Universidade de Lisboa
Faculdade de Ciências
Departamento de Física

Developing new machine learning methods more robust to inter-subject and inter-scanner variability in structural MRI

Ana Isabel Dos Santos Silva

Dissertação de
Mestrado Integrado em Engenharia Biomédica e Biofísica
Perfil em Sinais e Imagens Médicas

Orientador :Professor Alexandre Andrade, Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências da Universidade de Lisboa

Coorientador: Dr. Orla Doyle, Institute of Psychiatry, Psychology & Neurscience, King's College London

2015

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor in King's College London, Dr. Orla Doyle, for much advice, encouragement and immense knowledge. It was wonderful working with someone so inspirational and enthusiastic. Her guidance helped me in all the steps of the project and writing of the thesis. I would also like to thank my supervisor in Faculty of Science of University of Lisbon, professor Alexandre Andrade for providing much encouragement, support, feedback and advice and for letting me know about this project.

I would like to mention Dr. Maria João, in particularly, for guiding me in the beginning of the project and for being extremely receptive, supportive and helpful. I would like to thank Ahmed Abdulkadir, whose paper was as inspiration early on and for being receptive and helpful through several e-mail conversations.

A special thanks to my office mate Vasileia Kotoula, for all her support and friendship and for making me laugh in our walks to the coffee shop. Thanks to Richard Joules, who helped me several times in understanding the basis of Gaussian Process.

I would like to demonstrate my gratitude to all my friends that, even in a different country, were always there to support me and always made me laugh in our Skype conversations. Particularly, to my friends Eduardo Terças and Bernardo Emídio, for visiting me while I was in London and for being there whenever I needed a friend. A very special thanks to my dearest friend Joana Pereira, for being such a good friend and for all the support since the very beginning of our adventures outside Portugal and during our academic life.

Finally, I would like to express my deepest gratitude to my family for all their caring support and love. To my mother for having me hosted at her home in London, for all the family dinners and for keeping me calm in stressful situations. To my father for visiting me in London and for always being there for me. To my brother, for endless hours of stimulating discussions and for his friendship.

Abstract

Neuroimaging is a field that includes a wide range of brain-mapping techniques. Currently, massive univariate approach using Voxel-Based Morphometry (VBM) is a well-established method in neuroimaging data analysis. However, it may be insufficient to recognize multivariate relationships and obtain realistic analyses. This leads to the development of multivariate image analysis methods.

Machine learning can be seen as an alternative to inferential multivariate analyses. Classifiers were proved to be sufficiently sensitive to separate patients with a brain disease from cognitively normal subjects. To date, most machine learning based studies of neuroimaging data comprise images from a single imaging site. If imaging-based predictive models are to be adopted in the clinical setting, these models need to be robust to multi-site variation, typical of real-world datasets, which are acquired using different scanners and acquisition parameters.

Combining data from different scanners provides larger sample sizes, which leads to more robust classification results. However, this scanner-variability leads to systematic differences between images from different scanners introducing variance that is unrelated to disease. Between scanner bias affects both main stages of automated disease classification, the training phase and the testing phase. It is well known that other disease-unrelated factors such as age, sex and total intracranial volume would also influence the measured tissue properties.

As such, this thesis focuses on studying methods for correcting scanner and subject variability in machine learning based analyses of structural MRI data. The main goal of this project is to achieve better predictive models that are more robust to multi-site variation in data. For this purpose, three different datasets were used. Two datasets comprising only healthy subjects scanned twice, with different scanner conditions, were used in order to illustrate the scanner variability effect. Data from the ADNI project database were used to analyze the impact of scanner and subject variability in automated classification.

The ADNI project is a multisite study and 413 subjects (145 healthy, 148 MCI and 120 AD) from 6 scanners were selected for the VBM analysis, but only data from 4 scanners (1.5GE, 1.5SIEMENS, 1.5Philips and 3SIEMENS) were considered for the machine learning analyses. The diagnostic conditions were equally distributed, with approximately 30 scans per diagnostic group and scanner. First, a univariate VBM analysis was performed to illustrate the extent of scanner variability. In a multivariate way, a Gaussian process classification algorithm was used to study the extent and impact of scanner and subject variability. Then, two regression methods, Gaussian process regression (GPR) and Ridge ordinary least squares regression (ROLS) were applied to study their consistency to remove out the confound effect.

Results show that VBM analysis largely confirmed the presence of inter-scanner differences at group level for the three datasets. However, for the ADNI dataset, the effect of disease was spatially wider spread than effects of scanner and no significant interactions between scanner and disease were found. For the two first datasets, classification of scanner was possible with very high accuracy and area under the curve performances. For the ADNI dataset, classification of scanner was possible for all the combinations of scanners except 1.5SIEMENS and 3SIEMENS, where the classification performance was not significant. This suggests that

differences between data from scanners of different manufacturers are larger than between data from the same manufacturer.

With the aim of studying the impact of scanner and subject variability in automated classification, the four clinical scenarios formulated by Ahmed Abdukladir *et al*. in his study, were studied. For each scenario, the two correction methods were applied and permutation and McNemar tests were used to infer if there were any significant improvements after correction. The first scenario consisted in using controls and respective patients groups separately for each four sites. Here, MCI classification accuracies show a poor classification result when using 1.5 GE and 1.5 Philips scanners. The results from the second scenario show that a model trained with data from a single scanner generalized well to data from a different scanner. When confounding diagnostics groups and scanner during training, e.g, by using controls from one scanner and MCI from another, accuracy dropped significantly in many cases. By regressing out confounds with GPR and ROLS, performance levels were comparable to those obtained in scenarios without confound, except when using MCI subjects from 1.5 GE and 1.5 Philips scanners.

Despite the limitations due to the MCI classification performance of scanners 1.5 GE and 1.5 Philips and differences in gender when comparing diagnostic groups from 1.5 Philips to other scanners, results from this work are in agreement with Ahmed Abdulkadir *et al*. findings. Future work will focus in translating these findings to other neurodegenerative diseases and psychiatric disorders, such as schizophrenia.

Keywords: MRI, Gaussian Process, Between-scanner variability, Classification, Regression

Resumo

Neuroimagem é uma vasta área que inclui uma ampla gama de técnicas de mapeamento cerebral. A sofisticação dos métodos de neuroimagem nas últimas décadas possibilitou o avanço dos conhecimentos sobre alterações nas estruturas cerebrais, possibilitando o estudo de doenças neurodegenerativas como a doença de Alzheimer. Por exemplo, a técnica de Imagem de Ressonância Magnética (IRM) possibilita uma análise a nível estrutural e funcional, não invasiva, do cérebro. Não precisa de radiação ionizante, permitindo estudos longitudinais. Técnicas tradicionais de análise de scans de IRM consistem numa avaliação visual por parte de radiologistas experientes. Esta análise, para além de subjetiva, torna-se impraticável quando o número de sujeitos aumenta e as mudanças estruturais são mais subtis.

Nos últimos anos, têm sido feitos avanços no desenvolvimento de técnicas automáticas para identificar diferenças na anatomia do cérebro entre grupos de sujeitos. Atualmente, a estratégia mais utilizada para a análise de dados de neuroimagem é a análise utilizando o método VBM (Voxel-based Morphometry) e consiste numa abordagem univariada em massa usando métodos de mapeamento estatístico paramétricos para investigar diferenças na anatomia do cérebro. Contudo, devido à complexidade do cérebro e suas estruturas, esta abordagem pode não ser suficiente para se obter uma análise realista. A análise univariada poderá apresentar uma fraca sensibilidade quando o efeito de interesse tem uma distribuição multidimensional na ativação. Surge assim a necessidade do uso de métodos multivariados.

Machine learning consiste num método de aprendizagem automática para estudar padrões de atividade. O objetivo, numa aprendizagem supervisionada, é treinar o algoritmo com exemplos de treino e, depois de treinado, o algoritmo deve ser capaz de prever o resultado para novos exemplos. Estes métodos podem ser chamados regressão, se a variável for contínua ou classificação, se a variável for discreta. Vários estudos demonstraram que classificadores conseguiram ser suficientemente sensíveis para separar pacientes com uma doença cerebral de sujeitos cognitivamente normais. Um exemplo de algoritmos de machine learning são os algoritmos baseados em processos Gaussianos (Gaussian processes, GP) que permitem procedimentos não paramétricos Bayesianos para a classificação.

Até à data, a maioria dos estudos baseados em algoritmos de machine learning em neuroimagem analisam imagens de um único scanner de IRM. Contudo, para podermos adotar modelos preditivos num cenário clínico, estes modelos precisam de ser robustos a variações de scanners, típicas de conjuntos de dados reais, que são normalmente adquiridos utilizando diferentes scanners e parâmetros de aquisição. A combinação de dados de diferentes scanners origina amostras maiores, o que leva a um resultado de classificação mais robusto. Contudo, esta variabilidade de scanner leva a diferenças sistemáticas entre imagens de diferentes scanners que pode introduzir variância que não está relacionada com a variável de interesse. A variabilidade entre scanners afeta os dois estados da classificação, a fase de treino e a fase de teste. Outros factores que não estão relacionados com a doença, como a idade o género e o volume intracranial total podem influenciar os resultados da classificação.

O trabalho desenvolvido nesta tese foca-se no estudo de métodos para corrigir o efeito causado pela variabilidade entre scanners e pelas diferenças específicas entre sujeitos (idade, género, volume intracranial total) na análise de IRM estrutural em machine learning. O objetivo principal do projeto é alcançar melhores modelos preditivos mais robustos a variações de scanner. Para esse propósito três diferentes conjuntos de dados foram usados. Dois deles apenas continham sujeitos saudáveis com dois scans cada sujeito. Os dois scans foram adquiridos no mesmo scanner mas, no primeiro conjunto de dados com diferentes bobines e no segundo com diferentes parâmetros de scanner. Estes conjuntos de dados foram utilizados para ilustrar o efeito de variabilidade entre scanners, sem ter em conta diferenças entre sujeitos. Foram ainda selecionados dados da base de dados do projeto ADNI de forma a analisar o impacto da variabilidade entre scanners e sujeitos na classificação.

O projeto ADNI é um estudo que envolve vários scanners e 413 sujeitos (145 saudáveis, 148 com comprometimento cognitivo leve (CCL) e 120 com Alzheimer) adquiridos em 6 scanners diferentes foram selecionados para a análise VBM. Apenas dados de 4 scanners (1.5 GE, 1.5 SIEMENS, 1.5 Philips e 3 SIEMENS) foram considerados para a análise com machine learning. Os grupos de diagnóstico foram distribuídos igualmente, com aproximadamente 30 scans por grupo de diagnóstico e scanner. Em primeiro lugar, foi feita uma análise de VBM para ilustrar a extensão da variabilidade entre scanners. Depois, foi feita uma análise de classificação utilizando o processo Gaussian para demonstrar a extensão e o impacto da variabilidade entre scanners e entre sujeitos numa análise multivariada. Dois métodos de correção foram estudados, o processo de regressão Gaussiano (PRG) e o método de regressão dos mínimos quadrados ordinários regularizado (MMQOR).

Os resultados mostram que a análise de VBM confirma a presença de diferenças entre scanners para os primeiros dois conjuntos de dados, estando o efeito disperso por toda a matéria cinzenta. Para o conjunto de dados ADNI, o efeito da variabilidade entre scanners não estava disperso pela matéria cinzenta, mostrando fortes ativações localizadas. Contudo, o efeito causado pela doença de Alzheimer era especialmente mais disperso e não foram encontradas interações significantivas entre o efeito da doença e o efeito causado pelos scanners. Para os primeiros dois conjuntos de dados, a classificação de scanners foi possível mostrando uma acurácia e área sob a curva altas. Para o conjunto de dados ADNI, a classificação entre scanners foi também possível com uma alto desempenho para todas as combinações de scanners excepto quando combinando os scanners 1.5 SIEMENS com 3 SIEMENES, onde a desempenho da classificação não foi significativa. Isto sugere que diferenças entre scanners de diferentes fabricantes são maiores do que entre scanners do mesmo fabricante.

Com o objetivo de estudar o impacto da variabilidade entre scanners e entre sujeitos na classificação, quatro cenários clínicos formulados anteriormente no estudo de Ahmed Abdulkadir *et al*. foram estudados. O objectivo era investigar se, usando um estudo altamente controlado e normalizado como o da base de dados ADNI, os resultados encontrados no estudo de Abdulkadir eram replicados. Para cada cenário, os dois métodos de correção foram aplicados e testes de permutações e o teste de McNemar foram usados para inferir se a correção levava a melhorias significativas na classificação. O primeiro cenário consistiu em usar controlos e respetivos grupos de pacientes separadamente para cada um dos quatro cenários. Aqui, a classificação dos sujeitos com CCL adquiridos nos scanners 1.5 GE e 1.5 Philips foi pouco eficaz e não foi significativa. No segundo cenário, o modelo foi treinado com dados de um scanner

e testado com dados de um segundo scanner. Os resultados mostram que um modelo treinado com dados de um scanner generaliza bem para dados adquiridos num scanner diferente. Quando se misturam grupos de diagnóstico com scanners durante o treino, p. ex., usando controlos de um scanner e CCL de outro scanner, a acurácia desceu significativamente em vários casos. Depois da correção com os métodos PRG e MMQOR, o desempenho aumentou em vários casos tornando-se comparável aos desempenhos obtidos nos outros cenários, excepto quando utilizando dados dos sujeitos CCL adquiridos nos scanners 1.5 GE e 1.5 Philips.

Apesar das limitações devido ao desempenho da classificação dos sujeitos CCL destes dois scanners e diferenças no género quando comparando grupos de diagnóstico adquiridos no scanner 1.5 Philips com os outros scanners, os resultados deste trabalho estão de acordo com os resultados do estudo de Abdulkadir *et al*. Trabalho futuro consistirá em transpor estes resultados para outras doenças degenerativas e desordens psiquiátricas como o caso da esquizofrenia.

Palavras-chave: IRM, Processo Gaussiano, Variabilidade entre scanners, Classificação, Regressão

Symbols and Abreviations

Symbols

M_g	Net Magnetization
M_{gz}	Longitudinal Magnetic vector
M_{gxy}	Transversal Magnetic vector
B_0	Magnetic Field
T	Tesla
t, F	t and F statistics
ϵ	Residual errors matrix/vector
$\hat{\beta}$	Parameters estimates matrix/vector
y, Y	Observations vector, matrix
$\boldsymbol{x_d}, X_d$	Design vector, matrix
$\hat{Y}$	Predicted observations
X^+	Moore-Penrose pseudo inverse of X
h	Regularisation parameter
$\hat{\sigma}^2$	Residual variance estimates
c	Contrast vector/matrix
μ	Mean vector
K	Covariance matrix
σ_n^2	Variance of the noise
δ	Kronecker delta
θ	Hyperparameters vector/matrix
Φ(f)	Sigmoid function
χ^2	Qui-squared value
$\widehat{\boldsymbol{w}}$	Weight vector
x, X	GM concentrations vector, matrix
Q	Kernel function
ρ	p-value

Abreviations

AC	Anterior Commissure
Acc	Accuracy
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AUC	Area Under the Curve
CSF	Cerebrospinal Fluid
DARTEL	Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra
EP	Expectation Propagation
FID	Free Induction Decay
FN	False Negatives
FP	False Positives
FWHM	Full Width at Half-Maximum
GLM	General Linear Model
GM	Gray Matter
GP	Gaussian Process
GPC	Gaussian Process Classification
GPR	Gaussian Process Regression
MCI	Mild Cognitive Impairment
MNI	Montreal Neurological Institute
MP-RAGE	Magnetization-Prepared Rapid Gradient-Echo
MRI	Magnetic Resonance Imaging
RF	Radiofrequency
ROC	Receiver Operating Characteristic
ROLS	Ridge Ordinary Least Squares
RSS	Residual Sum of Squares
Se	Sensitivity
Sp	Specificity
SPM	Statistical Parametric Mapping
TE	Echo Time
TIV	Total Intracranial Volume
TN	True Negatives
TP	True Positives

TR	Repetition Time
TPM	Tissue Probability Maps
VBM	Voxel-Based Morphometry
WM	White Matter

List of Figures

CONTENTS

Chapter 1: Introduction

Neuroimaging is a vast field that covers a wide range of brain-mapping techniques with specific information about the brain. For instance, Magnetic Resonance Imaging (MRI) is used for structural analysis, functional MRI for functional analysis, and Positron Emission Tomography (PET) for metabolic and neurochemical analysis [1].

MRI offers a unique opportunity to non-invasively access the morphology of the human brain with high resolution. The anatomical complexity, however, makes purely visual interpretation of the obtained data challenging. In order to achieve an objective and operator-independent comparison, computational methods have been developed. One of the most commonly used is voxel-based morphometry (VBM) [2][3]. This technique involves tissue segmentation, spatial normalization and smoothing procedures, followed by voxel-by-voxel univariate statistical tests on feature changes across subjects [4].

Methods belonging to this family of mass-univariate methods have been fundamental tools in modern neuroimaging, by aiding in the detection of group differences structurally and in understanding of spatial patterns of functional activation. These methods are suitable for standard statistical inference techniques by associating a statistical significant measure, a p-value, with every voxel allowing an easy interpretation of the output [4]. However, recently the neuroimaging community has recognized that the presence of multivariate relationships between different brain regions may not be entirely explained by univariate analysis. This has lead to the development of multivariate image analysis methods [3].

In the past few years, interest in using machine learning classifiers and pattern recognition algorithms for analysing MRI data has been growing. They provide means to decode and characterize dependent and independent brain activity/structure, distinguishing it from non-informative brain signals [5]. Several studies have shown that automated analysis of structural magnetic resonance images is a promising way to improve early detection of neurodegenerative brain diseases [6][7][8]. Classifiers are sufficiently sensitive to separate patients with a brain disease from cognitively normal subjects, even if the patient diagnosed shows no symptoms [8]. The introduction of kernel machines, such as Support Vector Machines (SVM) and Gaussian Processes (GP) has opened the possibility of flexible models for classification and regression, which are practical to work with [9].

Gaussian processes, as they are applied in machine learning, provide a promising way of doing non-parametric Bayesian modelling in a supervised learning problem. They combine a great flexibility by working in high dimensional features spaces with the simplicity that all operations are performed in a lower dimensional input space using positive definite kernels [10]. This thesis addresses the mathematical basis behind Gaussian processes applied to classification and regression, once they were used during the study.

One of the many benefits of these methods is that their clinical application is independent from human experts and thereby can be applied outside specialized centres. However, if these imaging-based predictive models are to be adopted in the clinical setting, they need to be robust to multi-site variation, including different scanners and acquisition parameters.

Studies that use longitudinal and/or multi-site datasets have the potential to provide a wealth of information. Therefore, the large number of subjects resulting from pooling multi-scanner data-sets has several advantages such as: (1) improved sensitivity, allowing detection of subtle effects, (2) increased reliability and (3) higher confidence about the size of effect by averaging out unexpected confounds. However, two scanners, even from the same manufacturer and type cannot produce the exact same images [8].

Differences in hardware and/or acquisition sequenced may lead to systematic differences between images from different scanners. Even in longitudinal studies performed in one scanner over a period of time, it is very demanding to keep every scanner parameter constant during that period and exchanging of hardware components or upgrades of software is almost inevitable [3].

To some degree, pre-processing such as intensity normalization or tissue segmentation algorithms can remove scanner differences but systematic differences remain [11]. Between scanner bias is relevant in both main stages (training and testing phase) of automated disease classification. At the training phase, the classification algorithm learns a function that differentiates diagnostic groups. Here, between-scanner differences introduce variance that is unrelated to the disease and thus degrade the quality of the discriminative function. At the testing phase, scanner bias may favors the prediction of one of the classes [8].

Despite differences between scanners, other confounds can interfere with classifiers accuracy, such as age, gender and total intracranial volume (TIV). The TIV can be defined as the volume within the cranium, including the brain, meninges and cerebrospinal fluid (CSF). When comparing scans from different subjects we must be aware that bigger brains will have bigger grey matter (GM) and white matter (WM) volumes, which could confound comparisons. Measuring TIV allows whole-brain and regional volumetric measures to be normalized for head size [12].

1.1 Goals

This thesis addresses primarily the effect of scanner variability and subject-specific confounds in machine learning-base analysis of structural MRI data. The main goal of the project is to design better predictive models that are more robust to multi-site variation in the data, in order to improve classification results.

Hereupon, the project was divided as follows:

1. Illustrating and analysing the extent of scanner variability effect and other confounds (age, gender and TIV) by using voxel-based morphometry and machine learning based on Gaussian process classification method. Two different datasets comprising only healthy subjects scanned twice in different scanner conditions were used in order to illustrate the scanner variability effect and data from the ADNI project database was used to analyze the impact of scanner and subject variability in automated classification.
2. Two different methods for correcting data were used to study their reliability to remove the confound effect. The two regression methods were: Gaussian process regression and Ridge ordinary least squares regression.

1.2 State of the Art

Multi-centric setups, their implications in the context of systematic differences in scanner hardware, field strength and how they affect classification were studied in several investigations.

In 2008, a study conducted by Stonnington *et al*. [11], used a dataset including a total of 6 scanners and 136 subjects scanned over ten years. All the scans were done in the same platform, which underwent upgrades over time and there were minor variations the repetition time (TR), time echo (TE) and flip angles. The authors used Voxel-based morphometry (VBM) and found that scanner-related effects are substantially smaller than those caused by moderately progressed Alzheimer's disease. However, extents of univariate statistical differences on group level (as in VBM analysis) cannot be directly translated to discriminability of individual using multivariate methods (as in SVM and GP classification) [8]. On the other hand, several studies indicate that the effects of inter-scanner variability are far greater than intra-scanner variability[13][14].

In a previous study of Abdulkadir *et al*., 2011 [15], a total of 518 MRI sessions from 226 healthy controls and 191 individuals with probable Alzheimer's disease from the multicentre Alzheimer's Disease Neuroimaging Initiative (ADNI) were used. Using pooled data from this large-scale, highly standardized and controlled multi-site study, the authors did not find significant decline due to different field strengths. However, the highly standardized data acquisition procedure of the ADNI study is not representative for clinical scenarios.

