hypothesis in this case being that no accurate classification of healthy controls can be made based on the PET data. Across all three models, the class specific accuracies of the alcohol-dependent group are set at a (much) lower value than the class specific accuracies of the healthy controls. It seems that, when using the PET data, a lot of the alcohol-dependent participants are misclassified as healthy controls.

Table 10 displays the contributing regions of interest for the final MKL-rendition on the PET data. The regions that are defined as contributing to the discriminative power of the model vary greatly from those selected when using the MRI data, even though the included atlas has been the same for both models (AAL2). When using the PET data, only both angular gyri are selected as contributing to the model.

Even though both areas are also selected by the model using the MRI data, other regions are left out (Table 5). In an interesting manner, when using the PET data (unimodally), the angular gyrus on the right-side hemisphere is said to have the most impact on the performance of the model. However, when the MRI data is used, the angular gyrus on the left hemisphere is selected as having the most impact on the classifications. Images of the weight maps can be found in Appendix I.

Table 10
Weights of ROI's in finalized MKL-model using PET data

Region of interest	Contribution to performance
Angular Gyrus Right	87.03%
Angular Gyrus Left	12.97%

Note. Percentages are rounded to two decimals.

3.3 *Analyses using multimodal data*

In order to get insight into the (possible) impact of using heterogeneous but complimentary data (multimodal) within a single analysis, a final set of renditions had been made using both the MRI- and PET data together. Another reason for doing so is to receive insight into which of the algorithms is best used in a situation of using multimodal data. Just as with the unimodal analyses, the same parameters need to be optimized, starting with a proper value for scaling (if any). The process of finding this value is, however, slightly different compared to the unimodal analyses since different values for scaling each modality can interact with one another. Since the use of PET data requires the inclusion of a value for scaling the data, the different values have first tried out on this data whilst keeping the MRI data as is (no scaling value is applied). After finding the optimal value for the scaling of the PET data, different values of scaling for the MRI data are put to the test while keeping the scaling for PET data at the same (optimal) value. The results of the scaling experiments for the Gaussian process classifier and the support vector machine are summarized in Table 11.

Table 11
Scaling experiments for GPC and SVM using multimodal data

Machine	Modality	Atlas	Scaling MRI	Scaling PET	CV-Scheme	Balanced Accuracy	AUC
GPC	Multimodal	-	-	0.1	LOSO	85.94%	0.90
GPC	Multimodal	-	-	1	‘’	85.16%	0.91
GPC	Multimodal	-	-	50	‘’	81.16%	0.92
GPC	Multimodal	-	-	100	‘’	85.16%	0.92
GPC	Multimodal	-	0.1	0.1	‘’	85.16%	0.92
GPC	Multimodal	-	1	0.1	‘’	85.94%	0.90
GPC	Multimodal	-	50	0.1	‘’	83.59%	0.89
GPC	Multimodal	-	100	0.1	‘’	83.59%	0.89
SVM	Multimodal	-	-	0.1	LOSO	68.75%	0.73
SVM	Multimodal	-	-	1	‘’	77.34%	0.86
SVM	Multimodal	-	-	50	‘’	83.59%	0.91
SVM	Multimodal	-	-	100	‘’	83.59%	0.91
SVM	Multimodal	-	0.1	50	‘’	83.59%	0.91
SVM	Multimodal	-	1	50	‘’	83.59%	0.91
SVM	Multimodal	-	50	50	‘’	77.34%	0.86
SVM	Multimodal	-	100	50	‘’	75.00%	0.81

When looking at the results of the scaling experiments, a first thing to notice is that the different values of scaling pose little change in performance when using the GPC, which wasn’t the case when analyzing the unimodal PET data based on Gaussian process classification (Table 6). Regardless, the best performing model out of these renditions sets the scaling of the PET data at 0.1 and no value for scaling the MRI data. The same balanced accuracy was found when using a scaling value of 1 for the MRI data, but since there is no difference, it is chosen to keep the MRI data as is. As for the support vector machines, the optimal combination of balanced accuracy and area under the curve had been found when scaling the PET data to a value of 50 and no scaling for the MRI data. Overall, the performances of these multimodal analyses are set at higher values than unimodal ones using the PET data, but lower than those using only the MRI data. These values are, however, not definitive since other parameters still need to be optimized. The next step in the process had been to determine the right combination of scaling values and atlases for the multiple kernel learning algorithm. Results of these experiments are posted in Table 12.