In real-world clinical scenarios, data from a single specialized centre may have to be used to diagnose patients in other clinics using scan sequences that differ in several parameters. Classification in such a clinical setting could perform poorly and, in most cases, adaption of the protocol to a reference protocol is impractical.

Recently, Kostro and Abdulkadir *et al*. 2014 [8] carried out a study including patients with pre-manifested Huntington's disease (preHD), manifest Huntington's disease (mHD) and healthy controls. preHD patients were diagnosed with Huntington's disease but were not showing any symptoms. These patients had more subtle structural changes when compared with mHD patients. The structural magnetic resonance brain images were acquired at four sites. The authors formulated clinically relevant scenarios, taking into account inter-scanner differences and evaluated how well the classification SVM algorithm predicted the disease status. In addiction, they introduced a method based on Gaussian process regression (GPR) to correct the input data for confounding effects such as scanner, age, sex and total intracranial volume (TIV) and evaluate changes in performance after correction.

They concluded that a model trained with data from a single scanner generalized well to data from a different scanner. However, when confounding diagnostic groups and scanner during training classification accuracy dropped significantly, especially in preHD patients. When regressing out confounds with GPR, the overall performance increased and it was comparable to those obtained in scenarios without confound.

In the same study, the authors also looked at ordinary least squares regression for correcting confounds. This method was applied by Dukar *et al*. (2011) [16] for

correcting age in dementia, having substantially improved univariate detection (using VBM) of gray matter atrophy in AD patients. However, in Kostro and Abdulkadir et al. study, they claimed that, althought being a simple and fast method to compute, it does tend to overfit data. The authors present GPR as a method that overcomes some limitations of linear regression.

Concluding, this study suggests that bias due to scanner differences may be an issue and interfere with classification performance. Care should be taken when analysing data from different scanners and a careful analysis should be done to make sure that the performance does not drop due to transfer of the model. If classification accuracy decreases substantially, a procedure to correct the data must be done.

Chapter 2: Structural brain imaging in Alzheimer's disease

2.1 Alzheimer's disease and Mild Cognitive Impairment

Alzheimer's disease (AD) is the most common form of dementia. Clinical and neuropathological studies have greatly advanced in elucidating the underlying causes of AD. This disease is associated with the progressive accumulation of abnormal proteins (amyloid-β [Aβ] and hyperphosphorylated tau) in the brain, which leads to progressive synaptic, neuronal and axonal damage. Neurobiological changes normally occur years before symptoms appear, with a stereotypical pattern of early medial temporal lobe involvement, followed by progressive neocortical damage. Therefore, this delay suggests that the toxic effect of tau and/or Aβ progressively deteriorates the brain tissue, eroding the brain and cognitive reserve until a clinical threshold is surpassed and amnestic symptoms develop [17][18].

Mild Cognitive Impairment (MCI) consists in an intermediate stage between the expected cognitive decline of normal aging and the more serious decline of dementia. Amnestic MCI involves a memory disturbance in the absence of dementia and it is followed by widespread cognitive deficits in multiple domains. MCI may involve problems with memory, language, thinking and judgment that are greater than normal age related changes [19]. When this condition reaches a disability threshold, traditional diagnostic criteria for probable AD are fulfilled. The possibility of drugs that can slow or prevent disease progression has fortified increased interest in identifying individuals with AD more accurately and in early stages [18].

While many progresses have been made, studying the disorder *in vivo* faces several barriers. The structures affected by AD lie deep within the brain where biopsy is not practical and animal models do not develop AD naturally. In contrast to other techniques, Magnetic Resonance Imaging (MRI) offers a non-invasive method for analysing structural and functional brain characteristics. It does not need ionizing radiation, allowing longitudinal studies without significant health concerns [17].

2.1.1 Alzheimer's Disease Neuroimaging Initiative (ADNI)

Several MRI studies have been performed to understand the underlying pathology in patient populations already diagnosed with AD or MCI. Moreover, there are several databases, available online, with multimodal data collected during several studies. For instance, the Alzheimer's Disease Neuroimaging Initiative (ADNI) is an ongoing, longitudinal, multicentre study, designed to develop clinical, imaging, genetic and biochemical biomarkers for the early detection and tracking of Alzheimer's disease. The initiative aimed to enrol 400 subjects with early MCI, 200 subjects with early AD, and 200 normal control subjects [20].

Multiple research groups contribute their findings to the understanding of the progression of Alzheimer's disease in the human brain. Therefore, the MRI data were collected in different research centres, using different scanners, throughout the ADNI

project. Some data were acquired using 1.5T scanners or more modern and expanded 3T scanners, with different protocols. Moreover, scanners from different vendors were used. The ADNI project website provides the MRI scanner protocols from three different models (General Electric (GE) Healthcare, Philips Medical Systems, and Siemens Medical Solutions) [20].

In the interest of promoting consistency and rigor in analysis and meaningful comparisons of different algorithms, the AD MRI Core has created standardized analysis sets of data comprising scans that met minimum quality control requirements.

2.2 Magnetic Resonance Imaging

2.2.1 MRI signal formation

Magnetic Resonance Imaging (MRI) is a diagnostic imaging tool, that uses a combination of strong magnets and low-energy radiofrequency signals to gather information from certain atomic nuclei within the body. Therefore, this technique is based on the behaviour of a system of protons in the presence of a magnetic field. Magnetic strength is generally reported in units of Tesla (T), and MRI scanners have very high field strengths: normally 1.5T and 3T scanners predominate in the clinical setting [21].

In MRI, a particular type of nuclei is selected and its distribution in the body is monitored. Hydrogen is the most commonly imaged element in clinical imaging, not only due to its abundance in the body, but also because it gives the strongest MRI signal [22]. Therefore, clinical MRI is focused on these protons in both water and macromolecules, such as proteins and fat.

Quantum mechanics describes the behaviour of the individual protons. A classical mechanics model, on the other hand, is used to describe the changes for a large number of protons, where their quantum mechanical behaviour is averaged out [23]. Quantum mechanics describe that spin is an intrinsic form of angular momentum carried by the atomic nucleus, and all the protons spin around their own axes. This spinning protons act like tiny current loops and consequently generate, along the spin axis, a magnetic field. Under normal circumstances, these tiny magnets are randomly distributed in space, the magnetic moments cancel each other out, and thus the net magnetic vector (**Mg**) is zero [21][22].

However, if such a group of spinning protons is submitted to a strong external magnetic field, B_0, they preferentially align themselves either parallel or anti-parallel to the field (Figure 1, Left). This results in two different energy states E_1 (lower, parallel state) or E_2 (higher, antiparallel state) (Figure 1, Right) [21] [22].

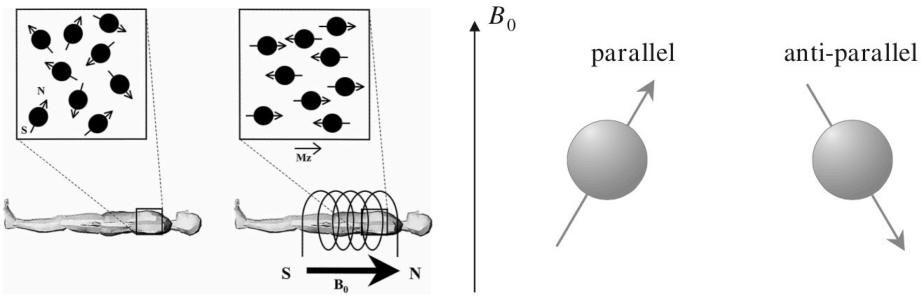

Figure 1 – (Left) Without a magnetic field, the magnetic moments of the nuclei are randomly distributed in space and the Net magnetization vector is zero. When they are submitted to a strong external magnetic field (B_0) they align themselves parallel or antiparallel to the external field, with a few more parallel than antiparallel [21]. (Right) The two possible orientations (parallel and antiparallel) for the proton in an external magnetic field [23].

Actually, the spinning protons do not align perfectly with the external magnetic field. They are constrained by the laws of quantum mechanics and therefore, experience a turning force which makes them precess around the direction of the field (Figure 2, A) [23]. They do so with an angular frequency, also known as natural or Larmor frequency, ω (radians per second), which is directly proportional to B_0. When in the external field B_0, more protons are in the lower energy state E_1 than in E_2. Hence, there is a small resultant longitudinal magnetic field $\boldsymbol{M}_{gz}$ developed by the protons. This magnetization is parallel to B_0, and therefore cannot be measured [22].

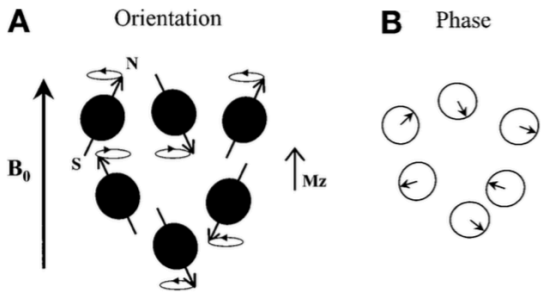

Figure 2 - (A) Individual nuclei spin around their own axes and precessing around the direction of the external field. (B) The phase precession around the axis of the external magnetic field [21].

If an external stimulus, in form of 90º radiofrenquency (RF) energy pulses of exactly the Larmor frequency is applied, the protons resonate, absorb the energy and 'flip' to a higher energy state [22]. Also, the protons start to precess in phase with the pulse, and thus with each other and all the protons' transverse components add up (Figure 2, B). The precession in phase is the origin of the MRI signal. The net magnetization rotates 90º from the positive axis to transverse plane, producing a transverse magnetic field $\mathbf{M}_{xy}$. This rotating transverse magnetization can be measured by detecting the alternating current (AC) it induces in a receiver coil (Figure 3), which is sensitive only to magnetization perpendicular to B_0 [21][23]. Leaving the RF pulse on for twice as long (or doubling its strength) would turn $\mathbf{M}_{gz}$ exactly 180º, and the pulse would then be known as a 180º pulse [23].

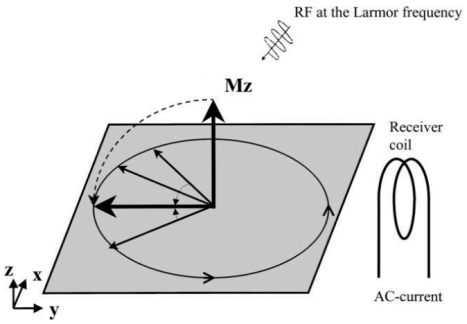

Figure 3 – When the spins are exited with an 90º RF pulse of exactly the Larmor frequency, the net magnetization flips 90º and the spins precess in phase. The rotating net magnetization vector induces an AC in a receiver coil [21].

After the RF frequency transmitter is switched off, the protons will seek the equilibrium state, returning to the low energy state. The protons will return to their original orientation, parallel to the B_0 field, by giving up the extra energy they acquired from the RF pulse, to their neighbouring atoms, in a process known as relaxation. The amplitude of the signal decays exponentially to zero, because the protons rapidly dephase with respect to each other. This signal is known as Free Induction Decay (FID) [23]. The way the neighbouring atoms will absorb this surplus energy depends on the exact nature of the tissue, and gives us a way of distinguishing one type of tissue from another. Therefore, the relaxation is the key to provide tissue contrast in MRI [22].

There are two relaxation processes: transverse relaxation and longitudinal relaxation and they are both independent. The longitudinal relaxation process is the process of realignment to the external magnetic field and is also known as T_1 relaxation time. During T_1 the energy is lost from the spinning protons to the surrounding molecules. A second effect is formed by the interactions between spins as they move around within tissues. This interaction is known as spin-spin relaxation, or T_2 relaxation time and is caused by the loss of the phase coherence amongst the precessing protons in the transverse plane [22][23].

2.2.2 Image Formation

Biological tissues have different T_1 and T_2 values, but T_2 is always shorter than T_1. MRI has the potential to visualize the differences in T_1 and T_2 of different tissues by manipulating the timing of the RF pulses. A second AC signal gives the opportunities to modify the contrast in the images depending on the T_1 and T_2 values of the tissues. To evoke a second AC signal, a second RF pulse is applied and flips the spin by 180º, reversing the dephasing process. As the spins rephase, the amplitude of the AC signal increases and this signal, called echo signal is measured at its maximum [21][22].

Therefore, the most fundamental timing parameters of relevance are repetition time (TR), echo time (TE), and in some cases inversion time (TI) (Figure 4). TR is the time between consecutive acquisitions and TE is the time from the onset of the excitation pulse, which is used for preparing the signal for detection, to the signal

acquisition. In an inversion recovery pulse sequence, TI refers to the time between the inversion pulse and the excitation pulse. For short TR and TE, the contrast in the image will be potentiated by the differences in T_1 value of the tissues (T_1-weighted sequences or T_1 images). Using long TR and TE, the contrast will be dependent on T_2 differences (T_2-weighted sequences or T_2 images) [21].

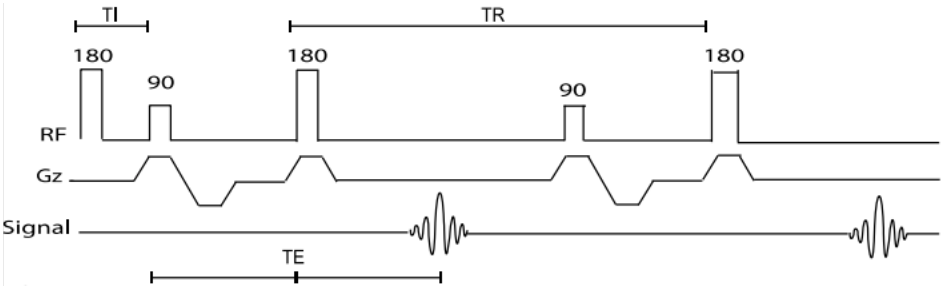

Figure 4 - Example of a pulse sequence showing timing parameters of the application of radio frequency pulse (RF), the onset of gradients in the Z direction (Gz), and the timing of signal acquisition (Signal) [24].

T_1 weighted imaging is normally used in structural imaging. MRI can provide a highly detailed three-dimensional image that allows the examination of brain structures and anatomy. Structural imaging based on MRI is an integral component of the clinical assessment of patients with suspected AD. The most differences reported in patients with AD progression are atrophy of the structures in the medial temporal lobe (MTL). Normally, patients diagnosed with MCI show change to the parahippocampal region. Patients that are at-risk for AD, but have no cognitive deficit, are much more difficult to identify. The presence of atrophy of medial temporal structures is a partially validated candidate marker for early diagnosis, and measurement of progression, of the disease at the MCI stage [24].

2.2.3 MR equipment

The magnet is the main component of the MR system and the magnetic field strength is measured in tesla (T). Most clinical studies to date have used 1.5 Tesla (T) scanners, however many medical centres now have 3T scanners. There are some 7T scanners worldwide, only used for research purposes. The high-field scanners allow for increased resolution for observing structural changes in the same scan time [23][24].

The RF system comprises a transmitter, coil and receiver. MRI radio frequency coils receive and/ or transmit the RF signal at the Larmor frequency. When this current is applied to the coil the alternating B_0 field is produced. Gradient coils are used to apply controlled variations in the main magnetic field, B_0, to provide spatial localization of the signals. The function of a receiver coil is to maximize signal detection, whilst minimizing the noise. Because there is a time delay between transmission and reception of signals, they are often the same coil, called transceivers. A variety of coils are available, which fit closely around parts of the body we are interested in visualize such as, the head, knee and breast [23].

Chapter 3: Voxel-Based Morphometry

Traditional techniques for analysing MRI brain scans include visual assessment by experienced radiologists, which is very time-consuming and is subjective to visual assessments. In the past few years, automated techniques have been developed to identify differences in brain anatomy across groups of subjects.

Brain morphometry is one of the most studied modalities in brain imaging. It is based on the study of the size and shape of the brain and its structures during development, ageing, learning and disease. Voxel-based Morphometry (VBM) is one of the automated techniques for analysing brain morphometry, using statistical parametric mapping approach to investigate focal differences in brain anatomy [25][26]

In particular, VBM has been widely used for studying differences of normal aging and Alzheimer's disease (AD) [27]. For structural MRI, VBM identifies differences in the local composition of brain tissue, discounting large scale differences in gross anatomy and position. This can be achieved by extracting the gray and white matter from all the images, spatially normalizing the gray and white matter images to the same stereotactic space, smoothing, and finally performing a statistical analysis. To do this statistical analysis, a mass-univariate approach is applied where the statistical model called "General Linear Model (GLM)" is performed on a voxel-by-voxel basis to make inference about group differences [28][29].

The GLM is used to identify regions in the brain that are different related to the particular effect that is being studied. Some covariates may introduce variance that is unrelated to the effect of interest in study, such as age, gender and total intracranial volume (TIV). Therefore, it is possible to identify the differences between groups of patients while controlling for the effects of nuisance covariates [25][29].

There are several approaches that could be used for preprocessing the data. In this thesis, the preprocessing was done in four main steps: (1) Realignment and set image origin, (2) segmentation, (3) normalization and (4) smoothing. For the ADNI project data the normalization and smoothing was done using the DARTEL toolbox in SPM.

3.1 Preprocessing of Structural MRI

3.1.1 Realignment and image origin

When comparing different scans from different subjects, some images may have different orientations. The goal of realignment is to align all images to one specific image, so that all MRI images are in the same orientation and position. In a first step of estimation, the first image in the list specified is used as a reference to which all subsequent scans are realigned. This routine realigns the different scans using last squares approach and a 6 parameter (rigid body) spatial transformation. In a second step, this function reslices the registered images to make them match the first image selecting voxel-by-voxel [30].

When opening a structural image in SPM, by default, the origin of the image is positioned in the centre of the image and not in a specific anatomical landmark. SPM's normalization is sensitive to the origin used as an initial alignment between the image and the template. In the MRI template used by SPM, the origin is the anterior commissure (AC), a thin white matter tract between the olfactory areas of each hemisphere. Thereby, if the location of the AC is not set, SPM's starting estimate will locate the AC at the centre of the volume. Therefore, it is important to set the anterior commissure as a landmark for all the images before normalization. The anterior commissure is represented in Figure 5. After this, all the images will have the same origin, which is crucial for a good normalization [31].

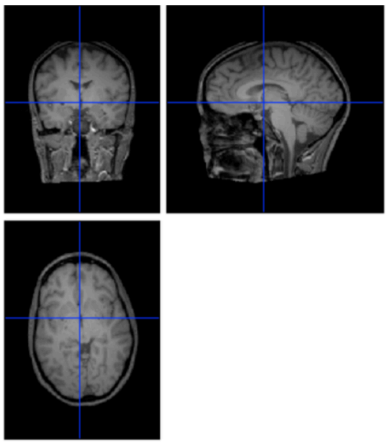

Figure 5 - Setting the anterior comissure as the origin in SPM (Images obtained in SPM8).

3.1.2 Segmentation

MRI scans may provide a detailed insight into the anatomy of the brain. However, not all of the anatomical information is interesting for analysis. Healthy brain tissues can generally be classified into three main tissue types: grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) [30][32]. This procedure generates new images, each image contain one of the tissue types (Figure 6). Tissue classification can be achieved by a statistical approach commonly known as segmentation and is performed by combining a priori probability maps or "Bayesian priors" with the data from images. These probability maps are tissue probability maps (TPMs, provided by the International Consortium for Brain Mapping), which encode knowledge of the spatial distribution of different tissues in normal subjects. The TPMs inform a mixture of Gaussians model that classifies each voxel as a tissue type by taking into account its position and image intensity [28][29]. In this thesis only the grey matter images were analysed.

Additionally, the segmentation also incorporates a bias correction component to account for smooth intensity variation caused by different positions of cranial structures within the MRI coil [28].

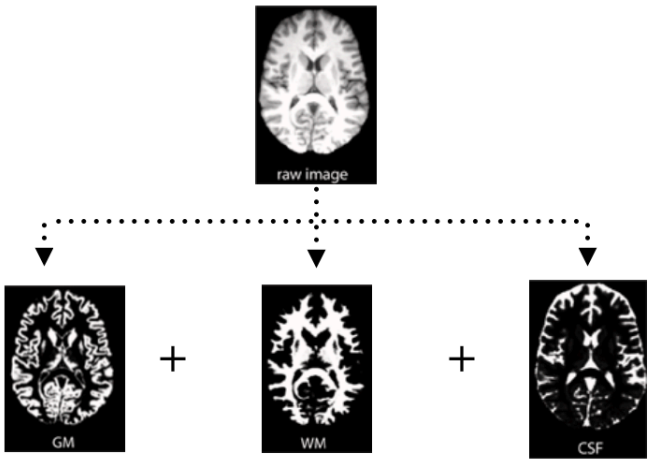

Figure 6 - Segmentation in VBM using SPM8 (Images obtained in SPM8).

3.1.3 Spatial Normalization

When performing a group analysis it is essential that all brains have the same size and orientation. Normalization is based on the principle that each voxel from different scans represents the same part of the brain, avoiding introducing artifactual changes in the voxel values [30]. In SPM, this is accomplished by spatially normalizing the images into the space defined by the Montreal Neurological Institute (MNI) template. Spatial normalization consists of registering each of the images to the same template and minimizing the residual sum of squared differences between them. In this way, both global and local structural differences between brains are removed [25][28].

Different algorithms can be used to perform this matching. The most commonly applied algorithm available in SPM involves performing a 12-parametric affine transformation (three translations, three rotations, three scales and three shearing) followed by a nonlinear registration using a mean squared difference matching function. The nonlinear registration minimizes the cost function between the MR image and the template and, simultaneously, maximizes the smoothness of the deformations [25][28].

3.1.4 Smoothing

Smoothing involves blurring the MRI images by convolving them with an isotropic Gaussian kernel of a specified width at half-maximum (FWHM) in millimetres (Figure 7). After smoothing, each voxel in the smoothed images will contain the average concentration of matter from around the voxel (where the region around the voxel is defined by the form of smoothing kernel), making the subsequent voxel-by-voxel analysis comparable to a region of interest [28][31].

One of the main reasons to use smoothing is rendering the data more normally distributed by the central limit theorem, increasing the validity of the parametric

statistical tests. The smoothing step also aids in compensating the inexact nature of the spatial normalization. During this procedure, the predominantly high-frequency noise is suppressed and the overlap of activation between subjects is increased. Finally, smoothing reduces the effective number of statistical comparisons by increasing the inter-voxel correlation in the image [28].

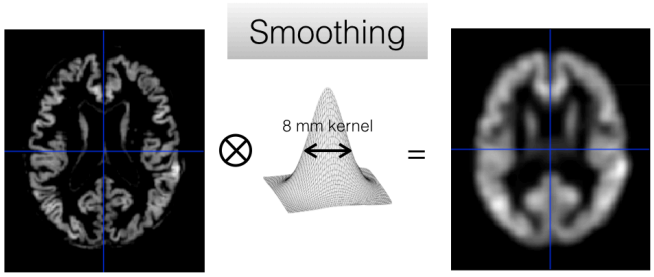

Figure 7 – Smoothing using a 8 mm gaussian kernel in VBM (Images obtained in SPM8).

3.1.5 DARTEL toolbox

In computational anatomy, several approaches have gained considerable interest due to their ability to effectively align structures across individuals in a way that respects anatomical constraints. The spatial normalization described earlier uses basis functions to deform brains to match one another without regard to the nature of the material being warped. This imperfect registration to a common template can lead to false estimates. New approaches in computational anatomy were developed to use models based on physical phenomena. For instance, Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) uses a specific kind of transformation known as diffeomorphism. This algorithm, implemented in SPM, allows a sharper normalization process and is a promising approach to solve the spatial normalization problem [33][34].

This diffeomorphic transformation from one image to another can be represented as a "flow field vector" describing, at every point, the movement of the data in that voxel from the original to the transformed image [34]. This algorithm only works with images with isotropic voxels, identical dimensions, and which are in approximate alignment with each other. Then, in a first step, images should be imported in order to write rigidly transformed versions of tissue class images in a close alignment as possible with the tissue probability maps [30].

The imported rigidly transformed images are used to generate a specific-subject template. DARTEL aims to increase the accuracy of inter-subject alignment by modelling the shape of each brain using millions of parameters. In DARTEL, a warping of all brain images is done by simultaneously matching grey matter among the images, while simultaneously aligning white matter. This can be achieved by generating the specific-subject template, to which the data are iteratively aligned [30][31]. During this process, an initial affine registration of the template with the TPM data released with SPM is done, ensuring the normalization to the MNI space. DARTEL also performs smoothing to the images. Smoothed spatially normalised images are generated in such a way that the original signal is preserved.

3.2 Statistical Analysis

Following preprocessing, the final step of a VBM analysis consists of applying a mass-univariate approach based on General Linear Models (GLMs) performed on a voxel-by-voxel basis. The GLM model is used to identify regions that are significantly related to the specific effects under study. This is a flexible framework and includes many different tests, ranging from group comparisons and identifying regions of gray matter concentration (or white matter) that are related to specific covariates (e.g. age or disease) to complex interactions between different effects of interest [28]. The statistical analysis comprises the following steps: (1) specification of the GLM design matrix and MRI data files, (2) estimation of GLM parameters using classical or Bayesian approaches and (3) examination of results using contrast vectors to produce Statistical Parametric Maps (SPMs) or Posterior Probability Maps (PPMs) [30].

Standard parametric statistical tests, such as t-tests and F-tests, are used to test the hypotheses. The null hypothesis assumes that there is no difference in tissue volume between the groups in question. These analyses generate statistical maps showing all voxels of the brain that refute the null hypothesis and show significance to a certain p-value. The maps are usually expressed as colour maps with a scale representing the t statistic (or F statistic) [25][28].

This analysis is designated as univariate, since only one dependent variable is at stake. However, this approach is not sensitive to the spatial covariance of the data. Yet, these simple tests are the building blocks of more complex approaches and it is can be informative to perform this analysis before as an alternative to multivariate approaches.

3.2.1 Univariate GLM

Given a data set, a linear regression model assumes that the relationship between the dependent variable and the regressors is linear. This relationship is modelled through a disturbance term or error variable ϵ (an unobserved random variable that adds noise to the linear relationship between the dependent variable and the regressors). Thus, a basic linear regression that explains a dependent continuous variable y by the behaviour of a single independent continuous variable x, can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon \quad (1)$$

where β_0 describes the value of y when x is equal to zero (also known as intercept), β_1 is a regression coefficient that represents the slope of the line, and ϵ is the residual error of the model. The regression coefficient β_1 describes the change in y that is associated with a unit change in x. The regressor is positive if the relation between x and y is direct, and negative if inverse. A p-value is attached to the regressor and the null hypothesis postulates that there is no relation between y and x [35].

The general linear model is an extension of this model, including more independent variables, each with its own regressor. Thus, the general linear model for a response variable can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon_k \tag{2}$$

In order to compute a general linear model for a set of observations, an observation vector $\boldsymbol{y}_{N\times 1}$, where N is the number of observations, is related to K unknown parameters, where K is the number of the predictor variables, represented by a vector $\boldsymbol{\beta}_{K\times 1}$ through a known design matrix $\boldsymbol{X}_{\boldsymbol{d}_{N\times K}}$. The error term is also included, to absorb the unexplained variance of the system. Summarizing, each observation can be described as a linear combination of independent covariates that influence the outcome [36][37]:

$$\boldsymbol{y}_{N\times 1} = \boldsymbol{X}_{\boldsymbol{d}_{N\times K}}\boldsymbol{\beta}_{K\times 1} + \boldsymbol{\epsilon}_{N\times 1} \tag{3}$$

where $\boldsymbol{y}$ the column vector of observations, $\boldsymbol{\beta}$ the column vector of parameters, $\boldsymbol{\epsilon}$ the column vector of error terms and $\boldsymbol{X_d}$ the design matrix. The design matrix specifies the experimental knowledge about the expected signal, defining the nature of hypothesis testing to be implemented. The design matrix has one row for each scan (observations) and one column for each effect of explanatory variable (predictor variables) [35][36]. The GLM system of equations can be expressed using matrix notation:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{d_{11}} & \cdots & x_{d_{1k}} \\ \vdots & \ddots & \vdots \\ x_{d_{N1}} & \cdots & x_{d_{NK}} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{4}$$

Data analysis often aims to evaluate the fit of the model to the data, test hypothesis regarding model parameters and predict future observations. The parameters in the GLM model need to be estimated in order to best fit the data. The least squares estimation is a well-known method that estimates the parameters by minimizing the residual sum squares.

3.2.2 Ordinary Least Squares and Multicollinearity

The ordinary least squares (OLS) algorithm is a method for estimating the unknown parameters in a linear regression model. Model fit can be determined by comparing the observed scores of **Y** with $\widehat{\mathbf{Y}}$ (values of **Y** predicted by the regression). The difference between the two values (also called as residual or deviation) provides an indication of how well a model predicts each data point. The residual sum of

squares (RSS) is the sum of the squared residuals and provides a measure of model fitting for the OLS regression model. A poorly fitting model will deviate from the data and will produce large RSS, whereas a good-fitting model will have a relatively small RSS [38]. OLS estimates the parameters by minimizing the RSS:

$$\min_{\boldsymbol{\beta}} RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X_d}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X_d}\boldsymbol{\beta}) \tag{5}$$

Assuming that the noise is normally distributed in the model, the parameters $\boldsymbol{\beta}$ are estimated from the data using [39]:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X_d}^T\boldsymbol{X_d}\right)^{-1}\boldsymbol{X_d}^T\boldsymbol{Y} \tag{6}$$

where $\boldsymbol{X_d}^+ = \left(\boldsymbol{X_d}^T\boldsymbol{X_d}\right)^{-1}\boldsymbol{X_d}^T$ denotes the (Moore-Penrose), pseudo inverse of $\boldsymbol{X_d}$. The fitted data $\widehat{\boldsymbol{Y}}$ (predicted by the model) is defined as:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X_d}\widehat{\boldsymbol{\beta}} = \boldsymbol{X_d}\boldsymbol{X_d}^+\boldsymbol{Y} \tag{7}$$

Therefore, the estimated noise is given by:

$$\boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I_n} - \boldsymbol{X_d}\boldsymbol{X_d}^+)\boldsymbol{Y} = \boldsymbol{r} \tag{8}$$

where

$$\boldsymbol{RSS} = \boldsymbol{r^T}\boldsymbol{r} \tag{9}$$

The ordinary least square method is based on several assumptions. One important assumption is that there is no linear relationship between the explanatory variables. However, if this is not true the multicollinearity problem will appear threatening both the assumption and usage of ordinary least square (OLS). When multicollinearity is present, two or more predictor variables in the regression model are highly correlated, meaning that one can be linearly predicted from the other. Also, the predictor matrix will be singular, and it will not be possible to find the inverse matrix of the explanatory variables. Consequently, the OLS method will provide infinite solutions [40].

Problems that suffer from this difficulty are known as ill conditioned, since there is not enough information in the data to precisely specify the solution. Collinearity makes it more difficult to achieve significance of the collinear parameters. In this situation an approach that is frequently adopted consists in restricting the choice of functions in some way. Such restriction or bias is referred as regularisation [41].

One popular method to overcome this problem and perform the regularisation is known as Ridge Ordinary Least Squares regression (ROLS, also known as the Tikhonov regularization). In ROLS, the parameter $\boldsymbol{\beta}$ is estimated by minimizing the object function:

$$\boldsymbol{\chi}^2 = \left|\left|\boldsymbol{X_d\beta} - \boldsymbol{Y}\right|\right|^2 + h^2\left|\left|\boldsymbol{\beta}\right|\right|^2 \tag{10}$$

The first term of the object function is the residual sum of squares and the second term penalizes a large norm of the parameter vector $\boldsymbol{\beta}$ (and therefore penalizes a long vector of parameters). The regularization parameter is h and it will determine the trade-off between minimizing the residual sum of squares and minimizing the norm of the estimate. This method overcomes the multicollinearity problem by adding a small quantity to the diagonal of $\boldsymbol{X_d}^T$. For a given regularization parameter the regularized estimation is:

$$\widehat{\boldsymbol{\beta}}_h = \left(\boldsymbol{X_d}^T\boldsymbol{X_d} + h^2\boldsymbol{I}\right)^{-1}\boldsymbol{X_d}^T\boldsymbol{Y} \tag{11}$$

If the regularization parameter is $h = 0$, the estimate $\widehat{\boldsymbol{\beta}}_h$ will be the ordinary least squares estimation. Otherwise, if $h \to \infty$, the estimate $\widehat{\boldsymbol{\beta}}_h$ will approach zero. For intermediate values of the h, the estimate $\widehat{\boldsymbol{\beta}}_h$ is "shrunk" toward zero when compared with the ordinary least squares estimate. Therefore, this is a biased estimation [42]. The problem with this method is that there is no way to know how much bias has been introduced.

An alternative approach to deal with collinearity is to transform the raw data $\boldsymbol{X_d}$ to create a new orthogonal design matrix. The orthogonality assumes that, if two variables are orthogonal, knowing the value of one variable does not provide information as to the value of the other and the correlation between them is zero, producing linearly independent regressors and meeting the criteria for OLS [43]. These regressors can subsequently be used instead of the original correlated regressor values. This technique can be done using a Gram Schmidt orthogonalization. Considering two linear independent vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ and $\boldsymbol{u}_1 = \boldsymbol{v}_1$, we want to find a vector $\boldsymbol{u}_2$, which is perpendicular to $\boldsymbol{u}_1$ and the distance of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ is the same as the distance of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Therefore, the goal is to find a number $a \in \mathbb{R}$ such that [44]:

$$\boldsymbol{u}_2 = a\boldsymbol{u}_1 + \boldsymbol{v}_2, \quad \boldsymbol{u}_2 \perp \boldsymbol{u}_1 \tag{12}$$

The inner product of $\boldsymbol{u}_1$ with $\boldsymbol{u}_2$ will be:

$$0 = (\boldsymbol{u}_2, \boldsymbol{u}_1) = a(\boldsymbol{u}_1, \boldsymbol{u}_1) + (\boldsymbol{v}_2, \boldsymbol{u}_1) = a\|\boldsymbol{u}_1\|^2 + (\boldsymbol{v}_2, \boldsymbol{u}_1) \tag{13}$$

$$a = -\frac{(\boldsymbol{v}_2, \boldsymbol{u}_1)}{\|\boldsymbol{u}_1\|^2} \tag{14}$$

For a third vector $\boldsymbol{v}_3$, after choosing $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, as above, $\boldsymbol{u}_3$ will be:

$$\boldsymbol{u}_3 = a_1 + a_2\boldsymbol{u}_2 + \boldsymbol{v}_3 \tag{15}$$

In order to find a_1, the inner product with $\boldsymbol{u}_1$ will be:

$$0 = (\boldsymbol{u}_3, \boldsymbol{u}_1) = a_1\|\boldsymbol{u}_1\|^2(\boldsymbol{v}_3, \boldsymbol{u}_1) \tag{16}$$

$$a_1 = -\frac{(\boldsymbol{v}_3, \boldsymbol{u}_1)}{\|\boldsymbol{u}_1\|^2} \tag{17}$$

and the inner product with $\boldsymbol{u}_2$ to find a_2:

$$0 = (\boldsymbol{u}_3, \boldsymbol{u}_2) = a_2\|\boldsymbol{u}_2\|^2(\boldsymbol{v}_3, \boldsymbol{u}_2) \tag{18}$$

$$a_2 = -\frac{(\boldsymbol{v}_3, \boldsymbol{u}_2)}{\|\boldsymbol{u}_2\|^2} \tag{19}$$

Thus:

$$\boldsymbol{u}_1 = \boldsymbol{v}_1$$

$$\boldsymbol{u}_2 = \boldsymbol{v}_2 - \frac{(\boldsymbol{v}_2, \boldsymbol{u}_1)}{\|\boldsymbol{u}_1\|^2}\boldsymbol{u}_1$$

$$\boldsymbol{u}_3 = \boldsymbol{v}_3 - \frac{(\boldsymbol{v}_3, \boldsymbol{u}_1)}{\|\boldsymbol{u}_1\|^2}\boldsymbol{u}_1 - \frac{(\boldsymbol{v}_3, \boldsymbol{u}_2)}{\|\boldsymbol{u}_2\|^2}\boldsymbol{u}_2 \tag{20}$$

The relative importance of the explanatory variables can be approximated by the relative importance of the orthogonal counterparts. However, the effects of orthogonalization on the interpretation of the resulting parameter estimates are often misunderstood. The orthogonal variables are only approximations of the original variables and may not be highly related to the original variables and therefore, there are some questions regarding the importance of the original variables [45].

3.2.3 Contrasts

After the parameter estimation, the t- and F-statistics may be used to make inferences from the data. The relationships postulated by the GLM are contained within a specified design matrix. Contrasts are used to test for a specific effect. Contrasts are defined as vectors (t-contrasts) or matrices (F-contrasts), which can be used to focus the inferential analysis on a subset of regressors, defining the relationship between them. All the other independent variables are ignored and are often seen as nuisance variables, meaning that their effect is taken into account but removed from the analysis [46].

For instance, assuming a certain GM volume as one dependent variable and 4 independent regressors $\boldsymbol{\beta} = \beta_1 + \beta_2 + \beta_3 + \beta_4$ corresponding to the independent variables related to: control, disease, age and gender. If the aim is to find where there is more GM volume in control subjects that disease subjects, excluding the nuisance variables TIV and age, the t-contrast may be defined as [1 -1 0 0]. The linear decrease of GM due to disease is represented as -1.

The F-contrast, by the other hand, is used to test whether any of several linear constrains is true. It can be considered to be the general form of the t-contrast. The F-contrast will test simultaneously the significance of including many (or even one) regression coefficients in the multiple linear regression model. It can be defined as an OR statement containing several t-contrasts, modelling multiple linear hypothesis.

As a simple example, assuming a model comprising T1-weighted images from healthy controls and patients from one scanner, and T1-weighted images from healthy controls and patients from a second scanner. F-contrast can be used to test specific hypotheses. The main effect of disease may be defined as a F-contrast, $\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$, and it will find the brain region where the GM atrophy is present in the scans from the first scanner or from the second scanner. F contrast can be also used to test for the interaction between different scans effect and disease.

Within the GLM, the t-test can be computed in order to make inferences about the linear combinations of regressors. To do so, the value of the error mean square (MS_E, $\hat{\sigma}^2$) is estimated:

$$\hat{\sigma}^2 = \frac{\epsilon^T \epsilon}{N - p} \tag{21}$$

where $\epsilon^T \epsilon$ is the residual sum of squares and $N - p$ are the degrees of freedom.

The parameters estimates are normally distribute, then $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \left(\boldsymbol{X_d}^T \boldsymbol{X_d}\right)^{-1}\right)$. Assuming that $\boldsymbol{c}$ is the contrast vector containing p weights, the following distribution is obtained:

$$\boldsymbol{c}^T \widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{X_d}^T \boldsymbol{\beta}, \hat{\sigma}^2 \boldsymbol{c}^T \left(\boldsymbol{X_d}^T \boldsymbol{X_d}\right)^{-1} \boldsymbol{c}\right) \tag{22}$$

and the t-value can be computed by:

$$T = \frac{\boldsymbol{c}^T \widehat{\boldsymbol{\beta}}}{\sqrt{\widehat{\boldsymbol{\sigma}}^2 \boldsymbol{c}^T \left(\boldsymbol{X_d}^T \boldsymbol{X_d}\right)^{-1} \boldsymbol{c}}} \tag{23}$$

Finally, the p-value can be calculated by comparing the t-value T with a t-distribution with $N - p$ degrees of freedom [43][45].

In SPM statistical analysis, when analyzing structural scans, a second level analysis is performed. There are several design models. In this thesis the paired sample t-test and a full factorial designs were used. The paired sample t-test is a design for within group comparisons. For instance, this test can be used if the purpose is to compare two scans from the same subject. This is generally used in fMRI to compare different times. Therefore, the number of pairs should match the number of subjects. The full factorial model is used when the purpose is to test all main effects and interactions between two conditions (e.g. disease and different scanners) [30].

Chapter 4: Machine Learning – Pattern Recognition

Univariate analysis has been widely used for making inferences on brain function and structure. However they are limited to the type of research questions that they can address. The univariate approach assumes that activity in one brain region occurs independently from activity in other regions. However, there is a growing recognition that the spatial dependencies among signal from different brain regions should be properly modelled. Consequently, univariate analysis may have poor sensitivity when effect of interest (e.g. cognitive paradigm, pathology, etc.) has a distributed multidimensional effect on activation [47][48]. Likewise, inference cannot be made at the single subject level - instead these methods were developed for group-based inference. Therefore, applying these methods for diagnosing purposes would be impractical since it is not possible to assess the generalisation of the methods to an independent cohort [49].

Multivariate statistical methods are an alternative to univariate methods. Instead of performing a series of univariate analysis each with only one dependent variable, multivariate models comprise a single analysis with multiple dependent variables [49]. Thus, it provides multiple levels of inference and allows researchers to test how distributed patterns of activation across multiple voxels relate to experimental variables.

Multi-voxel pattern analysis is a multivariate approach and has gained special interest in the neuroimaging community by focusing on the analysis and comparison of distributed patterns of activity. One example is machine learning based predictive models. The machine learning algorithms are first trained on a set of data (brain activation/anatomy) and corresponding variable of interest (e.g. the status of a patient, health vs. disease), and then aim to predict a variable of interest for a test case. Due to their multivariate properties, these methods can accomplish relatively high sensitivity, being able to detect subtle changes [48][50]. These methods are becoming increasingly popular and one major application has been to provide clinical diagnosis and prognosis in patients with diseases, such as Alzheimer's disease [5].

There is a wide range of machine learning algorithms that can be divided into three main classes: supervised, semi-supervised and unsupervised learning, based on whether the training data instances are labelled. In supervised learning, the learner is supplied with labelled training examples, where both the input and the groundtruth output are specified. The semi-supervised uses a combination of both labelled and unlabelled data for training and in unsupervised learning, the correct output (which may be unknown) is not provided with the input [51]. For the purposes of this thesis, the focus will be on supervised learning.

The aim of the learner, in supervised learning, is to learn the mapping from inputs to outputs. The learning process consists in learning a function f that accounts for the input/output pairs of the form $(\boldsymbol{x_i}, \boldsymbol{y}_i)$, where $\boldsymbol{x}_i$ is a possible input and $\boldsymbol{y}_i$ is the correctly labelled output associated with it. This function is called a classification if the output is discrete and regression if the output is continuous [51].

In case of MRI data, the input data are the scans from both groups. Thus, it is required to create a numerical representation of the example images. This process consists in transforming a 3 dimensional brain scan into a long vector of features (voxels) within the brain. This is called feature extraction and is expected that the

feature set will extract the relevant information from the input data in order to perform the desired classification task [52].

Having created a data matrix with all the subjects and all the features selected, the regression and/or classification procedures can be done. In the classification case, the process comprises two phases: training and testing (Figure 8). During the training phase, the algorithm finds a hyperplane that separates the examples in the input space according to their class labels. The classifier is trained by providing examples of the form $\{\boldsymbol{x}, \boldsymbol{c}\}$, where $\boldsymbol{x}$ represents a spatial pattern (e.g. features) and $\boldsymbol{c}$ is the class label (e.g. patient or control). Once the decision function is learned from the training data it can be used to predict the class of a new test sample [52].