Table 12

Atlas- and scaling experiments for MKL using multimodal data

Machine	Modality	Atlas	Scaling MRI	Scaling PET	CV-Scheme	Balanced Accuracy	AUC
MKL	Multimodal	AAL	-	0.1	LOSO	46.88%	0.47
MKL	Multimodal	AAL	-	1	‘’	37.50%	0.40
MKL	Multimodal	AAL	-	50	‘’	89.06%	0.96
MKL	Multimodal	AAL	-	100	‘’	89.06%	0.96
MKL	Multimodal	AAL	0.1	50	‘’	89.06%	0.96
MKL	Multimodal	AAL	1	50	‘’	89.06%	0.96
MKL	Multimodal	AAL	50	50	‘’	89.06%	0.96
MKL	Multimodal	AAL	100	50	‘’	89.06%	0.96
MKL	Multimodal	Brodmann	-	0.1	LOSO	96.88%	1.00
MKL	Multimodal	Brodmann	-	1	‘’	96.88%	1.00
MKL	Multimodal	Brodmann	-	50	‘’	96.88%	1.00
MKL	Multimodal	Brodmann	-	100	‘’	96.88%	1.00
MKL	Multimodal	Brodmann	0.1	1	‘’	96.88%	1.00
MKL	Multimodal	Brodmann	1	1	‘’	96.88%	1.00
MKL	Multimodal	Brodmann	50	1	‘’	96.88%	1.00
MKL	Multimodal	Brodmann	100	1	‘’	96.88%	1.00
MKL	Multimodal	Hammers	-	0.1	LOSO	43.75%	0.43
MKL	Multimodal	Hammers	-	1	‘’	32.81%	0.34
MKL	Multimodal	Hammers	-	50	‘’	98.44%	0.97
MKL	Multimodal	Hammers	-	100	‘’	98.44%	0.97
MKL	Multimodal	Hammers	0.1	50	‘’	98.44%	0.97
MKL	Multimodal	Hammers	1	50	‘’	98.44%	0.97
MKL	Multimodal	Hammers	50	50	‘’	98.44%	0.97
MKL	Multimodal	Hammers	100	50	‘’	98.44%	0.97
MKL	Multimodal	AAL2	-	0.1	LOSO	51.56%	0.69
MKL	Multimodal	AAL2	-	1	‘’	51.56%	0.69
MKL	Multimodal	AAL2	-	50	‘’	98.44%	0.97
MKL	Multimodal	AAL2	-	100	‘’	98.44%	0.97
MKL	Multimodal	AAL2	0.1	50	‘’	98.44%	0.97
MKL	Multimodal	AAL2	1	50	‘’	98.44%	0.97
MKL	Multimodal	AAL2	50	50	‘’	98.44%	0.97
MKL	Multimodal	AAL2	100	50	‘’	98.44%	0.97

As can be seen in Table 12, the highest value of balanced accuracy that has been found is set at the value of 98.44%, which is higher than the values found when using a support vector machine or Gaussian process classifier (Table 11) and is a very good score overall. This value has been found when using either the Hammers-atlas or the AAL2-atlas. It is chosen to use the AAL2-atlas for further analyses to keep the chosen atlas consistent (the best performing unimodal analyses also incorporated this atlas, see Table 2 and Table 7). Interestingly, different values for scaling do not seem to impact the performance of the model when using the Brodmann-atlas which wasn't the case for the unimodal analyses (Table 2 and Table 7). After finding optimal values for scaling, different combinations of first- and second-level masks are investigated. Results of these different renditions are summarized in Table 13.