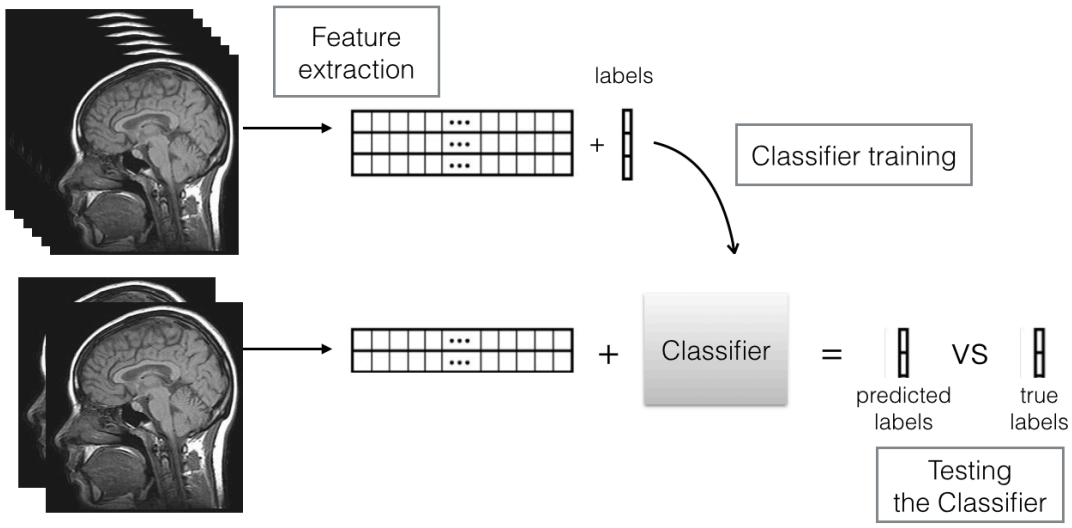

Figure 8 - The general process of classification algorithm. Figure adapted from F.Pereira et al [5].

Identifying patterns in a finite set of data presents very different and distinctive challenges. Therefore, a pattern analysis algorithm must fulfil certain requirements to be considered effective. First of all, pattern analysis must be computationally efficient, and therefore must be able to handle very large datasets. There is a computational shortcut, which makes it possible to represent linear patterns efficiently in high-dimensional spaces to ensure adequate representational power. This shortcut is also known as "kernel trick" and it consists in building a kernel matrix on both the training and test points so that the transformed kernel matrix contained all the information required. Most algorithms are implemented in such a way that the coordinates of the embedded points in the data matrix are not needed, only their pairwise inner products. The pairwise inner products can be computed directly from the original data using a kernel function [41].

The second challenge that a pattern analysis algorithm must address is the fact that in real-life applications data is often corrupted by noise. The algorithm must be able to handle noisy data and identify approximate patterns. An algorithm with this property is considered to be robust [41].

Finally, the patterns the algorithm identifies must be genuine patterns of data source and not just an accidental relation occurring in the finite training set [41].

Depending on the machine learning method used, there could be many possible decision boundaries or hyperplanes (e.g. linear discriminant analysis, support vector machine, Gaussian process, etc.). In this thesis Gaussian process learning will be studied.

4.1 Gaussian Process

Gaussian processes (GP) are one of the most widely used families of stochastic processes for modelling dependent data observed over time, or space, or time and space. GPs can be used to define prior distributions on functions. If combined with suitable noise models or likelihoods, Gaussian process models allow one to perform Bayesian nonparametric regression, classification, and other machine learning tasks. Such processes are quite popular due to two essential properties: (1) a GP is completely determined by its mean and covariance functions, facilitating the model fitting and (2) solving the prediction problem is relatively straightforward [53].

4.1.1 Definition of a Gaussian Process

A Gaussian distribution is given by:

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \quad (24)$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ the covariance matrix. Thus, a Gaussian distribution is fully specified by a mean vector and a covariance matrix [10].

A Gaussian process is a conditional probabilistic model defined as a set of random variables indexed by a continuous variable: $f(\boldsymbol{x})$, which all the variables have consistent Gaussian distribution. For instance, a particular finite subset of these random function variables $\boldsymbol{f} = \{f_1, f_2 \ldots f_N\}$ with corresponding inputs (indices) $\boldsymbol{X} = \{x_1, x_2, \ldots x_N\}$ are distributed multivariate GP [10][54]:

$$p(\boldsymbol{f}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{K}) \quad (25)$$

where $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{K})$ denotes a gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$.

A Gaussian process is then fully specified by a mean function $m(\boldsymbol{X})$ and covariance function $k(x_p, x_q)$, where x_p and x_q are samples from the dataset:

$$f(\boldsymbol{X}) \sim \mathcal{GP}\left(m(\boldsymbol{X}), k(x_p, x_q)\right) \quad (26)$$

If a Gaussian process is defined as a collection of random variables, the consistency requirement, also known as the marginalization property, is implied. This means that the examination of a larger set of variables does not change the distribution of the smaller set. For instance, if the GP specifies $(\boldsymbol{y}_1, \boldsymbol{y}_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$, then it must also specify $y_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{K}_{11})$ where $\boldsymbol{K}_{11}$ is the relevant submatrix of $\boldsymbol{K}$ [10]. Then, the marginalisation follows the equation:

$$p(f_1) = \int p(f_1, f_2) df_2 \tag{27}$$

In the context of machine learning, there is usually little or no prior knowledge about the mean function of a given Gaussian process. For notational simplicity, since GPs are a linear combination of random variables with normal distribution, the mean function is commonly assumed to be zero. Thus, what relates one observation to another is just the covariance function, $\boldsymbol{k}(x_p, x_q)$. One example is the linear covariance function. The linear kernel is non-stationary, meaning that it depends on the absolute location of each point and is defined as:

$$\boldsymbol{k}(x_p, x_q) = \sigma_f^2 (x_p - c)(x_q - c) \tag{28}$$

where σ_f^2 is the hyperparameter and c is a constant that determines the x-coordinate of the point that all the lines in the posterior go through [55].

The 'squared exponential' covariance function is probably the most popular choice. This function is stationary since it only depends on the relative position of two points. The covariance function specifies the covariance between pairs of random variables [56]:

$$\boldsymbol{k}(x_p, x_q) = \sigma_f^2 \exp\left[-\frac{(x_p - x_q)^2}{2l^2}\right] \tag{29}$$

where σ_f^2 and l are the hyperparameters. The maximum allowable value of the covariance function is defined as σ_f^2 and the effect of this separation will depend on the length parameter l. The length-scale (l) characterizes the distance in input space before the function value can change significantly. Short length-scales mean that the predictive variance can grow rapidly away from the data points, and all the predictions are little correlated. If $x_p \approx x_q$, then $\boldsymbol{k}(x_p, x_q)$ approaches the maximum value and $\boldsymbol{f}(x_p)$ is nearly perfectly correlated with $\boldsymbol{f}(x_q)$. Otherwise, if x_p is distant from x_q, then $\boldsymbol{k}(x_p, x_q) \approx 0$ and the two points are not correlated. Thus, during the interpolation at new x values, distant observations will have negligible effect. In most cases, the choice of hyperparameters can significantly influence the performance of the GP [56][57].

In more realistic modelling situations, where measurement errors may occur, data are often noisy. Each observation y can be thought of as related to an underlying function $f(x)$ through a Gaussian noise model [56]:

$$\boldsymbol{y} = f(x) + \mathcal{N}(0, \sigma_n^2) \tag{30}$$

and the covariance function will become:

$$\boldsymbol{k}(x_p, x_q) + \sigma_n^2 \boldsymbol{\delta}(x_p, x_q) = \sigma_f^2 \exp\left[-\frac{(x_p - x_q)^2}{2l^2}\right] + \sigma_n^2 \boldsymbol{\delta}(x_p, x_q) \tag{31}$$

where σ_n^2 is the variance of the noise and $\boldsymbol{\delta}(x_p, x_q)$ is a Kronecker delta which is one if $p = q$ and zero otherwise.

For Gaussian process regression/classification, the covariance function is computed among all possible combinations of these points, summarized in three matrices:

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{k}(x_1, x_1) & \boldsymbol{k}(x_1, x_2) & \dots & \boldsymbol{k}(x_1, x_n) \\ \boldsymbol{k}(x_2, x_1) & \boldsymbol{k}(x_2, x_2) & \dots & \boldsymbol{k}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{k}(x_n, x_1) & \boldsymbol{k}(x_n, x_2) & \dots & \boldsymbol{k}(x_n, x_n) \end{bmatrix} \tag{32}$$

$$\boldsymbol{K}_* = [\boldsymbol{k}(x_*, x_1)\, \boldsymbol{k}(x_*, x_2) \dots \boldsymbol{k}(x_*, x_n)] \qquad \boldsymbol{K}_{**} = \boldsymbol{k}(x_*, x_*)$$

where **K** denotes the $n \times n$ matrix of covariance evaluated at all pairs of training points, $\boldsymbol{K}_*$ denotes the $n_* \times n$ of covariance at all pairs of training and test points (n_* is the number of test cases) and $\boldsymbol{K}_{**}$ denotes the $n_* \times n_*$ matrix of the covariance evaluated at all pairs of testing points [10][56].

4.1.2 Gaussian Process Regression

Gaussian process regression (GPR) is at the core of many machine learning tasks. Given n observations (Figure 9), the aim of regression is to incorporate the knowledge that the training data provides about the function and then predict y_*, based on the testing data point x_* [56].

The training data set consists of n input vectors $\boldsymbol{X} = x_1, x_2, \dots, x_n$ and corresponding continuous outputs $\boldsymbol{y} = y_1, y_2, \dots, y_n$. For notational convenience, quantities carrying an asterisk refer to test points. For instance f_* contains the latent function values for the n_* test points $\boldsymbol{X}_* = [x_{*,1}, \dots, x_{*,n_*}]$ and $\boldsymbol{y}_*$ are the corresponding predicted outputs.

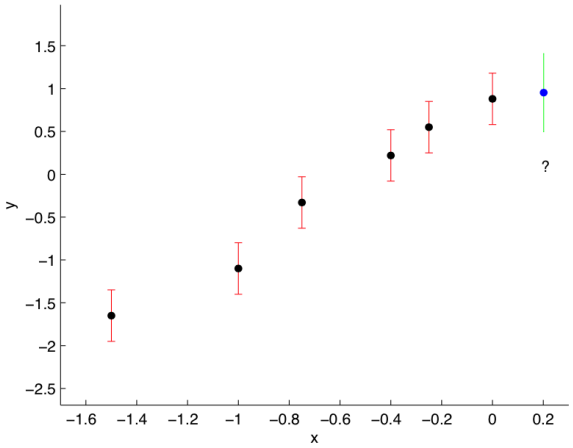

Figure 9 – Six noisy data points, with error bars indicated with vertical lines, the seventh point at $x_* = 0.2$ is the point to be estimated [56].

Assuming the outputs are noisily observed, a single estimation of $\boldsymbol{f}(x)$ is not enough, and a probability distribution over likely functions is needed. Since the Gaussian process regression model is a fully probabilistic Bayesian model, it allows the definition of a probability distribution on functions, $p(\boldsymbol{f})$. This can be used as a Bayesian prior for regression and Bayesian inference can be applied to make prediction from data, and the posterior is:

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})} \tag{33}$$

where $p(\boldsymbol{y}|\boldsymbol{f})$ is the likelihood, $p(\boldsymbol{f}|\boldsymbol{X})$ is the prior and $p(\boldsymbol{y}|\boldsymbol{X})$ is the marginal likelihood. Thus, GP solves the regression problem by giving probabilistic predictions of possible interpolating functions $\boldsymbol{f}$ [54].

We assume that the continuous outputs observations $\boldsymbol{y}$ are contaminated with Gaussian noise ϵ with variance σ_n. The noise that affects the process is supposed to be random, thus no correlation between different inputs is expected, and so the term σ_n is only presented on the diagonals of the covariance matrix. Therefore $\boldsymbol{y}$ is represented as being the sum of a function and the Gaussian noise:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \epsilon, \qquad \epsilon = \sigma_n^2 \boldsymbol{\delta}(x_p, x_q) \tag{34}$$

Equivalently, the noise model, or likelihood is:

$$p(\boldsymbol{y}|\boldsymbol{f}) = N(\boldsymbol{f}, {\sigma_n}^2 \boldsymbol{I}) \tag{35}$$

where $\boldsymbol{I}$ is the identity matrix. Integrating over the unobserved functions variables $\boldsymbol{f}$ gives the marginal likelihood [54]:

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} = N(0, \boldsymbol{K} + {\sigma_n}^2\boldsymbol{I}) \tag{36}$$

The performance of the regression model is dependent on how well the covariance function is selected, and how well the parameters are estimated. On selecting the squared exponential as the covariance function, the hyperparameters of the model are represented as $\boldsymbol{\theta} = \{l, \sigma_f, \sigma_n\}$ [57].

The GP framework provides access to the model evidence (the marginal likelihood) which provides an elegant solution for selecting the hyperparameters from the training data obviating the need for schemes such as nested cross validation. This is one of the major advantages in Gaussian process models. This is possible since the GP is a fully probabilistic model and the maximum estimate of $\boldsymbol{\theta}$ occurs when $p(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y})$ is at its greatest. Often we don't have strong prior evidence for $\boldsymbol{\theta}$, therefore finding the best estimate of $\boldsymbol{\theta}$ corresponds to minimizing the negative log marginal likelihood with respect to $\boldsymbol{\theta}$ (the hyperparameters of the covariance). Therefore, the result of the integration over $\boldsymbol{f}$ using the logarithmic identity, gives the negative log marginal likelihood:

$$-\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{y}^{\boldsymbol{T}}(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y} + \frac{1}{2}\log|\boldsymbol{K} + \sigma_n^2\boldsymbol{I}| + \frac{n}{2}\log(2\pi) \tag{37}$$

The problem of learning with Gaussian process is exactly the problem of learning the hyperparameters. The optimization occurs using only the training data, helping to preventing overfitting. The three different terms in the marginal likelihood play different roles. The first term involves the past observations $\boldsymbol{y}$ and is known as data-fit term. The second term depends only on the covariance matrix, working in an analogous way to the regularization terms in linear regression, adding a penalty as the complexity increases. The last term is only a normalizing constant, and does not have a specific role [57].

Given this, the purpose is to predict f_* expected value given the test input x_*. Recalling that a Gaussian Process is a set of random variables which have a consistent Gaussian distribution (and the mean can be considered as 0), the problem can be represented as [10][56][57]:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \boldsymbol{K} + \sigma_n^2\boldsymbol{I} & \boldsymbol{K}_*^T \\ \boldsymbol{K}_* & \boldsymbol{K}_{**} \end{bmatrix}\right) \tag{38}$$

where T indicates matrix transposition.

Prediction consists in estimating the mean value and the variance of f_*. Considering equation (38), what is desired is the conditional probability $p(y_*|\boldsymbol{y})$: "given the observed outputs, how likely is a certain prediction for y_*?" Thus, the conditional distribution of $\boldsymbol{y}_*$ given $\boldsymbol{y}$ is given by:

$$\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{X}_* \sim N\left(\overline{f_*}, cov(f_*)\right) \tag{39}$$

where

$$\overline{\boldsymbol{f}_*} = \boldsymbol{K}_*(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y}$$

$$cov(\boldsymbol{f}_*) = \boldsymbol{K}_{**} - \boldsymbol{K}_*[\boldsymbol{K} + \sigma_n^2\boldsymbol{I}]^{-1}\boldsymbol{K}_*^{\boldsymbol{T}} \tag{40}$$

The mean value of the prediction f_* (also known as the matrix of regression coefficients) gives the best estimate value for y_*. The $cov(f_*)$ represents the uncertainty of the estimation [56][57].

4.1.3 Gaussian Process Classification

Gaussian processes can be applied to problems other than regression. Another popular application is classification, where the purpose is to assign an input pattern $\boldsymbol{X}$ to one of C classes, C_1, …, C_c. Classification problems can either be binary (two-class, C=2) or multi-class (C>2) [10]. This thesis will be focused on the binary problem. The aim of the classification problem is to output the correct class label for a new data point. Generalization to test cases involves some inherent level of uncertainty. To present this uncertainty over class labels, one may want a method that outputs probabilities over the different labels for each new data point [10].

The basic idea behind Gaussian process prediction is to convert the output of a regression model (the latent function $\boldsymbol{f}$, which can lay in the domain $(-\infty, \infty)$) into a class probability by means of a sigmoid function, squashing its argument into the range [0,1]. This guarantees a valid probabilistic interpretation. An example of a "squashed" function is presented in Figure 10. Summarizing, this can be done in two steps [10][56]:

1. Evaluate a "latent function" $\boldsymbol{f}$, responsible for modelling qualitatively how the likelihood of one class versus the other changes over the x axis. This is the GP.
2. Squash the output of this latent function onto [0,1] using a sigmoidal function, to obtain a prior on $\pi(x) \triangleq p(y = +1|f(\boldsymbol{x})) = \sigma\big(f(\boldsymbol{x})\big)$.

Two common choices for this response function are the logistic response function and the cumulative response function. The latter is known as probit regression and is represented as $\Phi(\boldsymbol{f})$. This S-shaped function maps high values of $\boldsymbol{f}$ into $\pi(\boldsymbol{f}) \approx 1$, and low values of $\boldsymbol{f}$ into $\pi(\boldsymbol{f}) \approx 0$ [56]. Probit regression was used in this thesis and the sigmoid function will be represented as $\boldsymbol{\Phi}$(f).

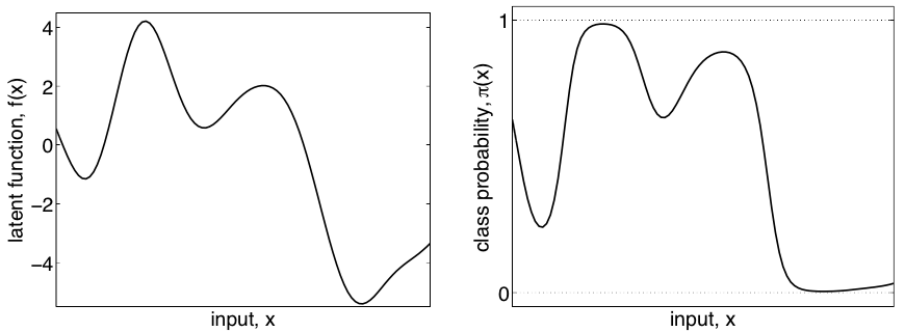

Figure 10- (a) Describes a sample latent function $f(x)$ drawn from a Gaussian process as a function of x. (b) Shows the result of squashing this sample function through the sigmoid function (in this example, a logistic was used) to obtain the class probability [10].

The latent function ($\boldsymbol{f}$) is also known as the nuisance function. This means that the values of $\boldsymbol{f}$ are not observed directly, only the inputs $\boldsymbol{X}$ and the class labels $\boldsymbol{y}$ are observed. The purpose of $\boldsymbol{f}$ is only to allow a convenient formulation of the model and after it will be integrated out. The interest will lay on the values of π, particularly on the value of the test case $\pi(x_*)$ [10].

The same notation from the regression problem will be used. Here, the latent function values are summarized by $\mathbf{f} = [f_1, \dots, f_n]^T$ with $f_i = f(x_i)$. Assuming that the latent Gaussian process is noise-free, and similar to the regression problem, the two equations can be written as:

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}_*^{\boldsymbol{T}} \\ \boldsymbol{K}_* & \boldsymbol{K}_{**} \end{bmatrix}\right) \tag{41}$$

When making predictions, and marginalizing the training set latent variables:

$$p(\boldsymbol{f}_*|\boldsymbol{X}_*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) = \int p(\boldsymbol{f}_*, \mathbf{f}|\boldsymbol{X}_*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta})\, d\mathbf{f} = \int p(\boldsymbol{f}_*|\mathbf{f}, \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{\theta})\, p(\mathbf{f}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) d\mathbf{f} \tag{42}$$

where the joint posterior is factored into the product of the posterior , the conditional prior is

$$p(\boldsymbol{f}_*|\mathbf{f}, \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}_*|\boldsymbol{K}_*\boldsymbol{K}^{-1}, \boldsymbol{K}_{**} - \boldsymbol{K}_*\boldsymbol{K}^{-1}\boldsymbol{K}_*^{\boldsymbol{T}}) \tag{43}$$

Finally the predictive class membership probability $p_* \coloneqq p(\boldsymbol{y}_* = +1|\boldsymbol{X}_*, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta})$ is obtained by averaging out the test set latent variables:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*,\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}) = \int p(\boldsymbol{y}_*|\boldsymbol{f}_*)\, p(\boldsymbol{f}_*|\boldsymbol{X}_*,\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}) d\boldsymbol{f}_* = \int \Phi(\boldsymbol{y}_*,\boldsymbol{f}_*) p(\boldsymbol{f}_*|\boldsymbol{X}_*,\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta}) d\boldsymbol{f}_* \tag{44}$$

where Φ is the cumulative Gaussian sigmoid function [10][56][58].

Then, considering the outputs **f** of a certain GP, how likely they are to be appropriate for the training data can be decomposed using Bayes' theorem, and the posterior is:

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})} \tag{45}$$

The first factor in the numerator, $p(\boldsymbol{y}|\boldsymbol{f})$, is the likelihood and is informed by the sigmoid function, $\Phi(\boldsymbol{f})$. Since the probability of the two classes must sum 1, $p(y=+1|\boldsymbol{f})$ is $\Phi(\boldsymbol{f})$ and , $p(y=-1|f)$ is $1-\Phi(\boldsymbol{f})$ [56].

Given the latent functions **f**, the observations are assumed to be independent, which gives rise to a factorial likelihood, factorizing over data points:

$$p(\boldsymbol{y}|\mathrm{f}) = \prod_{i=1}^{n} p(y_i|f_i) = \prod_{i=1}^{n} \Phi(y_i f_i) \tag{46}$$

and using the Gaussian prior, $\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\theta} \sim \mathcal{N}(0,\boldsymbol{K})$, for the latent function, the posterior becomes:

$$p(f|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})} = \frac{\mathcal{N}(\mathbf{f}|0,\boldsymbol{K})}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})} \prod_{i=1}^{n} \Phi(y_i f_i) \tag{47}$$

which is non-Gaussian [59]. A graphical representation of the binary Gaussian process classification is shown in Figure 11.