Table 13
Mask experiments for analyses using multimodal data

Machine	Modality	Atlas	Scaling MRI	Scaling PET	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	AUC
GPC	Multimodal	-	-	0.1	Mask A	-	LOSO	85.94%	0.90
GPC	Multimodal	-	-	0.1	Mask B	-	‘’	84.38%	0.92
GPC	Multimodal	-	-	0.1	Mask B	YES	‘’	82.03%	0.90
GPC	Multimodal	-	-	0.1	Mask A	YES	‘’	85.16%	0.89
SVM	Multimodal	-	-	50	Mask A	-	LOSO	83.59%	0.91
SVM	Multimodal	-	-	50	Mask B	-	‘’	83.59%	0.91
SVM	Multimodal	-	-	50	Mask B	YES	‘’	82.13%	0.87
SVM	Multimodal	-	-	50	Mask A	YES	‘’	83.59%	0.88
MKL	Multimodal	AAL2	-	50	Mask A	-	LOSO	98.44%	0.97
MKL	Multimodal	AAL2	-	50	Mask B	-	‘’	98.44%	0.97
MKL	Multimodal	AAL2	-	50	Mask B	YES	‘’	95.31%	0.97
MKL	Multimodal	AAL2	-	50	Mask A	YES	‘’	92.19%	0.97

Table 13 displays the results of the mask-optimization process of the multimodal analyses. In all three cases, Mask-A has been chosen for the final analyses. Also, across all renditions, the inclusion of a second-level mask seems to decrease the value of balanced accuracy. Interestingly, the variation of the different mask-combinations doesn't impact the performance of the different models as much as it does in the unimodal analyses. Especially in the unimodal analyses for the PET data, the mask seems to have a large impact (Table 8). The final multimodal analyses (including *P*-values for accuracies) are displayed in Table 14.

Table 14

Final outcomes of analyses of models using multimodal data

Machine	Modality	Atlas	Scaling MRI	Scaling PET	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	Class AC HC	Class AC ALC	AUC
GPC	Multimodal	-	-	0.1	Mask A	-	LOSO	85.94% *P* = .0033	96.88% *P* = .7176	75.00% *P* = .0033	0.90
SVM	Multimodal	-	-	50	Mask A	-	LOSO	83.59% *P* = .0033	85.94% *P* = .0066	81.25% *P* = .0033	0.91
MKL	Multimodal	AAL2	-	50	Mask A	-	LOSO	98.44% *P* = .0033	96.88% *P* = .0033	100% *P* = .0033	0.97

As mentioned before, the best performing algorithm (model) out of the three is the multiple kernel learning method. For the measures regarding accuracy of this model, the results seem to be very stable. *P*-values are set at the lowest possible point when using the predefined number of permutations (300). This *P*-value indicates that the values of accuracies set by the model are clearly more extreme (i.e., larger) than the distribution of plausible values found during the permutation process under the null-hypothesis. The *P*-values of the accuracies in the other models are also set at low values, except for the class accuracy of the healthy controls in the model using Gaussian process classification. Interestingly, the class accuracy of healthy controls has also been the most unstable of the accuracies in the unimodal analyses as well (Table 4 and Table 9).

In Table 15, the percentage of contribution of different regions of interest (ROI's) to the discriminative power of the MKL-model as listed in Table 14 are posted. The different areas within the brain that are selected as contributing to the power of the model are the same as have been found in the MKL-analysis using the MRI data. The corresponding percentages (rounded to two decimals) differ only slightly between the two models (Table 5). The weights as presented in this table do however, vary greatly from those found when using the PET data (Table 10). Images of the corresponding weight maps are found in Appendix I.

Table 15
Weights of ROI's in finalized MKL-models using multimodal data

Region of interest	Contribution to performance
Angular Gyrus Left	45.19%
Hippocampus Right	25.48%
Hippocampus Left	15.55%
Angular Gyrus Right	12.36%
Frontal Inferior Operculum Right	1.42%

Note. Percentages are rounded to two decimals.

3.4 *Concluding remarks*

Three main research questions had been set within this project. Those three questions are as follows:

- Are the different algorithms able to cope with the challenges of high dimensionality and the use of heterogeneous but complimentary data (multiple modalities)?
- To what degree do the performances of the different classification methods differ from each other?
- Does the use of heterogeneous but complimentary data (multiple modalities) lead to better results in classification compared to using a single modality, and if so, to what extent?

A table displaying the finalized results for each of the three classification problems and the three algorithms, is presented below and will serve as a basis for shedding light onto these questions.