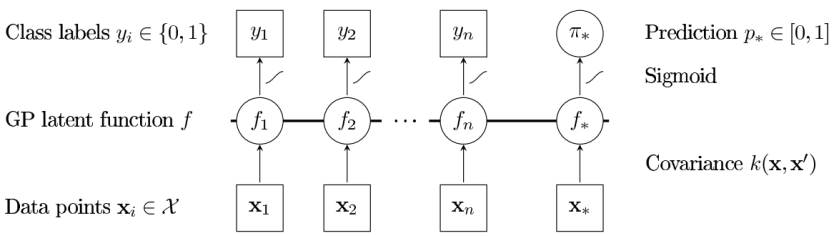

Figure 11 – Graphical representation for the binary Gaussian process classification where circles represent unknown quantities and squares observed variables. An observed label y_i is conditionally

independent of all other nodes given the corresponding latent variable f_i. Labels y_i and latent function values f_i are connected through the sigmoid likelihood: all latent functions values f_i are fully connected, since they are drawn from the same P. The labels y_i are binary, whereas the prediction p_* is a probability in the interval [0,1] [58].

The solution of classification problems using Gaussian processes is more computationally demanding than for the regression problems. In the regression problems, the likelihood function is assumed to be Gaussian. For classification models, where the targets are discrete class labels, the Gaussian likelihood is inappropriate and the exact inference is not feasible. However, even though the likelihood is non-Gaussian, the posterior process can be approximated by a GP [10].

There are two analytic approximations for approximating the non-Gaussian joint posterior with a Gaussian one. The first one, and more straightforward is the Laplace approximation method. The second one, and more sophisticated is the expectation propagation (EP) method [10].

In general, approximating the non-Gaussian posterior by a Gaussian is defined by:

$$p(\mathbf{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) \simeq q(\mathbf{f},\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m},\boldsymbol{A}) \tag{48}$$

where $\boldsymbol{A}^{-1} = \boldsymbol{K}^{-1} + \boldsymbol{W}$ and **W** denotes the precision of the effective likelihood. Therefore, the approximation methods will correspond to particular choices of **m** and **A** [58]. Assuming that a Gaussian approximation to the posterior with mean **m** and covariance **A** was found. Consequently, the latent distribution for a test point becomes a tractable one-dimension Gaussian with $q(\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{X}_*) = \mathcal{N}(\boldsymbol{f}_*|\mu_*,\sigma_*^2)$, where:

$$\begin{aligned} \mu_* &= \boldsymbol{K}_*\boldsymbol{K}^{-1}m \\ \sigma_*^2 &= \boldsymbol{K}_{**} - \boldsymbol{K}_*(\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{K}^{-1})\boldsymbol{K}_*^T \end{aligned} \tag{49}$$

Using this approximation, the probability of the test point belonging to the positive class is computed as [59]:

$$q(\boldsymbol{y}_* = +1|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{X}_*) = \int \Phi(\boldsymbol{f}_*)\,\mathcal{N}(\boldsymbol{f}_*|\mu_*,\sigma_*^2)d\boldsymbol{f}_* = \Phi\left(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\right) \tag{50}$$

In this thesis, the expectation propagation (EP) algorithm was used to find the approximate parameters m and **A**. The EP algorithm has been shown to have superior performance when compared to other approximation methods (Nickisch and Rasmussen, 2008). A detailed description of EP can be found in Rasmussen and Williams (2006, pp. 52-60) or in Nickisch and Rasmussen (2008) work [10][58].

Basically, the EP algorithm is an iterative method to find approximations based on approximate marginal moments, which can be applied to Gaussian

processes. The posterior is given by Bayes' rule, as the product of a normalization term, the prior and the likelihood [58][59]:

$$p(\mathbf{f}|\boldsymbol{X},\boldsymbol{y}) = \frac{1}{Z}p(\mathbf{f}|\boldsymbol{X})\prod_{i=1}^{n}p(y_i|f_i) \tag{51}$$

where the normalizing term is the marginal likelihood:

$$Z = p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\mathbf{f}|\boldsymbol{X})\prod_{i=1}^{n}p(y_i|f_i)d\mathbf{f} \tag{52}$$

The individual likelihood terms are replaced by site functions $t_i(f_i)$ being unnormalized Gaussians:

$$p(y_i|f_i) \approx t_i(f_i,\mu_i,\sigma_i^2,Z_i) \coloneqq Z_i\mathcal{N}(f_i|\mu_i,\sigma_i^2) \tag{53}$$

During the procedure, the three quantities μ_i, σ_i^2 and Z_i are iteratively optimized. Based on local approximations, the approximated posterior can be written as:

$$p(\mathbf{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta}) \approx \mathcal{N}(\mathbf{f}|m,\boldsymbol{A}) = \mathcal{N}(\mathbf{f}|m,(\boldsymbol{K}^{-1}+\boldsymbol{W})^{-1}) \tag{54}$$

where $\boldsymbol{W} = [\sigma_i^{-2}]_{ii}$ and

$$m = \boldsymbol{A}W\mu_i = [I-\boldsymbol{K}(\boldsymbol{K}+\boldsymbol{W}^{-1})]\boldsymbol{K}\boldsymbol{W}\mu_i\,, \qquad \mu_i = (\mu_1,\dots,\mu_n)^T \tag{55}$$

4.2 Classification performance

The output of the Gaussian process classifier is the probability of a certain test point belonging to the positive class +1, and it is given by equation (50). Thereby, it is required to find a threshold in order to decide when a probability belongs to class +1 or -1. The probabilities that exceed this threshold are assigned as +1 and the others to -1. This threshold takes into account the number of subjects belonging to each class in the training set. Therefore, if the number of subjects belonging to class +1 is equal to the number of subjects belonging to class -1, the threshold will be 0.5. The threshold can then be defined as:

$$threshold = \frac{\sum y = 1}{\sum y = 1 + \sum y = -1} \tag{56}$$

After this, the predicted labels are compared with the true labels and the accuracy of the classifier can be achieved (Figure 8). Ideally, the classifier would be trained on one dataset and test on a completely independent dataset. If only a single dataset is available for both training and test then cross-validation can be used to help overcome this problem. For cross-validation, some data are removed from the training set, and the model is trained with the remaining ones. Then, after the training, the removed or "left-out" data are used to test the learned model. This process is repeated for all the examples in turn. Finally, the accuracy of the prediction is computed for all the examples. There are several types of cross-validation. The "leave-one-out" cross-validation approach consists in removing one subject in each turn, making a prediction for each subject in the dataset [5]. Figure 12 shows a schematic representation of a cross-validation procedure.

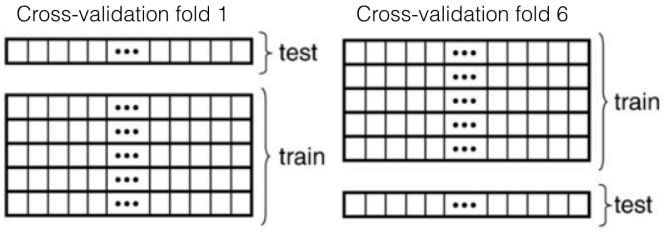

Figure 12 – Cross-validation procedure for 6 groups of examples. Each group takes a turn as the test set while the rest serves as the training set [5].

Testing a classifier involves determining if the features used contain information about class membership. If the classifier truly captured the relationship between features and classes, it will be able to predict the classes of examples it has not seen before. The most common way to measure how well a classifier predicts the label for the class test is its accuracy. The accuracy defines the percentage of predictions that are correct, and is given by:

$$acc = \frac{sensitivity + specificity}{2} = \frac{|TP| + |TN|}{n} \tag{57}$$

where TP is the number of true positives, TN is the number of true negatives and n the number of subjects. Sensitivity, or true positive rate, measures the proportion of actual positives, which are correctly identified. Specificity, or true negative rate, measures the proportion of negatives that are correctly identified as such. Sensitivity and specificity are computed as follows:

$$se = \frac{|TP|}{|TP| + |FN|} \qquad sp = \frac{|TN|}{|TN| + |FP|} \tag{58}$$

where FN is the number of false negatives and FP the number of false positives [5].

Other widely used method to evaluate the performance of a classifier is the area under the receiver operating characteristic (ROC) curve, from which the area under the curve (AUC) can be extracted. This method avoids the potential subjectivity in the threshold selection process, by summarizing overall performance over all possible thresholds. The ROC curve is a plot of the test true-positive rate (y-axis) against the corresponding false-positive rate (x-axis), e.g. the ROC curves plots sensitivity as a function of commission error (1-specificity). The curve is computed group test performance at different "diagnostic thresholds" [60].

The AUC score provides a measure of discrimination, that is, the ability of the test to correctly classify those with and without a certain disease, across all possible thresholds. A perfect test would have 100% sensitivity with zero false positives (100/ specificity), across all thresholds. No discrimination would result in AUC=0.5. A real world scenario will be in the between [60][61].

ROC AUC is a single metric facilitating comparisons between tests. However, AUC lacks clinical interpretability. Clinicians are not interested in performance across all thresholds; they are rather focus on clinically relevant thresholds. AUC will include clinically relevant but also clinically illogical thresholds [61]. Therefore, in order to evaluate the performance of the classifier, both accuracy and AUC should be reported.

4.3 Evaluating results

4.3.1 Comparing if two groups are homogeneous in age and gender

When comparing two groups of patients with the purpose of predicting a variable of interest (e.g. the status of a patient), the variability that is not related with the status of the patient must be reduced. Therefore, when picking subjects to train the model, homogeneity in age and gender is desired.

Age is a continuous variable, which can assume infinite range of values. In order to infer if two groups are homogeneous in age, a two-sample t-test may be used. In this test the means from the two different patient groups are compared [62]:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{rp}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{59}$$

where s_{rp} is the pooled standard deviation for both groups, $\bar{x}_1$ and $\bar{x}_2$ are the mean ages for each group respectively.

Gender is a categorical variable, which can take one of two fixed values (e.g. 1 for female and 0 for male) assigning each individual to a particular group or category. In this case, a chi-square test is used to examine the differences in gender between the

two groups. Therefore, the Chi-square test will analyse the frequency of each category in each group, and the frequencies can be presented in a frequency table (Table 1).

Table 1 - Frequency table representing the frequencies in gender for group 1 and group 2.

	Group 1	Group 2	Total
Female	12	15	27
Male	18	15	33
Total	30	30	60

The formula for calculating the chi-square is:

$$\chi^2 = \frac{(observed - expected)^2}{expected} \tag{60}$$

The value of the chi squared is found by determining whether the observed frequencies differ significantly from the expected frequencies. To find the expected frequencies, the independence of the rows and columns is assumed. For instance, the expected frequency for females in Group 1 is given by: multiplying the row total (27) with the column total (30) and dividing by the overall total (60).

With the Chi-square value and the degrees of freedom is possible to obtain a p-value. If the p-value is less than 0.05, then the two groups are significantly different in gender.

4.3.2 Permutation test

After evaluating the classifier performance the accuracy and AUC need to be tested for significance. A statistically significant classification result implies the null hypothesis to be rejected. The null hypothesis claims that there is no information about the variable of interest in the data from which is being predicted. Establishing statistical significance is usually done by determining how improbable the observed result would be if the null hypothesis were true. This probability is also known as p-value.

Therefore, this is done by determining how likely this accuracy could be observed if the classifier was operating at chance (i.e. if the null hypothesis were satisfied). Permutation-based p-value is a method widely used to measure the accuracy when the classifier is operating at chance. The permutation test is a non-parametric technic, and can be of great value when the distribution of the data is unknown. The permutation test operates as follows:

1. The training labels are permuted

2. The classifier is trained with the permuted labels
3. The classifier is tested with the true test set
4. This procedure is repeated many times (normally, 1000 times) with a different shuffling each time.
5. Over many repetitions, this yields a sample of accuracy results under the null hypothesis.

The p-value will represent the fraction of random data sets under a certain null hypothesis where the classifier behaved as well as or better than in the original data [5][63]. A significant classifier will have a small p-value, rejecting the null hypothesis and suggesting that the accuracy of the classifier is significant.

The permutation test is often used to test the classifier significance. However, it can be used to compare the performance of two different classifiers using a paired permutation test. Assuming two different models, M_1 with accuracy a and M_2 with accuracy b, the null hypothesis is that model M_1 and model M_2 are identical. This test can also be applied to compare area under the curve measures.

Assuming 1000 permutations, the same permuted training labels are used in each permutation for both models. Therefore, this results in 1000 pairs of accuracy estimates $\{(a_1, b_1), (a_2, b_2), \dots (a_{1000}, b_{1000})\}$ for M_1 and M_2. Here, (a_1, b_1) correspond to the accuracy of the true label. Under the null hypothesis, any permutation of the labels is an equally likely output.

The difference between M_1 and M_2 accuracies is measured for each permutation, $\{(|a_1 - b_1|), (|a_2 - b_2|), \dots (|a_{1000} - b_{1000}|)\}$. Thus, the number of times a difference between the accuracies was equal or greater than the difference using the true labels, divided by 1000 will give the two-side p-value or achieved significance level for the null hypothesis [64]. Considering that Model M_1 as an accuracy of 0.775 and M_2 an accuracy equal to 0.450, the true difference between accuracies will be 0.325. Figure 13 shows an example of a distribution of 1000 differences in accuracies between random permutations of model M_1 and model M_2. This difference is significant and the p value is 0.0130.

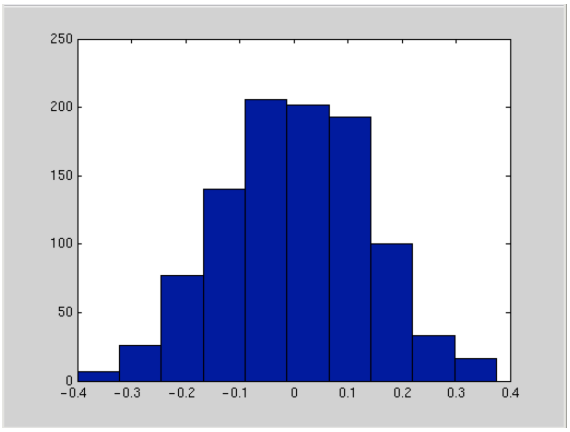

Figure 13 - Distribution of 1000 differences in accuracies between random permutations of model M1 (acc=0.7750) and Model M2 (acc=0,450). (Image obtained in 2014a®.

4.3.3 Mc Nemar's Test

Other possible solution for comparing the accuracy of two classifiers is Mc Nemar's test. Mc Nemar's test is a variant of χ^2 test and is a non-parametric test used to analyze matched pairs of data. According to Mc Nemar's test, two algorithms can have 4 possible outcomes arranged in a 2×2 contingency table:

Table 2 - Contingency table with the possible results of two algorithms.

	Algorithm M_1 failed	Algorithm M_1 failed
Algorithm M_2 failed	N_{ff}	N_{sf}
Algorithm M_2 succeeded	N_{fs}	N_{ss}

where N_{ff} denotes the number of times when both algorithms failed and N_{ss} denotes the success for both algorithms. The other two parameters, $N_{sf}N_{fs}$, show cases where one of the algorithms failed and the other succeeded indicating the performance discrepancies.

Under the null hypothesis, the two algorithms should have the same error rate, which means that $N_{sf} = N_{fs}$. The Mc Nemar's test is distributed approximately as χ^2 with 1 degree of freedom:

$$\chi^2 = \frac{\left(\left|N_{sf} - N_{fs}\right| - 1\right)^2}{N_{sf} + N_{fs}} \tag{61}$$

The null hypothesis is correct when the probability that this quantity is greater than $\chi^2_{1,0.95} = 3.841459$ is less than 0.05 [65][66].

4.4 Weight Maps

One of the most important goals in neuroimaging is the interpretability for neuroscience and clinical use. Therefore, basic neuroscience research is often concerned with determining the brain regions, frequencies, or time intervals reflecting a certain cognitive process [67]. Unlike univariate methods, described earlier in chapter (…), machine learning based analysis does not naturally provide statistical test (and corresponding p-values) associated with every voxel/region of an image. Rather, these models are evaluated as "black boxes" and provide an overall estimate of the separability between two groups or conditions [68].

A natural approach for addressing this issue is based on finding the voxels that contribute most strongly and reliably to the classifier's success. During the

classification, the model generates weights for each voxel and the combination of all weights defines the model. The multivariate machine learning models predictions are based on the whole pattern and the weights at each voxel are thus dependent on one another and therefore, no direct localization inferences or voxel-wise statistical test assuming independence can be performed on them. However, the weights of each voxel can be used in order to build multivariate maps, which represent the behaviour of the underlying classifier and describe which brain regions carry discriminative information [68].

The weight vector, $\hat{\boldsymbol{w}}$, can be understood as providing a spatial representation of the decision boundary. Figure 14 provides a toy example and a geometric interpretation of $\hat{\boldsymbol{w}}$ [69].

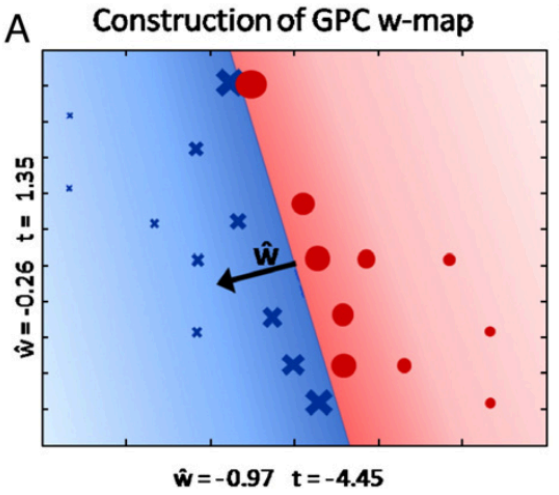

Figure 14 – In a GPC w-mapping, samples closer to the decision boundary carry higher weight and the weight vector ($\hat{w}$) is orthogonal to the decision boundary [69].

5.1 Materials and Methods

5.1.1 Subjects/ Database

The first step of the project was to illustrate the extent of scanner variability effect. Two different datasets comprising only healthy subjects were used for this purpose. These subjects were scanned using a repeated measures design whereby a single participant is scanned using different coils therefore the effect of intra-subject variability was negligible.

The first dataset included 14 healthy subjects (9 women and 5 men) scanned twice with a different coil. The MR brain scans were acquired on a 3T GE Medical Systems, Signa HDx. The first scan was acquired with a Quad Coil and the second scan with an 8-channel coil. The second dataset included MRI brain scans from 41 young healthy subjects (19 women and 22 men) and was acquired on a 3.0T GE scanner. Here, patients were scanned twice with different acquisition parameters. Parameters used in both acquisitions are summarized on Table 3.

Table 3 - Scan parameters for the 41 subjects dataset.

Parameters	Scanner 1	Scanner 2
Coil	8-channel	8-channel
TR (ms)	650	400
TE (ms)	Min full	Min full
FA	8º	11º
Slice thickness (mm)	1.2	12.3
Resolution (mm)	256x256	256x256
FOV	26	26

For the second part of the study, participants from the ADNI project database were used with the aim of studying the impact of scanner and subject variability in automated classification in different clinical scenarios. The ADNI project is a multisite study comprising three different scanner models with field strengths of 1.5T and 3T. Data from the same model and same strength field were considered from the same scanner (even if the site was different) since the ADNI protocols on each scanner type were adjusted such that all sites report comparable results. Therefore, the subjects' scans were chosen considering the vendor and the field of strength.

The dataset included a total of 413 subjects (145 healthy, 148 MCI and 120 AD) and 6 scanners. The whole dataset was used for the VBM analysis but only 4 scanners were used for the machine learning based analysis. Subjects were selected by

taking into account the distribution of the diagnostic conditions and scanners for the machine learning based analysis, in order not to add a potential confound. Therefore, the diagnostic conditions were equally distributed, with approximately 30 scans per diagnostic group and scanner. The 3T scanners, GE and Philips, were excluded from the machine learning analysis since there were not enough subjects scanned. Table 4 summarizes the number of subjects per scan and diagnostic group with mean age and gender distribution.

Table 4 - Mean age and gender distribution across sites. M=Male F=Female.

	NL		MCI		AD	
Mean	**Age**	**Gender**	**Age**	**Gender**	**Age**	**Gender**
1.5 GE	75.92	M=18 F=12	74.64	M=20 F=10	76.16	M=15 F=15
3 GE	74.78	M=6 F=6	80.19	M=4 F=3	75.08	M=0 F=4
1.5 SIEMENS	75.50	M=12 F=18	73.74	M=14 F=16	78.09	M=10 F=20
3 SIEMENS	75.59	M=13 F=17	75.25	M=18 F=12	75.65	M=8 F=17
1.5 Philips	75.24	M=21 F=5	74.64	M=23 F=7	72.64	M=12 F=8
3 Philips	76.44	M=8 F=9	71.69	M=14 F=7	74.06	M=6 F=5
TOTAL	75.62	M=78 F=67	74.43	M=93 F=55	75.72	M=51 F=69

For the machine learning analysis, the 1.5T scanners used were General Electric (GE) Healthcare (1.5GE), SIEMENS Medical Solutions and Philips Medical Systems (1.5SIEMENS) and the only 3T scanner chosen was SIEMENS Medical Solutions (3SIEMENS). The T1-weighted brain images selected were obtained using the 3D anatomical magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence and the parameters were slightly different in each scanner.

5.1.2 SPM Analyses

Images were processed with SPM8, running on Matlab 2014a® (The Math-Works, Inc., Natick, MA), in order to perform the VBM analysis. This included realignment of all scans, spatial normalisation, segmentation and smoothing using the steps explained before, on Chapter 3. For the ADNI dataset, a study-specific template including 30 subjects (10 healthy, 10 MCI and 10 AD) from scanner 1.5SIEMENS was obtained and used on the DARTEL algorithm. Only data from one scanner were used in order to not add scanner variability effect into normalisation. Normalisation, using the study-specific template and modulation were then performed. The modulated gray matter segments were then smoothed with an 8mm full width at half maximum (FWHM) three-dimensional Gaussian kernel to ensure the local normality of the data.

The spatially normalized and smoothed GM maps were used in a mass univariate analysis on each voxel of the GM mask. For the first two datasets, comprising two scans from each healthy subject, a standard parametric statistical

paired t-test between scans of each subject was performed. For the ADNI project dataset, a factorial design was specified in order to combine the disease condition and different scanners. The factorial design allows the study of the interaction between the disease condition and scanner variability. The diagnostic group condition was defined as one factor with 3 levels (controls, MCI and AD) and the scanners as one factor with 6 levels (6 different scanners). Statistical parametric maps of F-test were computed to assess the main effects of disease and scanner. The specifications for both models are specified on table 5.

Table 5 – Specifications of the General Linear Model for the Same Subject dataset and for the ADNI project dataset.

	Same Subject dataset	**ADNI dataset**
Design	Paired t-test	Full Factorial
Scans	smwc1.nii	smwc1.nii
Intercept	Omitted	-
Covariates	NA	Age,Sex and TIV
Masking	Implicit absolute masking with threshold = 0.4	None
Global Calculation	Yes, using TIV as a global	Omit
Global Normalisation	Propotional	none

For the VBM, the "global normalisation" is about dealing with brains of different sizes and bigger brains are likely to have larger structures. Since VBM is a voxel-wise assessment of volumetric differences, it is useful to consider how the regional volumes are likely to vary as a function of whole brain volume. Computing globals can be done in a number of ways. One option consists of using the Total Intracranial Volume.

The presence of widespread "global" decreases in GM in case of Alzheimer's is well established. When interested in an effect that means a significant decrease in gray matter, computing the TIV as a global can lead to misleading results. Therefore, for the Factorial model, TIV was inserted as a covariate and not a global [70].

5.1.3 Machine learning

5.1.3.1 Classification

As stated before, a classification algorithm based on Gaussian process probabilistic classification was used. GM concentrations of each voxel from the normalized and modulated GM maps were used as input for classification. The matrix of GM concentrations is defined as **X**.

The Gaussian process classification is described in chapter 4. A GP is determined by its mean and covariance functions. A linear covariance function and a constant mean function, both with one hyperparameter each initialized at zero, were

used. The likelihood function specifies the probability of the observations given the latent function $\boldsymbol{f}$ (i.e. the GP and the hyperparameters). The likelihood function used was the error function or cumulative Gaussian likelihood and is normally used for probit regression. The inference method specifies how to infer the posterior process (by an approximation), evaluate the log marginal likelihood and make predictions. The expectation propagation approximation to the posterior GP was used.

Before classification, data were centred voxelwise in the feature space, meaning that the origin of the feature space was moved to the centre of mass of the training samples. Because the dimensionality of these data is very high, it is not computationally feasible to work in the original voxel space. Fortunately, all information necessary to train GP models is contained in the kernel matrix and targets. The kernel matrix is $N \times$N and, for a linear covariance function, it only needs to be computed once and can be computed by: $\boldsymbol{Q}_{raw} = \boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$. The transformed kernel matrix contains all the information required.

5.1.3.2 Global compensation of confounding effects

GM density depends on disease status. As stated before, one of the challenges in pattern recognition is that an algorithm must be robust and handle data that is corrupted by noise. There are several factors unrelated to the disease that could influence the measured tissue properties. The purpose of this study is to investigate the impact of these factors, aiming to design more robust machine learning algorithms, which can handle the scanner and subject variability

The principle of compensating for non-disease specific factors consisted of (1) learning a model that estimates the GM density values for every voxel based on examples of confound covariates and corresponding GM density maps, (2) applying the model to new data to obtain a confound template, and (3) subtracting the template from the observed GM map [8].

The Gaussian process regression method proposed by Ahmed Abdulkadir *et al.* [8] as well as the Ridge Ordinary Least Squares regression were used as correction methods. The goal of regression was to estimate a map of GM density based on the characteristic of each subject on a design matrix. Age, TIV, sex and a scanner specific constant term were used as covariates. In the design matrix $\boldsymbol{X_d} = [\boldsymbol{X_{d_{intercept}}}, \boldsymbol{X_{d_{age}}}, \boldsymbol{X_{d_{TIV}}}, \boldsymbol{X_{d_{gender}}}, \boldsymbol{X_{d_{scanner}}}]$, the design variables form the columns and the observations form the rows. The variables age and TIV were scaled to the interval [0,1] and sex and site scanner (S1, … , S4) were coded as {0,1} multi-level categorical variables. Data were centred by including a vector of ones: $\boldsymbol{X_{d_{intercept}}}$. Only data from controls were used in other to avoid confounding disease with non-disease effects during the estimation of the regression parameters.

Ridge Ordinary Least Squares

Following the basic linear regression presented on equation (1), the general linear model for voxel *v* in matrix form is:

$$\boldsymbol{x}^{(v)} = \boldsymbol{X_d}\boldsymbol{\beta}^{(v)} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon}\sim\mathcal{N}(0,\sigma^2) \tag{62}$$

where $x^{(v)} = \left[x^{(1,v)}, x^{(2,v)}, \dots, x^{(N,v)}\right]^T$ is the vector of observed GM concentrations at voxel *v*, ε is the residual error with zero mean and variance σ^2, and $\boldsymbol{\beta}^{(v)} = \left[\beta_{intercept}^{(v)}, \beta_{age}^{(v)}, \beta_{TIV}^{(v)}, \beta_{gender}^{(v)}, \beta_{scanner}^{(v)}\right]$ represents the parameters associated to design elements for voxel *v*. The residual data of a test example is then computed using the pseudo-inverse (equation 11) as follows:

$$\widehat{\boldsymbol{x}}_{OLS}^{(*,v)} = \boldsymbol{x}^{(*,v)} - \boldsymbol{x_d}^*\widehat{\boldsymbol{\beta}}^{(v)} = x^{(*,v)} - \boldsymbol{x_d}^{(*)}(\boldsymbol{X_d}^T\boldsymbol{X_d} + h^2\boldsymbol{I})^{-1}\boldsymbol{X_d}^T\boldsymbol{X} \tag{63}$$

where $\boldsymbol{x}^{(*,v)}$ and $\widehat{\boldsymbol{x}}_{OLS}^{(*,v)}$ are the original and corrected data vectors of the test example at voxel *v*, respectively [8]. In order to prevent multicolionearity, the parameters in the design matrix were orthogonalized by using Gram-Schmidt orthogonalization.

The regularization parameter, *h*, was estimated for each scenario. Normally, nested cross-validation is used to select attributes or parameters for the learning process. Nested cross-validation consists in splitting the training data in folds, one to train and the other to test the model with different parameters. The best set of parameters is chosen. The testing data is not considered for the nested cross-validation in order to prevent overfitting.

However, for this study different clinical scenarios were considered where testing data were different from training data (e.g, testing data were acquired in a different scanner that training data). Therefore, if we only used training data, the scanner involved in the testing set would not be considered for the parameter estimation, and correction of scanners would not be possible. Hence, the regularisation parameter *h*, was estimated using training and testing data from one example in each clinical scenario and the same parameter was used for all the examples on that clinical scenario. In this way, overfitting was only present in the first example.

Gaussian Process Regression

The conditional probability density function of the Gaussian process is entirely specified by its mean function $m(\boldsymbol{x_d})$ and covariance function $\boldsymbol{k}_{\theta,\alpha(\boldsymbol{x_d}.\boldsymbol{x_d}^*)}$. The zero mean function and the squared exponential covariance function were considered. Given the covariance function and its parameters and the zero mean function, the Gaussian process regression correction of the input dataset is:

$$\widehat{\boldsymbol{x}}_{GPR}^{(*,v)} = \boldsymbol{x}^{(*,v)} - \left(\boldsymbol{k}_{\theta,\sigma}^{(*)}\right)^T\boldsymbol{K}_{\theta,\sigma}^{-1}\boldsymbol{x}^{(v)} \tag{64}$$

where $\widehat{\boldsymbol{x}}_{GPR}^{(*,v)}$ and $\boldsymbol{x}^{(*,v)}$ are the corrected and original data matrices respectively of test example for voxel v.

The vector $\boldsymbol{k}_{\theta,\sigma}^{(*)} = \left[k_{\theta,\sigma}(\boldsymbol{x_d}^{(*)}, \boldsymbol{x_d}^{(1)}), k_{\theta,\sigma}^{(*)}(\boldsymbol{x_d}^{(*)}, \boldsymbol{x_d}^{(2)}), \ldots, k_{\theta,\sigma}(\boldsymbol{x_d}^{(*)}, \boldsymbol{x_d}^{(N)})\right]^T$ is a column vector of covariance function evaluations of the test example $x^{(*)}$; $\left[\boldsymbol{K}_{\theta,\sigma}^{(*)}\right]_{ij} = \boldsymbol{k}_{\theta,\sigma}(\boldsymbol{x_d}^{(i)}, \boldsymbol{x_d}^{(j)}), i, j \in \{1, \ldots, N\}$ is the covariance kernel matrix of the training examples. The covariance function of two input patterns $\boldsymbol{x_d}^{(i)}$and $\boldsymbol{x_d}^{(j)}$ was set as:

$$k_{\theta,\sigma}(\boldsymbol{x_d}^{(i)}, \boldsymbol{x_d}^{(j)}) = \theta_1^2 \exp\left(-\theta_2^2 \boldsymbol{x_d}^{(i)} - \boldsymbol{x_d}^{(j)^2}\right) + \theta_3^2 + \theta_4^2 (\boldsymbol{x_d}^{(i)})^T \boldsymbol{x_d}^{(j)} + \sigma^2 \delta_{ij} \tag{65}$$

where $\theta_k, k = \{1, \ldots, 4\}$ and σ are scalar model hyperparameters. δ_{ij} is the delta function and is one if $i = j$, and zero otherwise. The σ^2 is a high noise term, also known as bias term, and regularizes the solution preventing overfitting, while $\theta_1 \ldots, \theta_4$ determine the covariance characteristics, such as linear, non-linear and constant terms. The hyperparameters were all initialized at one and quasi-Newton optimization was performed for 50 iterations or until the relative change of the function value was smaller than $1e-6$. Then optimization was further continued using the trust-region algorithm as provided by Matlab's *fminunc* function and the Hessian was computed analytically [8].

Performing ordinary least-squares regression or Gaussian Process Regression on the evaluated kernel matrix is equivalent to first mapping the inputs to feature space, performing linear regression and then computing the pair-wise dot-product matrix. Considering a kernel function $q: x \times x \to \mathbb{R}, \left(x^{(i)}, x^{(j)}\right) \mapsto q\left(x^{(i)} - x^{(j)}\right) = \langle \Phi\left(x^{(i)}\right), \Phi\left(x^{(j)}\right) \rangle$, this function returns a number characterizing similarity between two input patterns x_i and x_j where $\Phi: x \times H, x \longmapsto h \coloneqq \Phi(x)$ is a map from input space x to some feature space H. The corrected kernel matrix can be computed directly in terms of the original kernel matrix **Q** and a matrix **R**:

$$\begin{aligned}
\boldsymbol{Q_{corr}} &= \langle \boldsymbol{\Phi}(\mathbf{X}) - \boldsymbol{X_d}\widehat{\boldsymbol{\beta}}, \boldsymbol{\Phi}(\mathbf{X}) - \boldsymbol{X_d}\widehat{\boldsymbol{\beta}} \rangle \\
&= \langle \boldsymbol{\Phi}(\mathbf{X}) - \boldsymbol{X_d}(\boldsymbol{X_d}^T \boldsymbol{X_d})^{-1} \boldsymbol{X_d}^T \boldsymbol{\Phi}(\mathbf{X}), \boldsymbol{\Phi}(\mathbf{X}) - \boldsymbol{X_d}(\boldsymbol{X_d}^T \boldsymbol{X_d})^{-1} \boldsymbol{X_d}^T \boldsymbol{\Phi}(\mathbf{X}) \rangle \\
&= \langle \boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{I} - \boldsymbol{X_d}(\boldsymbol{X_d}^T \boldsymbol{X_d})^{-1} \boldsymbol{X_d}^T), \boldsymbol{\Phi}(\mathbf{X})(\boldsymbol{I} - \boldsymbol{X_d}(\boldsymbol{X_d}^T \boldsymbol{X_d})^{-1} \boldsymbol{X_d}^T) \rangle \\
&= \langle \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{R}, \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{R} \rangle \\
&= \boldsymbol{R} \langle \boldsymbol{\Phi}(\mathbf{X}), \boldsymbol{\Phi}(\mathbf{X}) \rangle \boldsymbol{R}^T = \boldsymbol{R}\boldsymbol{Q}\boldsymbol{R}^T
\end{aligned} \tag{66}$$

Depending on the regression method, the matrix R differs:

$$\boldsymbol{R}_{OLS} = \boldsymbol{I} - \boldsymbol{X_d}(\boldsymbol{X_d}^T \boldsymbol{X_d})^{-1} \boldsymbol{X_d}^T \tag{67}$$

$$\boldsymbol{R}_{GPR} = \boldsymbol{I} - \boldsymbol{K}^T (\mathbf{K} + \alpha^{-1} \boldsymbol{I})^{-1}, \tag{68}$$

where K is given by equation (65) [8].

5.1.3.3 Scenarios of clinical diagnosis

The same clinical scenarios formulated by Ahmed Abdulkadir *et al*. were used in this study. The purpose was to investigate if, using a highly standardized and controlled multi-site study like ADNI project database, the same findings in Abdulkadir study could be replicated, evaluating how well the classification pipeline predicted the disease status using the ADNI dataset. For each scenario, the two correction methods were applied and permutation and McNemar tests were used to infer if there were any significant improvements after correction. Four clinical scenarios are described as follows:

Single site training: On-site application

In this scenario data from controls and the respective patients group were used separately for each of the four sites. This avoids any potential issues with heterogeneous hardware and/or protocol. Performance was measured with cross-validation. For this scenario, correction was applied only taking into account subject covariates, such as age, sex and TIV.

Single site training: Transfer to other site

In this scenario, the classifier was trained with data from controls and respective patients group from one scanner and it was further tested with data from another scanner (that was not involved in the training stage). Correction was applied taking into account scanner, age, sex and TIV covariates.

In order to correct for scanner variability, data from healthy subjects from both scanners were needed for the hyperparameter estimation in regression. However, subjects from the scanner involved in the testing phase cannot be used for estimating hyperparameters in regression, since labels have to be assumed as unknown. Therefore, in order to perform correction in this case, a cross-validation procedure was applied to the testing subjects, where one subject was evaluated during the testing phase and the other subjects were used for hyperparameter estimation for correction. In this way, the fold with the test example was excluded from the hyperparameter estimation.

Group-specific sites for training: On-site application

In this scenario two scanners were involved in the training phase, providing data from only one diagnostic group. In the testing phase, data from the same two scanners were used, but with reversed diagnostic groups. For example, controls from 1.5GE and MCI from 1.5Philips for training and MCI from 1.5GE and controls from 1.5Philips for testing. As stated in the previous scenario, only data from controls were used for correction and testing data cannot be used for correction. In order to correct for scanner variability, the controls from the scanner used in the testing phase were needed. Therefore, a cross-validation procedure was applied in order to split the

control subjects data for correcting and for testing, ensuring that the controls used for correction were not used for testing.

Group-specific sites for training: Transfer to another site

As in the previous setting, the training set in this scenario was composed of controls from one scanner and the disease group from another scanner. However the test data came from a scanner not used in the training set. For example: controls from 1.5GE and AD from 1.5Philips for training and controls and AD from 3SIEMENS for testing. In this case, the controls from the scanner that provides the disease group for training were also used for correction and a cross-validation procedure was used to take some controls for correction and the others for testing.

6.1 First Dataset: Comparing two scans from the same subject

6.1.1 Voxel Based Morphometry - Univariate Analysis

For all the VBM results, after defining the contrasts, a gray mask, a voxel threshold equal to 0 and a p-value equal to 0.05 after correction for multiple comparisons using family-wise error (FEW) correction were used. The outcome is a map of p-values (Figure 15 and Figure 16) with statistically differences between each scan.

Comparing two coils

For this data a paired sample t-test was done between the two scans of each subject. The effect of scanner was visualized by using a F-contrast. The results are presented in Figure 15.

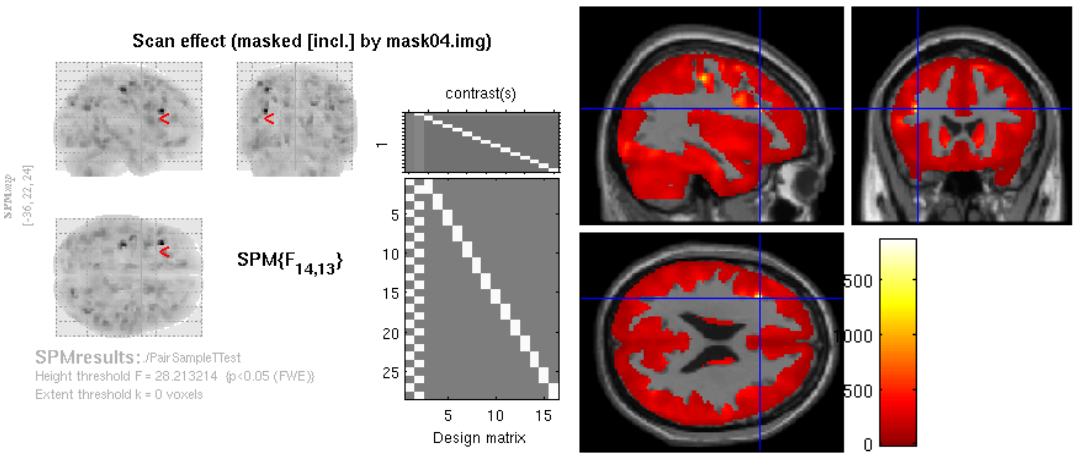

Figure 15 – VBM analysis of effect of scanner (FEW $\rho < 0.05$). Left: Brain Glass and Design matrix. Right: Effect of scanner-variability using a Paired sample t-test between the scans and a F-contrast. Color bars indicate F-scores for each contrast.

Figure 15 shows that the scanner variability effect is dispersed all over the gray matter with strong activation in localized regions.

Comparing scans with different parameters

As stated before, a paired sample t-test was done between the two scans of each subject. To access the effect of scanner variability a F-contrast was also used.

Results can be visualized in Figure 16. The scanner variability effect has a strong activation dispersed all over the gray matter.

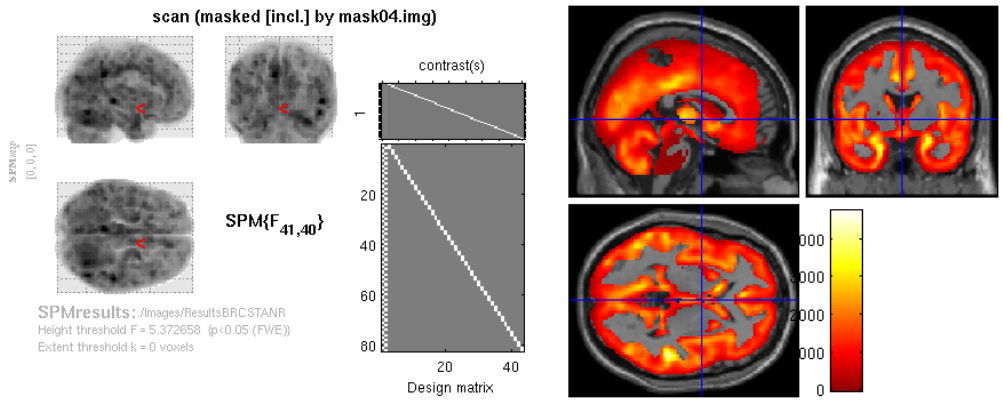

Figure 16 - VBM analysis of effect of scanner (FEW $\rho < 0.05$). Left: Brain Glass and Design matrix. Right: Effect of scanner-variability using a Paired sample t-test between the scans and a F-contrast. Color bars indicate F-scores for each contrast.

6.1.2 Classification – Multivariate Analysis

Classification of scanner was conducted in order to demonstrate how well the scanner was distinguishable, using two scans from the same subject.

Comparing two coils

The Gaussian process classification was used to access the effect of scanner-variability on machine learning based analysis. Thus, the label was defined as 1 for scans from the Quad coil and -1 for scans from the 8-channel coil. Permutation tests were used to infer the significance of the results.

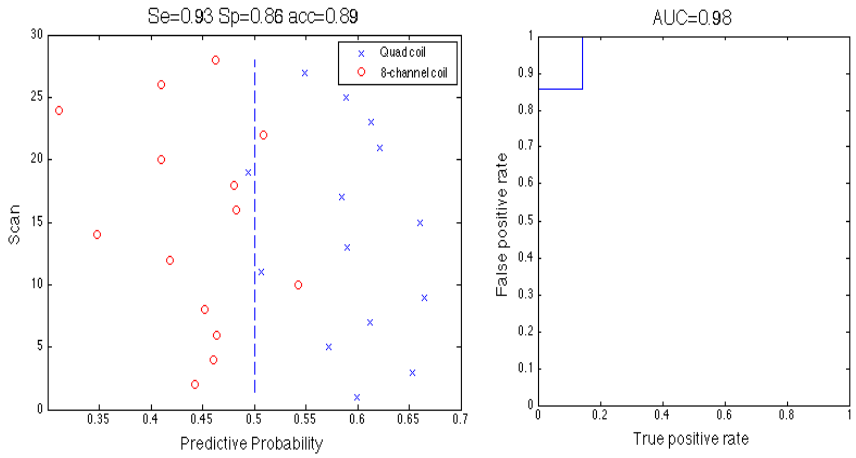

Figure 17 – (Left) Classification accuracies for GPC predictors for classifying scans into either the quad coil group (assigned as 1) or 8-channel coil group (assigned as -1). (Right) Receiver operating characteristic (ROC) curve.

The accuracy (acc) for this classification was 0.89 and the area under the curve (AUC) was 0.98 and both were significant. GPC predictions for classifying scans are presented in Figure 17. The weight maps of the classification of Quad coil scan versus 8-channel coil scan are presented in Figure 18.