Table 16

Final results for all classification problems

Machine	Modality	Atlas	Scaling MRI	Scaling PET	1-st Level Mask	2-nd Level Mask	CV-Scheme	Balanced Accuracy	Class AC HC	Class AC ALC	AUC
GPC	MRI	-	-	-	Mask A	-	LOSO	100% *P* = .0033	100% *P* = .3056	100% *P* = .0033	1.00
GPC	PET	-	-	50	Mask A	YES	‘’	76.56% *P* = .0066	90.63% *P* = .9336	62.50% *P* = .0033	0.79
GPC	Multimodal	-	-	0.1	Mask A	-	‘’	85.94% *P* = .0033	96.88% *P* = .7176	75.00% *P* = .0033	0.90
SVM	MRI	-	-	-	Mask A	-	LOSO	98.44% *P* = .0033	96.88% *P* = .0033	100% *P* = .0033	1.00
SVM	PET	-	-	50	Mask A	YES	‘’	79.69% *P* = .0033	90.63% *P* = .0399	68.75% *P* = .0033	0.84
SVM	Multimodal	-		50	Mask A	-	‘’	83.59% *P* = .0033	85.94% *P* = .0066	81.25% *P* = .0033	0.91
MKL	MRI	AAL2	50	-	Mask A	-	LOSO	98.44% *P* = .0033	96.88% *P* = .0033	100% *P* = .0033	0.97
MKL	PET	AAL2	-	50	Mask A	YES	‘’	78.13% *P* = .0033	93.75% *P* = .0166	62.50% *P* = .0266	0.84
MKL	Multimodal	AAL2		50	Mask A	-	‘’	98.44% *P* = .0033	96.88% *P* = .0033	100% *P* = .0033	0.97

With regard to the first research question, the chosen methods are well equipped in dealing with the challenges of high dimensional data and the inclusion of multiple modalities within a single model as high BA and AUC values are encountered for all three machines for both unimodal- and multimodal datasets. As for the second question, the choice of machine learning technique does impact the accuracy of a model. In the unimodal analyses, the performance of the three algorithms differs only slightly (when using the same dataset). There is, however, a measurable difference. In terms of the multimodal analyses, the difference in accuracy is far greater. The best performing model out of the three is the one based on multiple kernel learning (BA = 98.44%). The difference between the models based on GPC and SVM is much smaller with balanced accuracies of 85.94% and 83.59%, respectively. It was already hypothesized for the multiple kernel learning method to outperform the other algorithms in a situation of multimodality due to the fact it computes specialized kernels for each of those modalities. Based on these results, this seems to be the case.

Concerning the last research question, the use of multiple modalities (both MRI- and PET data) within a single analysis does not seem to increase the performance of the rendered models. It is, however, to be noted that (almost) perfect classifications has been found with a single modality only, leaving (almost) no room for the multimodal analysis to improve. Perhaps in a situation where the classification would prove to be 'more difficult', a larger effect of combining heterogeneous, but complimentary, data could be seen. Regardless, the best result of all has still been found in one of the unimodal analyses. Namely the classification using Gaussian process classification based on MRI data only.

With regard to the regions of interest contributing to the discriminative power of the models, the results/regions as selected by the model based on PET data vary quite heavily from results found in analyses using the other datasets (unimodal MRI or multimodal data). The regions as selected in the latter two classification problems are very similar. It could be that these regions and corresponding measures regarding contribution are the most accurate since the overall predictive power of the models are set at higher values than the one based on PET data. Regardless, it seems that the angular gyri plays a significant role in the discrimination between alcohol-dependent respondents and healthy controls.

4. Discussion

The aim of this project has been to make comparisons in the performance of three different machine learning algorithms (support vector machines, Gaussian process classification and multiple kernel learning) in their ability to make classifications between alcohol-dependent respondents and healthy controls (N=48) based on either unimodal- or multimodal neuroimaging data (magnetic resonance imaging and positron emission tomography). Also, the ability of the techniques to cope with the challenge of high dimensional data (as is the case when using neuroimaging data) has been assessed.

The three different algorithms have been used in three different classification problems. All models and corresponding statistics have been rendered in the 'Pattern Recognition for Neuroimaging Toolbox (PRoNTo)' (Schrouff et al., 2013). The first classification problem has been the discrimination between the two groups based on MRI data. The analyses based on this data pose very good performance measures and when Gaussian process classification had been used, a perfect balanced accuracy had been found. The support vector machine and multiple kernel learning algorithms had also shown an almost perfect performance in classification. It is interesting to see that the best performing algorithm out of the three has been the GPC as it was hypothesized that this algorithm would not perform as well as the others due to its tendency to underperform in situations of high dimensionality. In the second classification problem, the PET data has been used. Overall, models using the unimodal PET data where found being less powerful than ones using the MRI data (balanced accuracies between 75% and 80%). Out of the three machine learning techniques, the best performing one had been the support vector machine. The difference between the three final models was, however, not large. Lastly, the methods have been applied to a multimodal dataset combining both MRI and PET data. As had been hypothesized, the best performing algorithm when using multimodal data has been the multiple kernel learning technique (BA, after optimizing the different parameters, of 98.44%), followed, by some distance, by the support vector machine and the Gaussian process classifier (BA's around 85%).

In the end, the best classification performances have been found when using the unimodal MRI data. Performance measures when using multimodal data where, however, set at higher values in comparison to unimodal PET data. It is to be noted that this might be dependent on the type of research question and data at hand. It cannot be said that it will always be best to use unimodal MRI data in the classification of individual clinical status.

For each of the classification problems, weight maps have been computed on the basis of the best performing multiple kernel learning model. This has been done in order to get insight into the contribution of different regions of interest to the power of the model. Even though variation had occurred between the different models in the results of these computations, it seems that the angular gyri play a significant role in the discrimination between alcohol-dependent respondents and healthy controls.

These results indicate that the use of machine learning algorithms can be very useful and accurate in classifying individual clinical status, which is a very promising conclusion on its own. Also, machine learning is able to overcome challenges in classification such as the use of high-dimensional data. Different situations, however, are cause for different solutions. This project shows that the right choice of technique is dependent on, for instance, the fact if unimodal- or multimodal data is used. Also, the settings/optimization of parameters within the model can make a grand impact on the accuracy of an analysis.

This is, for instance, indicated by a previous analysis that had used the same PET data as has been used within this project. In this study, a support vector machine was used to make discriminations between the two groups (alcohol-dependent respondents and healthy controls). In this analysis, a balanced accuracy of 79.7% had been found (Devrome et al., n.d.). This is also the highest BA that has been rendered within this project as well, when using the PET data in a unimodal manner. The different variations of the models trained on this data, however, show that this accuracy could differ when parameter settings are changed or when another algorithm is chosen. Also, the inclusion of MRI data to this model has made a large improvement to the classification performance of the algorithm (SVM).

Even though the project at hand is a step in the right direction of filling in the gap of knowledge in the comparison of different techniques in the classification of individual clinical status, further research into the topic is still needed. First of all, the current study uses neuroimaging data to classify alcohol-dependent respondents from healthy controls. As has been said before, different situations cause for different solutions. It could very well be that with a different classification problem (e.g., discriminating between depressed respondents and healthy controls), a different algorithm would be the best fit. The same goes for the parameters and settings such as the choice for a mask and atlas. In order to get insight into these possibilities, the three techniques (and perhaps others) need to be compared using different datasets and hence, different classification problems. Also, the impact on performance of the algorithms of using different (combinations of) neuroimaging modalities should be assessed in future research.

The project at hand has used neuroimaging data from 16 alcohol-dependent respondents and 32 healthy controls (N=48). Ideally, future research would use data from a larger pool of respondents in order to draw conclusions with more confidence.

In the end, the results as found within this project are very promising. Even though research on the topic is still incomplete and further analyses need to be rendered, evidence has been gathered that the three different techniques are well equipped in the classification of individual clinical status based on neuroimaging data and overcoming challenges within this process. The choice for the best fitting technique seems to be dependent on the situation at hand which is why I would recommend researchers to try different options before selecting an algorithm for making classifications. This also holds true for values on parameters within the model and the selection of, for instance, a mask. Different options need to be weighed and assessed in order to receive the highest possible accuracy.

5. References

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., & Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage*, *112*, 232-243.