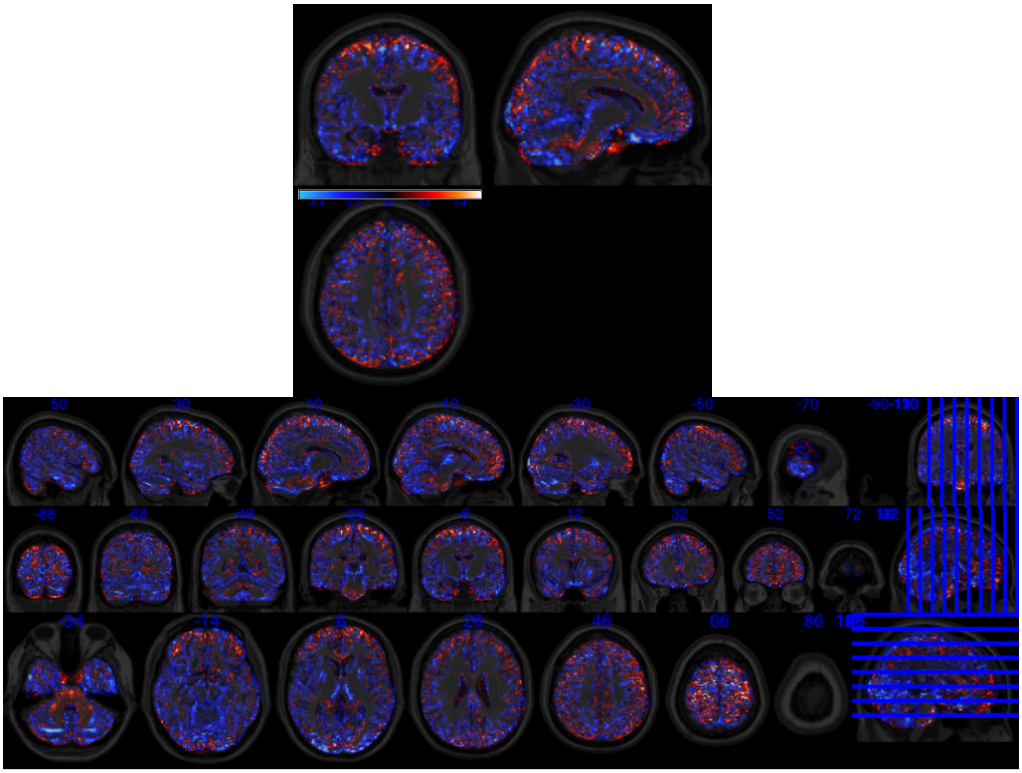

Figure 18 – Multivariate discrimination weight map. Unthresholded GPC weights overlaid on an anatomical template. The color code shows the relative weight on each voxel for the decision boundary (red/yellows scales: higher weights for the quad coil and blue scales: higher weights for the 8-channel coil).

Comparing scans with different parameters

For this dataset, the procedure was similar to the previous one. For testing the scanner-variability effect, the labels were defined as 1 for Scanner 1 and -1 for Scanner 2.

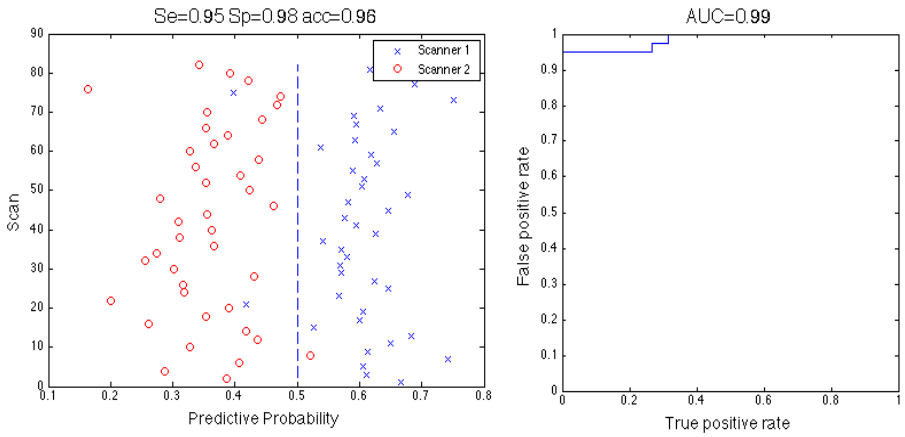

Figure 19 - (Left) Classification accuracies for GPC predictors for classifying scans into either the Scan 1 group (assigned as 1) or Scan 2 group (assigned as -1). (Right) Receiver operating characteristic (ROC) curve.

The accuracy (acc) for this classification was 0.96 and the area under the curve (AUC) was 0.99 and both were significant. GPC predictions for classifying scans are presented in Figure 19. The multivariate discrimination weight maps of the classification of Scanner 1 versus Scanner 2 are presented in Figure 20.

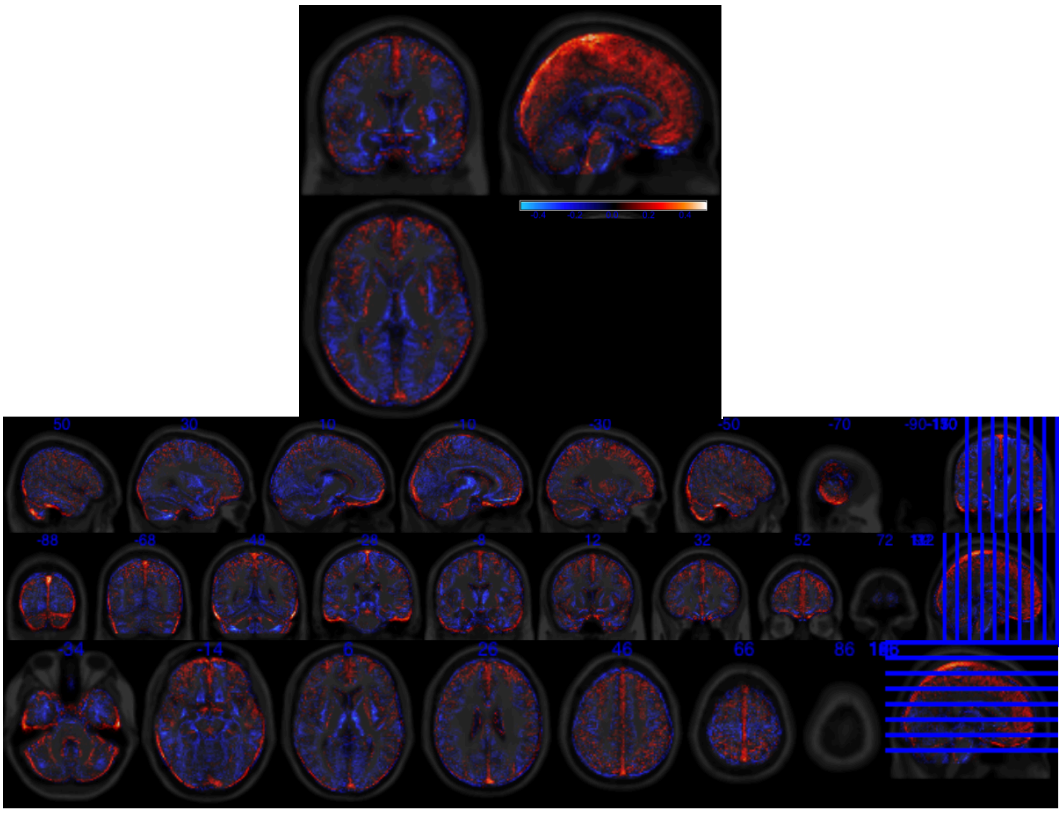

Figure 20 - Multivariate discrimination weight map. Unthresholded GPC weights overlaid on an anatomical template. The color code shows the relative weight on each voxel for the decision boundary (red/yellows scales: higher weights for the Scanner 1 parameters and blue scales: higher weights for the Scanner 2 parameters).

6.2 Second dataset: ADNI Project database

When picking subjects to train the model, homogeneity in age and gender is desired in order to reduce the variability that is not related with the status of the patient.

Therefore, a two-sample t-test was used to compare each combination of diagnostic group and scanner, inferring if two groups were homogeneous in age. The two-sample t-test was performed using a function in Matlab 2014a®. Almost all the combinations of groups were homogeneous in age, except AD patients from 1.5SIEMENS with AD patients from 1.5Philips and healthy subjects from 1.5Philips

with AD patients from 1.5SIEMENS. These groups differed significantly in age (ρ<0.05).

Since gender is a categorical variable, a chi-squared test was used to analyze the frequency of each category in each group, using a function from the software R2014 (version 3.1.2). Here, diagnostic groups from 1.5GE, 1.5 SIEMENS and 3SIEMENS did not differ significantly in gender. However, some diagnostic groups from 1.5Philips differ significantly (ρ <0.05) from the diagnostic groups from the other three scanners.

6.2.1 Voxel Based Morphometry - Univariate Analysis

Statistical parametric maps of F-tests were computed to assess the main effects of disease and scanner, as well as their interaction. Here, statistically significant F-scores are reported (ρ <0.05 after FWE correction at the voxel-level) for the two main effects while a more liberal threshold (ρ <0.001 uncorrected) was applied to the interaction.

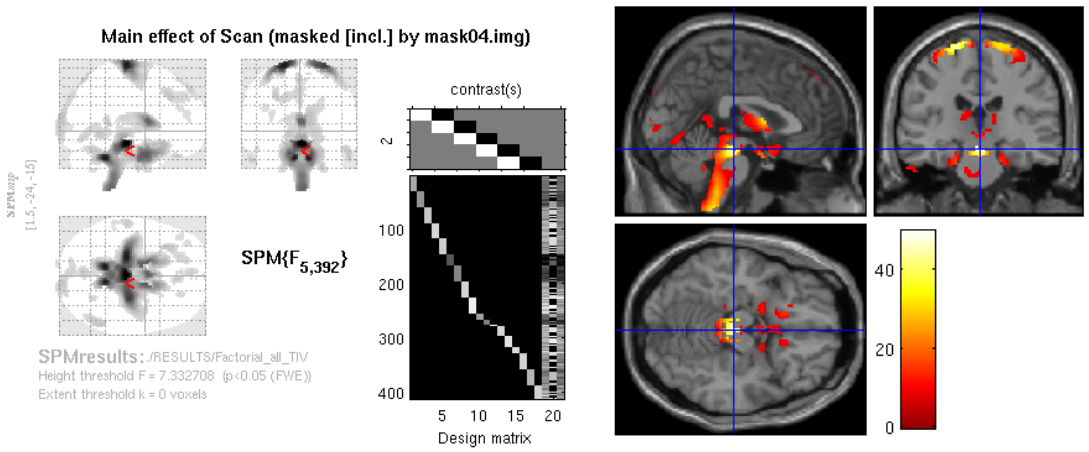

Figure 21 – VBM analysis of main effect of scanner (FWE ρ <0.05). Left: Brain Glass and Design matrix. Right: Main effect of scanner and F scores. Color bars indicate F-scores for each contrast.

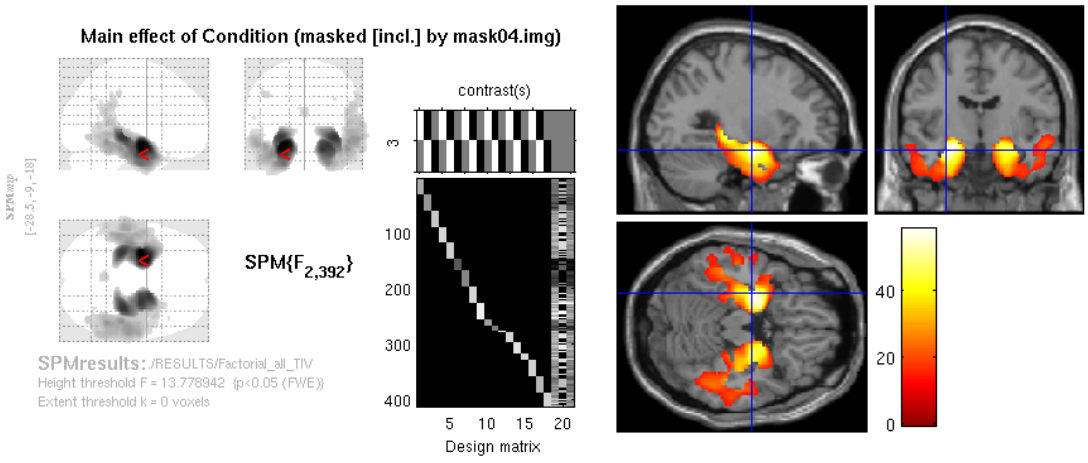

Figure 22 - VBM analysis of main effect of disease (FWE ρ <0.05). Left: Brain Glass and Design matrix. Right: Main effect of disease (MCI and AD) and F scores. Color bars indicate F-scores for each contrast.

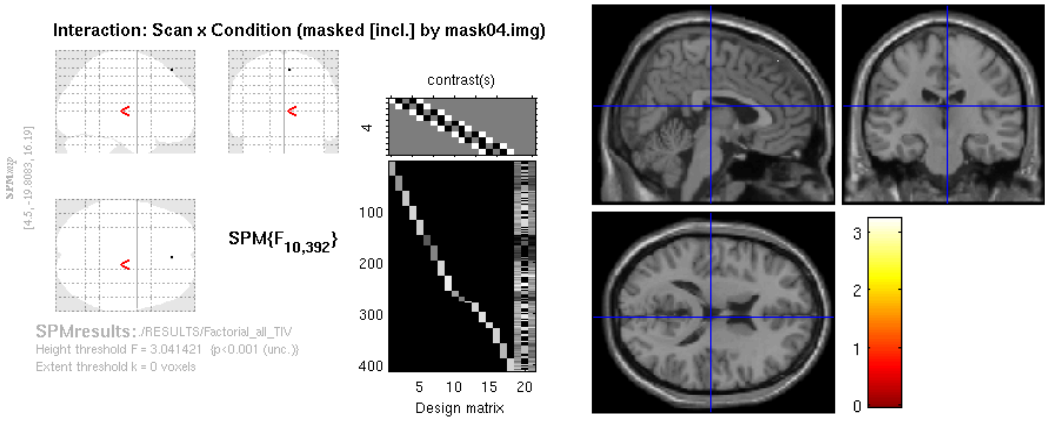

Figure 23 – Interaction between disease and scanner (ρ <0.001 uncorrected). Left: Brain Glass and Design matrix. Right: Interaction between disease and scanner and F-scores. Color bars indicate F-scores for each contrast.

6.2.2 Classification – Multivariate Analysis

Classification of scanner

Classification of scanner was conducted in order to demonstrate how well the scanner was distinguishable when using only healthy subjects (Table 6). Also, the correction of subject-specific covariates age, sex and TIV had little effect in classification performance. From Table 4 it is noticeable that healthy subjects from each scanner are well matched for age and gender.

Table 6 – Classification of scanner using data from healthy controls, with and without correction of age, sex and total intracranial volume using Gaussian process regression. Performance of scanner is reported as accuracy (acc) followed by sensitivity and specificity in brackets and area under the curve (auc). Statistically significant classification results (ρ <0.05) are presented with *.

	Correction	Method	
1.5 GE and 1.5 Philips	No		acc=0.69* (0.61, 0.77) auc=0.72*
	Yes	GPR	acc=0.65* (0.62, 0.69) auc=0.72*
1.5 GE and 1.5 SIEMENS	No		acc=0.83* (0.87, 0.80) auc=0.85*
	Yes	GPR	acc=0.70* (0.70, 0.70) auc=0.75*
1.5 GE and 3 SIEMENS	No		acc=0.87* (0.93, 0.80) auc=0.94*
	Yes	GPR	acc=0.85* (0.90, 0.80) auc=0.91*
1.5 SIEMENS and 1.5 Philips	No		acc=0.73* (0.65, 0.80) auc=0.81*
	Yes	GPR	acc=0.67 (0.65, 0.69) auc=0.75*
1.5 SIEMENS and 3 SIEMENS	No		acc=0.45 (0.40, 0.50) auc=0.36
	Yes	GPR	acc=0.45 (0.40, 0.50) auc=0.36
3 SIEMENS and 1.5 Philips	No		acc=0.71* (0.77, 0.65) auc=0.77*
	Yes	GPR	acc=0.63 (0.65, 0.62) auc=0.65

Scanner classification was higher when comparing 1.5 GE with 3 SIEMENS and it was not significant when comparing 1.5 SIEMENS with 3 SIEMENS. The multivariate discrimination weight maps for the significant classifications are presented in Figure 24.

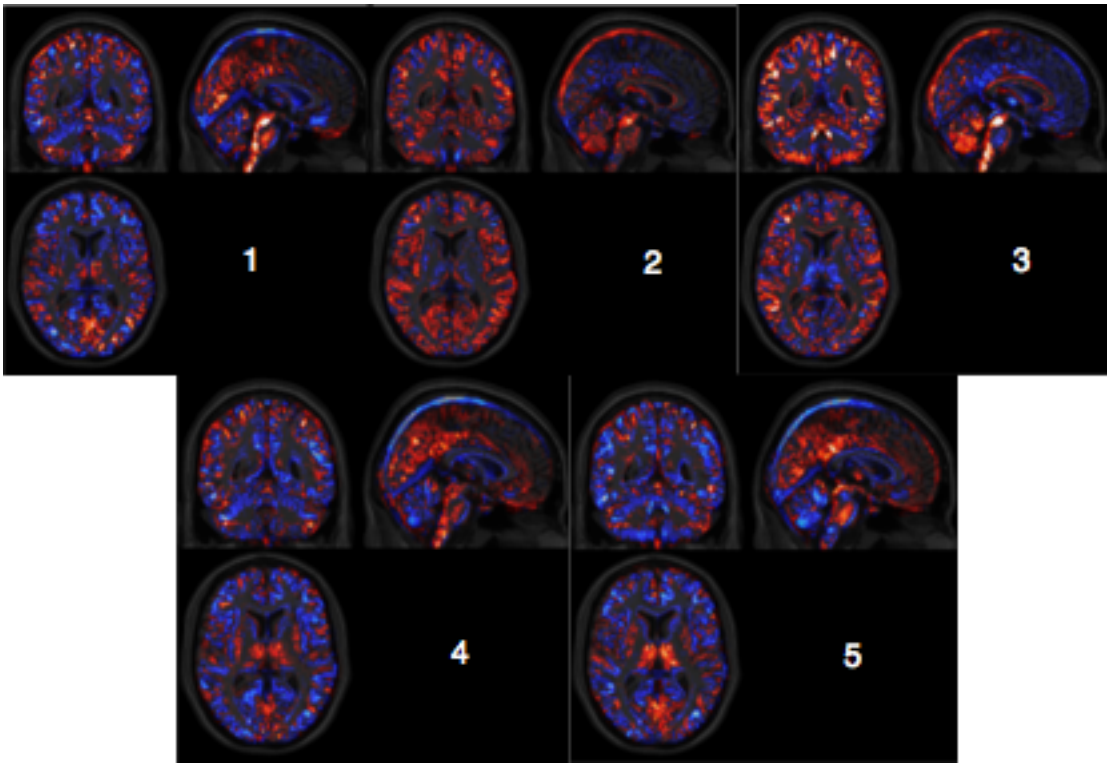

Figure 24 - Multivariate discrimination weight maps. Unthresholded GPC weights overlaid on an anatomical template. Weight map when performing scanner classification of (1) 1.5GE vs 1.5Philips, (2) 1.5GE vs 1.5SIEMENS, (3) 1.5GE vs 3SIEMENS, (4) 1.5SIEMENS vs 1.5Philips and (5) 3SIEMENS vs 1.5Philips. The color code shows the relative weight on each voxel for the decision boundary (red/yellows scales: higher weights for the first scanner and blue scales: higher weights for the second scanner).

Scenario 1 - Single site training: On-site application

This scenario is representative for research studies that avoid any potential issues with heterogeneous hardware and/or protocol (Table 7). Scanner specific cross-validation accuracies without correction ranged between 48% (1.5GE) and 70% (1.5 SIEMENS and 3 SIEMENS) for MCI and from 74% (3SIEMENS) to 88% (1.5SIEMENS) for AD.

Both scanners 1.5GE and 1.5Philips performed poorly when classifying the disease status and neither of them produced significant classification results, even after correction. Correction of covariates (sex, age, TIV) improved the AD classification but it did not improve MCI classification.

Table 7 – Single-scanner predictive performance of on-site classification without and with correction of age, sex and total intracranial volume using Gaussian process regression. Performance of Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD) classification. Performance of disease status is reported as accuracy (acc), followed by sensitivity and specificity, in brackets, and area under the curve (auc). Statistically significant classification results (ρ <0.05) are presented with *.

	Correction	Method	MCI	AD
1.5 GE	No		acc=0.48 (0.50, 0.47) auc=0.44	acc=0.75* (0.83, 0.67) auc=0.84*
	Yes	OLS	acc=0.50 (0.47, 0.48) auc=0.48	acc=0.77* (0.83, 0.70) auc=0.88*
		GPR	acc=0.53 (0.47, 0.60) auc=0.56	acc=0.80* (0.83, 0.77) auc=0.85*
1.5 SIEMENS	No		acc=0.70* (0.73, 0.67) auc=0.71*	acc=0.88* (0.90, 0.87) auc=0.93*
	Yes	OLS	acc=0.70* (0.67, 0.73) auc=0.73*	acc=0.92* (0.97, 0.87) auc=0.94*
		GPR	acc=0.70* (0.67, 0.73) auc=0.76*	acc=0.90* (0.87, 0.93) auc=0.92*
1.5 Philips	No		acc=0.54 (0.58, 0.50) auc=0.56	acc=0.73* (0.80, 0.65) auc=0.76*
	Yes	OLS	acc=0.56 (0.58, 0.54) auc=0.54	acc=0.77* (0.85, 0.70) auc=0.77*
		GPR	acc=0.60 (0.50, 0.69) auc=0.58	acc=0.72* (0.75, 0.70) auc=0.76*
3 SIEMENS	No		acc=0.70* (0.70, 0.70) auc=0.67*	acc=0.74* (0.72, 0.76) auc=0.77*
	Yes	OLS	acc=0.63* (0.63, 0.63) auc=0.66*	acc=0.76* (0.76, 0.76) auc=0.79*
		GPR	acc=0.60 (0.57, 0.63) auc=0.66*	acc=0.74* (0.72, 0.76) auc=0.77*

Scenario 2- Single site training: Transfer to another site

This scenario consisted in evaluating how well a model learned on one scanner can be transferred to another scanner that was not involved in the training stage. As demonstrated in Table 8, when applying a classifier that was trained on one scanner to data from another, a significant decline in improvement in performance was not observed for MCI or for AD.