Costafreda, S. G., Fu, C. H., Picchioni, M., Toulopoulou, T., McDonald, C., Kravariti, E., Walshe, M., Prata, D., Murray, R. & McGuire, P. K. (2011). Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. *BMC psychiatry*, *11*(1), 1-10.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., & Resnick, S. M. (2008). Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of aging*, *29*(4), 514-523.

Ferreira, L. K., Rondina, J. M., Kubo, R., Ono, C. R., Leite, C. C., Smid, J., Bottino, C., Nitrini, R., Busatto, G. F., Duran, F.L. & Buchpiguel, C. A. (2018). Support vector machine-based classification of neuroimages in Alzheimer's disease: direct comparison of FDG-PET, rCBF-SPECT and MRI data acquired from the same individuals. *Brazilian Journal of Psychiatry*, *40*(2), 181-191.

Fernandes, O., Portugal, L. C. L., Rita de Cássia, S. A., Arruda-Sanchez, T., Volchan, E., Pereira, M. G., Mourão-Miranda, J. & Oliveira, L. (2020). How do you perceive threat? It's all in your pattern of brain activity. *Brain imaging and behavior*, *14*(6), 2251-2266.

Gallangher, J. (2017). Autism detectable in brain long before symptoms appear. *BBC news website. https://www.bbc.com/news/health-38955872*

Goodwin, G. M. (2012). Bipolar depression and treatment with antidepressants. *The British Journal of Psychiatry*, *200*(1), 5-6.

Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., Elison, J.T., Swanson, M. R., Zhu, H., Botteron, K. N., Collins, D. L., Contstantino, J. N., Dager, S. R., Estes, A. M., Evans, A. C., Fonov, V. S., Gerig, G., Kostopoulos, P., McKinstry, R. C., Pandey, J., Paterson, S., Pruett, J. R., Schultz, R. T., Shaw, D. W., Zwaigenbaum, L. & Piven, J. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, *542*(7641), 348-351.

Kuss, M., Rasmussen, C. E., & Herbrich, R. (2005). Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of machine learning research*, *6*(10).

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24

Leurquin-Sterk, G., Ceccarini, J., Crunelle, C. L., de Laat, B., Verbeek, J., Deman, S., Neels, H., Bormans, G., Peuskens, H. & Van Laere, K. (2018). Lower limbic metabotropic glutamate receptor 5 availability in alcohol dependence. *Journal of Nuclear Medicine*, *59*(4), 682-690.

Long, Z., Jing, B., Yan, H., Dong, J., Liu, H., Mo, X., Han, Y. & Li, H. (2016). A support vector machine-based method to identify mild cognitive impairment with multi-level characteristics of magnetic resonance imaging. *Neuroscience*, *331*, 169-176.

Nicholson, A. A., Densmore, M., McKinnon, M. C., Neufeld, R. W., Frewen, P. A., Théberge, J., Jetly, R., Richardson, J. D. & Lanius, R. A. (2019). Machine learning multivariate pattern analysis predicts classification of posttraumatic stress disorder and its dissociative subtype: a multimodal neuroimaging approach. *Psychological medicine*, *49*(12), 2049-2059.

Portugal, L. C., Schrouff, J., Stiffler, R., Bertocci, M., Bebko, G., Chase, H., Lockovitch, J., Aslam, H., Graur, S., Greenberg, T., Pereira, M., Oliveira, L., Phillips, M. & Mourão-Miranda, J. (2019). Predicting anxiety from wholebrain activity patterns to emotional faces in young adults: a machine learning approach. *NeuroImage: Clinical*, *23*, 101813.

Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. *Journal of Machine Learning Research*, *9*, 2491-2521.

Ranlund, S., Rosa, M. J., de Jong, S., Cole, J. H., Kyriakopoulos, M., Fu, C. H., Mehta, M.A. & Dima, D. (2018). Associations between polygenic risk scores for four psychiatric illnesses and brain structure using multivariate pattern recognition. *NeuroImage: Clinical*, *20*, 1026-1036.

Resnick, S. M., Goldszal, A. F., Davatzikos, C., Golski, S., Kraut, M. A., Metter, E. J., Bryan, N. & Zonderman, A. B. (2000). One-year age changes in MRI brain volumes in older adults. *Cerebral cortex*, *10*(5), 464-472.

Sample, I. (2017). Brain scans could identify babies most at risk of developing autism, study shows. *The Guardian news website.* https://www.theguardian.com/society/2017/feb/15/brain-scans-could-identify-babies-most-at-risk-of-developing-autism-study-shows

Schouten, T. M., Koini, M., De Vos, F., Seiler, S., Van Der Grond, J., Lechner, A., Hafkemeijer, A., Möller, C., Schmidt, R., de Rooij, M. & Rombouts, S. A. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage: Clinical*, *11*, 46-51.

Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, Phillips C, Richiardi J & Mourao-Miranda J. (2013). *PR*oNTo: Pattern Recognition for Neuroimaging Toolbox**.** *Neuroinformatics*, 1, 319–337.

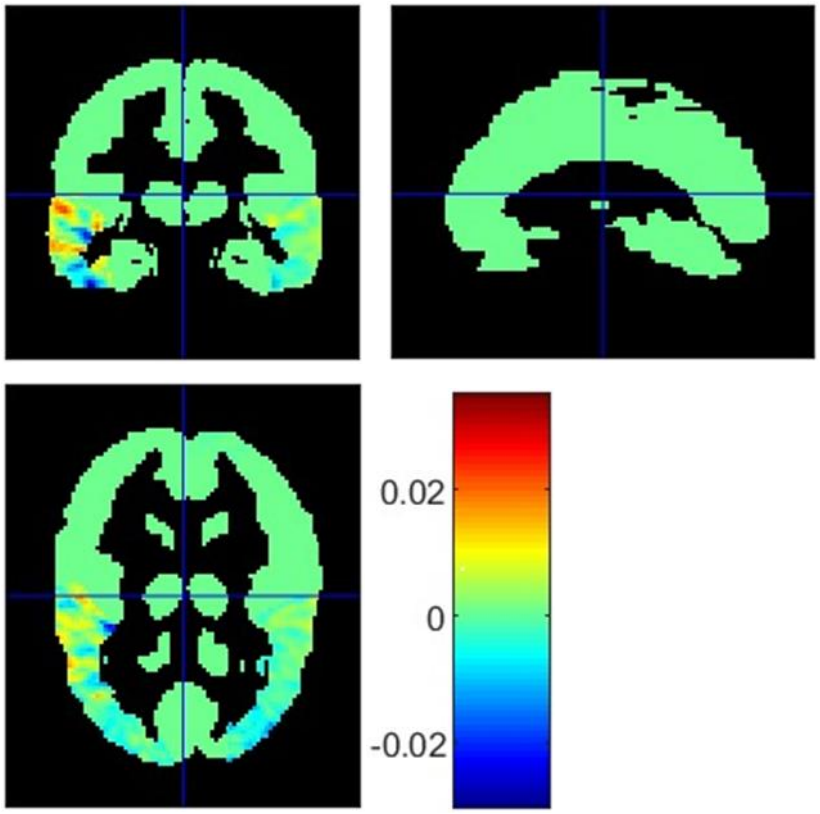
Vai, B., Parenti, L., Bollettini, I., Cara, C., Verga, C., Melloni, E., Mazza, E., Poletti, S., Colombo, C. & Benedetti, F. (2020). Predicting differential diagnosis between bipolar and unipolar depression with multiple kernel learning on multimodal structural neuroimaging. *European Neuropsychopharmacology*, *34*, 28-38.

Youssofzadeh, V., McGuinness, B., Maguire, L. P., & Wong-Lin, K. (2017). Multi-kernel learning with dartel improves combined MRI-PET classification of Alzheimer's disease in AIBL data: group and individual analyses. *Frontiers in human neuroscience*, *11*, 380.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, *55*(3), 856-867.

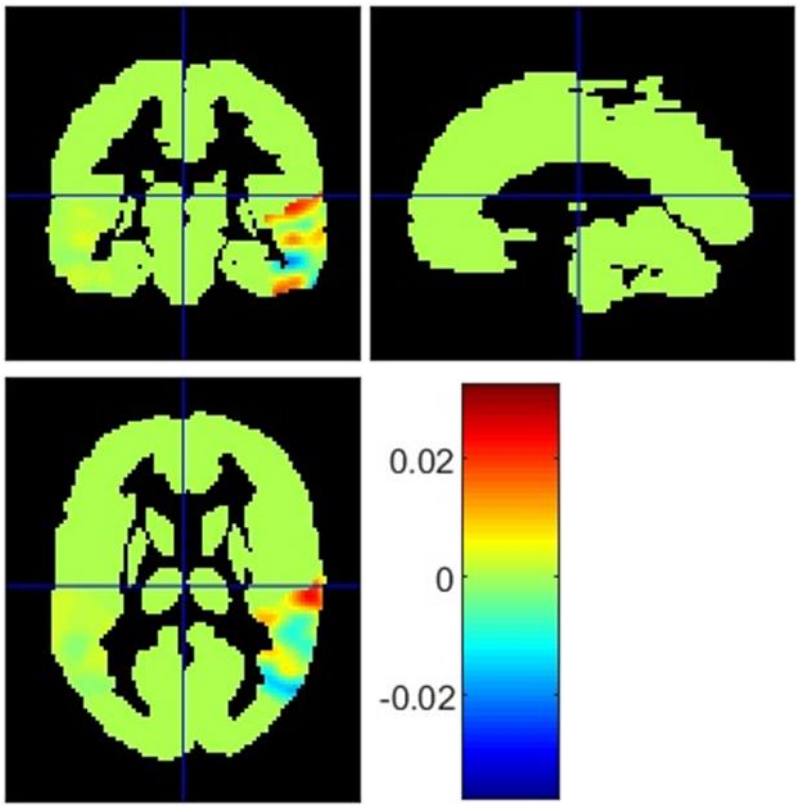
4. Appendix I

Figure 1.
Weight map of the finalized MKL-model using MRI data



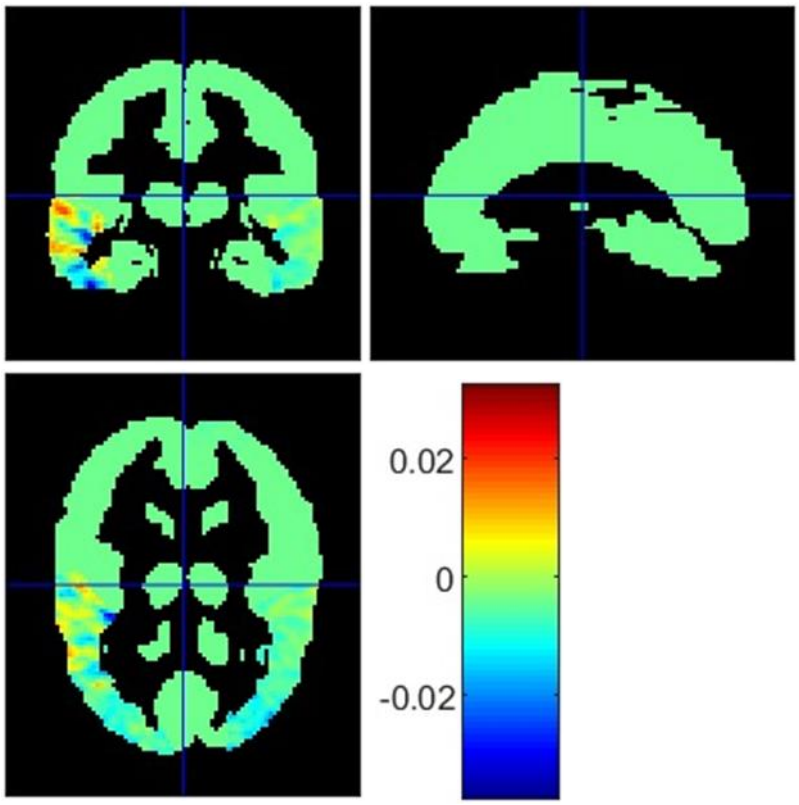
Note. Images represent T-values of the different regions

Figure 2.
Weight map of the finalized MKL-model using PET data



Note. Images represent T-values of the different regions

Figure 3.
Weight map of the finalized MKL-model using multimodal data



Note. Images represent T-values of the different regions