In MCI classification, the average accuracy and area under the curve without correction were 0.66 and 0.71. With correction and using ROLS, average acc and auc were 0.67 and 0.71 and using GPR were 0.58 and 0.72, respectively. In AD classification, the average accuracy and area under the curve without correction were 0.82 and 0.89. With correction and using ROLS, average acc and auc were 0.80 and 0.90 and using GPR were 0.77 and 0.89, respectively. Therefore, correction was not overall beneficial for MCI and AD classification in this scenario.

Table 8 – Single scanner training is applied to a data from a new scanner for testing. Performance of Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD) classification. Performance of disease status is reported as accuracy (acc) and area under the curve (auc) without and with correction of scanner plus age, sex and total intracranial volume using Ridge Ordinary Least Squares regression (ROLS) and Gaussian Process regression (GPR). Performances after correction were compared with performances without correction. Results presented with * are significant performance classifications without correction (ρ <0,05). Accuracies presented with * are accuracies that show significantly improvement (ρ <0,05) after correction by the McNemar Test. Accuracies and auc presented with * show significant improvement (ρ <0,05) after correction by the Permutation Test.

		Testing			
	Corr.	**1.5 GE**	**1.5 SIEMENS**	**1.5 Philips**	**3 SIEMENS**
Training	**MCI**				
1.5GE	No		acc=0.70* auc=0.71*	acc=0.58 auc=0.59	acc=0.72* auc=0.75*
	ROLS		acc=0.68 auc=0.72	acc=0.62 auc=0.59	acc=0.68 auc=0.77
	GPR		acc=0.52 auc=0.73	acc=0.52 auc=0.62	acc=0.48 auc=0.79
1.5 SIEMENS	No	acc=0.65 auc=0.66		acc=0.58 auc=0.67	acc=0.78* auc=0.88*
	ROLS	acc=0.70 auc=0.65		acc=0.65 auc=0.65	acc=0.80 auc=0.88
	GPR	acc=0.55 auc=0.65		acc=0.58 auc=0.72	acc=0.80 auc=0.88
1.5 Philips	No	acc=0.48 auc=0.56	acc=0.62 auc=0.70*		acc=0.68* auc=0.72*
	ROLS	acc=0.50 auc=0.57	acc=0.63 auc=0.68		acc=0.67 auc=0.71
	GPR	acc=0.48 auc=0.56	acc=0.55 auc=0.68		acc=0.57 auc=0.72
3 SIEMENS	No	acc=0.68* auc=0.68*	acc=0.80* auc=0.89*	acc=0.65 auc=0.72*	
	ROLS	acc=0.63 auc=0.68	acc=0.85 auc=0.89	acc=0.67 auc=0.69	
	GPR	acc=0.53 auc=0.69	acc=0.82 auc=0.88	acc=0.58 auc=0.73	
Training	**AD**				
1.5GE	No		acc=0.83* auc=0.95*	acc=0.78* auc=0.86*	acc=0.78* auc=0.87*
	ROLS		acc=0.92** auc=0.94	acc=0.80 auc=0.87	acc=0.78 auc=0.88
	GPR		acc=0.80 auc=0.88	acc=0.78 auc=0.87	acc=0.72 auc=0.87
1.5 SIEMENS	No	acc=0.78* auc=0.89*		acc=0.82* auc=0.87*	acc=0.82* auc=0.92*
	ROLS	acc=0.70 auc=0.65		acc=0.82 auc=0.88	acc=0.86 auc=0.94
	GPR	acc=0.75 auc=0.91		acc=0.75 auc=0.86	acc=0.84 auc=0.92
1.5 Philips	No	acc=0.80* auc=0.88*	acc=0.72* auc=0.88*		acc=0.76* auc=0.89*
	ROLS	acc=0.80 auc=0.90	acc=0.87** auc=0.87		acc=0.82* auc=0.89
	GPR	acc=0.72 auc=0.89	acc=0.87** auc=0.86		acc=0.74 auc=0.88
3 SIEMENS	No	acc=0.75* auc=0.87*	acc=0.87* auc=0.93*	acc=0.82* auc=0.88*	
	ROLS	acc=0.75 auc=0.88	acc=0.88 auc=0.93	acc=0.80 auc=0.90	
	GPR	acc=0.68 auc=0.88	acc=0.85 auc=0.93	acc=0.78 auc=0.89	

Correction for scanner and subject-specific covariates did not increase accuracy performance for the MCI subjects; instead it led to a decrease in accuracy in the majority of the cases, specially using Gaussian Process regression method for correction. However, the area under the curve showed that performance was constant or slightly increased in most cases for both correction methods.

As stated before in Table 7, both scanners 1.5 GE and 1.5 Philips performed poorly when classifying the disease status for MCI when compared with 1.5 SIEMENS and 3 SIEMENS. As shown in Table 8, only the accuracies involving 1.5 GE and 1.5 Philips have decreased after correction. One interesting effect when correcting with GPR, in these cases, is that accuracy dropped to near 0.5, losing the ability to discriminate between classes. Figure 25 shows the GPC predictions before and after GPR correction with a model trained using healthy and MCI subjects from 1.5 GE and tested using healthy and MCI subjects from 3 SIEMENS.

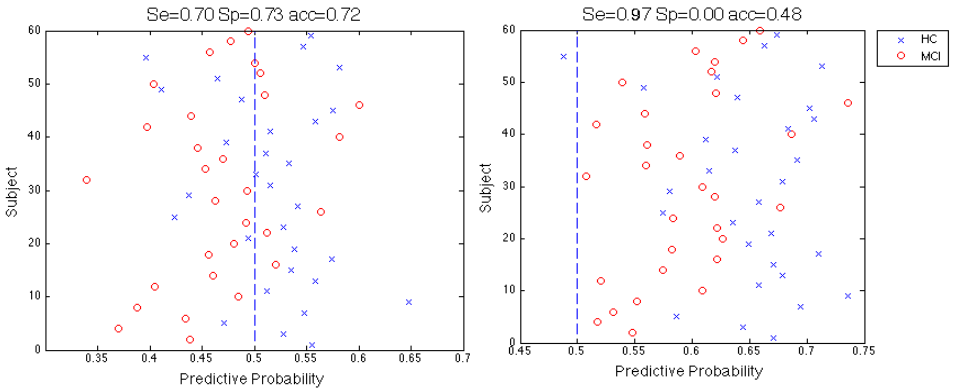

Figure 25 – (Left) Classification MCI accuracies for GPC predictors before correction. (Right) Classification MCI accuracies for GPC predictors after GPR correction. The model was trained using healthy and MCI subjects from 1.5GE and tested using healthy and MCI subjects from 3 SIEMENS. MCI subjects were assigned as -1 and Healthy controls (HC) were assigned as 1.

Figure 25 shows that after correction the threshold, previously defined as 0,50, is no longer suitable. Area under the curve summarizes overall performance over all possible thresholds and shows an increase from 0.75 to 0,9, in this case.

However, when training and testing the model with 1.5 SIEMENS and 3 SIEMENS, the accuracy slightly improved with both correction methods. Though, this improvement was not significant and classification of scanner showed that scanner variability effect is not strong between these two scanners.

Scenario 3- Group-specific sites for training: On-site application

In this scenario, disease and scanner were confounded in the training and the testing data sets and the reciprocal data sets were classified. Test performance significantly declined in both MCI and AD classifiers.

In MCI classification, the average accuracy and area under the curve without correction were 0.33 and 0.29. With correction and using ROLS, average acc and auc were 0.56 and 0.59 and using GPR were 0.56 and 0.57 respectively. In AD classification, the average accuracy and area under the curve without correction were 0.65 and 0.70. With correction and using ROLS, average acc and auc were 0.79 and 0.87 and using GPR were 0.79 and 0.86, respectively.

Table 9 – Disease and scanner were confounded in the training and testing data sets and the reciprocal data sets were classified. Here, controls from Scanner 1 and patients from Scanner 2 were used in the training phase and controls from Scanner 2 and patients from Scanner 1 were used in the testing phase. Performance of Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD) classification. Performance of disease status is reported as accuracy (acc) and area under the curve (auc) without and with correction of scanner plus age, sex and total intracranial volume using Ridge Ordinary Least Squares regression (ROLS) and Gaussian Process regression (GPR). Performances after correction were compared with performances without correction. Results presented with * are significant performance classifications without correction (ρ <0,05). Accuracies presented with * are accuracies that show significantly improvement (ρ <0,05) after correction by the McNemar Test. Accuracies and auc presented with * show significant improvement (ρ <0,05) after correction by the Permutation Test.

		Scanner 2			
	Corr.	**1.5 GE**	**1.5 SIEMENS**	**1.5 Philips**	**3 SIEMENS**
Scanner 1		**MCI**			
1.5GE	No		acc=0.35 auc=0.22	acc=0.27 auc=0.17	acc=0.13 auc=0.05
	OLS		acc=0.57 auc=0.61	acc=0.48 auc=0.45	acc=0.43 auc=0.46
	GPR		acc=0.57 auc=0.59	acc=0.50 auc=0.52	acc=0.50 auc=0.52
1.5 SIEMENS	No	acc=0.35 auc=0.21		acc=0.23 auc=0.29	acc=0.62* auc=0.74*
	OLS	acc=0.57 auc=0.61		acc=0.50 auc=0.54	acc=0.77** auc=0.83*
	GPR	acc=0.63** auc=0.65*		acc=0.54 auc=0.54	acc=0.73** auc=0.85*
1.5 Philips	No	acc=0.31 auc=0.24	acc=0.31 auc=0.25		acc=0.25 auc=0.22
	OLS	acc=0.52 auc=0.51	acc=0.58 auc=0.62		acc=0.67** auc=0.65
	GPR	acc=0.50 auc=0.52	acc=0.60 auc=0.52		acc=0.60 auc=0.60
3 SIEMENS	No	acc=0.13 auc=0.08	acc=0.70* auc=0.76*	acc=0.35 auc=0.22	
	OLS	acc=0.37 auc=0.36	acc=0.77* auc=0.85*	acc=0.50 auc=0.53	
	GPR	acc=0.25 auc=0.20	acc=0.77* auc=0.84*	acc=0.52 auc=0.54	
Scanner 1		**AD**			
1.5GE	No		acc=0.67* auc=0.70*	acc=0.75* auc=0.83*	acc=0.52 auc=0.56
	OLS		acc=0.83** auc=0.88*	acc=0.85 auc=0.94	acc=0.62 auc=0.76*
	GPR		acc=0.77 auc=0.86*	acc=0.88 auc=0.93	acc=0.66* auc=0.76*
1.5 SIEMENS	No	acc=0.82* auc=0.87*		acc=0.45 auc=0.56	acc=0.82* auc=0.92*
	OLS	acc=0.87 auc=0.98		acc=0.88** auc=0.96*	acc=0.86 auc=0.96
	GPR	acc=0.88 auc=0.96		acc=0.78** auc=0.98*	acc=0.88 auc=0.96
1.5 Philips	No	acc=0.68 auc=0.67	acc=0.48 auc=0.51		acc=0.57 auc=0.61
	OLS	acc=0.72 auc=0.78	acc=0.70** auc=0.71*		acc=0.73** auc=0.79*
	GPR	acc=0.70 auc=0.74	acc=0.70 auc=0.71		acc=0.72** auc=0.76*
3 SIEMENS	No	acc=0.54 auc=0.60	acc=0.78* auc=0.84*	acc=0.72* auc=0.76*	
	OLS	acc=0.78** auc=0.85*	acc=0.82 auc=0.90	acc=0.83 auc=0.95*	
	GPR	acc=0.76** auc=0.84*	acc=0.84 auc=0.87	acc=0.85 auc=0.94*	

After correction, MCI and AD classifications significantly improved in most cases. In MCI classification only statistically significant classifications after correction were reported as a significant improve in Table 9.

Scenario 4- Group-specific site for training: Transfer to another site

In this scenario, a model learned with data from group-specific scanners was tested with data from a third scanner.

In MCI classification, the average accuracy and area under the curve without correction were 0.61 and 0.67. With correction and using ROLS, average acc and auc were 0.60 and 0.66 and using GPR were 0.59 and 0.71 respectively.

Similar to what happened in Scenario 2, correction for scanner and subject-specific covariates did not increase accuracy performance for the MCI subjects. In some cases, it led to a decrease in accuracies, especially after Gaussian Process regression. However, after Gaussian Process regression, area under the curve performances increased significantly in several cases.

In the same way, when the MCI subjects from 1.5GE and 1.5Philips scanners were not considered in the training or testing set, the accuracy and area under the curve improved with both correction methods (Table 10).

Table 10 – Disease and scanner were confounded in the training set and a third scanner was used to test the classifier. Performance of Mild Cognitive Impairment (MCI) classification. Performance of disease status is reported as accuracy (acc) and area under the curve (auc) without and with correction of scanner plus age, sex and total intracranial volume using Ridge Ordinary Least Squares regression (ROLS) and Gaussian Process regression (GPR). Performances after correction were compared with performances without correction. Results presented with * are significant performance classifications without correction (ρ <0,05). Accuracies presented with * are accuracies that show significantly improvement (ρ <0,05) after correction by the McNemar Test. Accuracies and auc presented with * show significant improvement (ρ <0,05) after correction by the Permutation Test.

		Testing			
	Corr.	**1.5 GE**	**1.5 SIEMENS**	**1.5 Philips**	**3 SIEMENS**
Training		**MCI**			
1.5GE (NL) 1.5SIEMENS (MCI)	No			acc=0.58 auc=0.64	acc=0.68* auc=0.78*
	OLS			acc=0.60 auc=0.62	acc=0.70 auc=0.82*
	GPR			acc=0.58 auc=0.64	acc=0.75 auc=0.86*
1.5GE (NL) 1.5Philips (MCI)	No		acc=0.63* auc=0.70*		acc=0.60 auc=0.65*
	OLS		acc=0.57 auc=0.62		acc=0.55 auc=0.60
	GPR		acc=0.57 auc=0.74*		acc=0.60 auc=0.72*
1.5GE (NL) 3SIEMENS (MCI)	No		acc=0.65* auc=0.78*	acc=0.60 auc=0.66*	
	OLS		acc=0.72 auc=0.79	acc=0.67 auc=0.64	
	GPR		acc=0.75 auc=0.84*	acc=0.54 auc=0.71*	
1.5SIEMENS (NL) 1.5GE (MCI)	No			acc=0.54 auc=0.59	acc=0.70* auc=0.84*
	OLS			acc=0.46 auc=0.52	acc=0.63 auc=0.76
	GPR			acc=0.50 auc=0.64*	acc=0.52 auc=0.85
1.5SIEMENS (NL) 1.5Philips (MCI)	No	acc=0.50 auc=0.51			acc=0.58 auc=0.69*
	OLS	acc=0.47 auc=0.49			acc=0.60 auc=0.61
	GPR	acc=0.52 auc=0.53			acc=0.62 auc=0.73*

1.5SIEMENS (NL) 3SIEMENS (MCI)	No	acc=0.58 auc=0.65*		acc=0.63* auc=0.69*	
	OLS	acc=0.58 auc=0.58		acc=0.52 auc=0.63	
	GPR	acc=0.60 auc=0.66		acc=0.60 auc=0.72	
1.5Philips (NL) 1.5GE (MCI)	No		acc=0.55 auc=0.55		acc=0.70* auc=0.70*
	OLS		acc=0.68* auc=0.71*		acc=0.73 auc=0.79*
	GPR		acc=0.53 auc=0.60		acc=0.55 auc=0.76*
1.5Philips (NL) 1.5SIEMENS (MCI)	No	acc=0.57 auc=0.57			acc=0.65* auc=0.81*
	OLS	acc=0.53 auc=0.62			acc=0.82** auc=0.88*
	GPR	acc=0.57 auc=0.63			acc=0.75** auc=0.89*
1.5Philips (NL) 3SIEMENS (MCI)	No	acc=0.50 auc=0.61*	acc=0.65* auc=0.76*		
	OLS	acc=0.57 auc=0.58	acc=0.68 auc=0.78		
	GPR	acc=0.53 auc=0.71	acc=0.77** auc=0.83*		
3SIEMENS (NL) 1.5GE (MCI)	No		acc=0.73* auc=0.81*	acc=0.62 auc=0.57	
	OLS		acc=0.73 auc=0.79	acc=0.50 auc=0.56	
	GPR		acc=0.52 auc=0.83	acc=0.52 auc=0.66*	
3SIEMENS (NL) 1.5SIEMENS (MCI)	No	acc=0.63* auc=0.67*		acc=0.62 auc=0.67*	
	OLS	acc=0.52 auc=0.59		acc=0.52 auc=0.61	
	GPR	acc=0.60 auc=0.66		acc=0.60 auc=0.71	
3SIEMENS (NL) 1.5Philips (MCI)	No	acc=0.57 auc=0.54	acc=0.63 auc=0.75*		
	OLS	acc=0.48 auc=0.50	acc=0.65 auc=0.70		
	GPR	acc=0.52 auc=0.55	acc=0.58 auc=0.76		

In AD classification, the average accuracy and area under the curve without correction were 0.77 and 0.86. With correction and using ROLS, average acc and auc were 0.77 and 0.86 and using GPR were 0.75 and 0.88 respectively. Overall, correction of covariates were not significantly beneficial for AD classification in this scenario (Table 11).

Table 11 - Disease and scanner were confounded in the training set and a third scanner was used to test the classifier. Performance of Alzheimer's disease (AD) classification. Performance of disease status is reported as accuracy (acc) and area under the curve (auc) without and with correction of scanner plus age, sex and total intracranial volume using Ridge Ordinary Least Squares regression (ROLS) and Gaussian Process regression (GPR). Performances after correction were compared with performances without correction. Results presented with * are significant performance classifications without correction (p <0,05). Accuracies presented with * are accuracies that show significantly improvement (p<0,05) after correction by the McNemar Test. Accuracies and auc presented with * show significant improvement (p<0,05) after correction by the Permutation Test.

		Testing			
	Corr.	**1.5 GE**	**1.5 SIEMENS**	**1.5 Philips**	**3 SIEMENS**
Training	**AD**				
1.5GE (NL) 1.5SIEMENS (AD)	No			acc=0.72* auc=0.85*	acc=0.75* auc=0.87*
	OLS			acc=0.82 auc=0.88	acc=0.80 auc=0.92
	GPR			acc=0.75 auc=0.85	acc=0.84* auc=0.90

1.5GE (NL) 1.5Philips (AD)	No		acc=0.83* auc=0.91*		acc=0.70* auc=0.86*
	OLS		acc=0.85 auc=0.92		acc=0.76 auc=0.86
	GPR		acc=0.77 auc=0.90		acc=0.72 auc=0.87
1.5GE (NL) 3SIEMENS (AD)	No		acc=0.75* auc=0.85*	acc=0.77* auc=0.81*	
	OLS		acc=0.85** auc=0.91*	acc=0.80 auc=0.90*	
	GPR		acc=0.82 auc=0.89	acc=0.80 auc=0.86	
1.5SIEMENS (NL) 1.5GE (AD)	No			acc=0.88* auc=0.89*	acc=0.78* auc=0.93*
	OLS			acc=0.68 auc=0.79	acc=0.68 auc=0.86
	GPR			acc=0.78 auc=0.89	acc=0.74 auc=0.91
1.5SIEMENS (NL) 1.5Philips (AD)	No	acc=0.65* auc=0.79*			acc=0.72* auc=0.82*
	OLS	acc=0.52 auc=0.57			acc=0.64 auc=0.68
	GPR	acc=0.63 auc=0.88			acc=0.66 auc=0.93
1.5SIEMENS (NL) 3SIEMENS (AD)	No	acc=0.75* auc=0.87*		acc=0.85* auc=0.88*	
	OLS	acc=0.77 auc=0.92		acc=0.85 auc=0.87	
	GPR	acc=0.68 auc=0.89		acc=0.80 auc=0.89	
1.5Philips (NL) 1.5GE (AD)	No		acc=0.72* auc=0.82*		acc=0.76* auc=0.82*
	OLS		acc=0.88** auc=0.91*		acc=0.78 auc=0.88
	GPR		acc=0.75 auc=0.84		acc=0.76 auc=0.83
1.5Philips (NL) 1.5SIEMENS (AD)	No	acc=0.75* auc=0.81*			acc=0.72* auc=0.82*
	OLS	acc=0.75 auc=0.87			acc=0.80 auc=0.89
	GPR	acc=0.73 auc=0.86			acc=0.78 auc=0.86*
1.5Philips (NL) 3SIEMENS (AD)	No	acc=0.73* auc=0.78*	acc=0.72* auc=0.81*		
	OLS	acc=0.65 auc=0.84*	acc=0.78 auc=0.86		
	GPR	acc=0.68 auc=0.82	acc=0.78 auc=0.85		
3SIEMENS (NL) 1.5GE (AD)	No		acc=0.92* auc=0.96*	acc=0.80* auc=0.91*	
	OLS		acc=0.87 auc=0.96	acc=0.77 auc=0.82	
	GPR		acc=0.82 auc=0.95	acc=0.80 auc=0.90	
3SIEMENS (NL) 1.5SIEMENS (AD)	No	acc=0.75* auc=0.87*		acc=0.75* auc=0.87*	
	OLS	acc=0.85* auc=0.92		acc=0.85 auc=0.89	
	GPR	acc=0.75 auc=0.88		acc=0.75 auc=0.91	
3SIEMENS (NL) 1.5Philips (AD)	No	acc=0.87* auc=0.94*	acc=0.90* auc=0.96*		
	OLS	acc=0.70 auc=0.79	acc=0.80 auc=0.89		
	GPR	acc=0.70 auc=0.95	acc=0.70 auc=0.92		

Chapter 7: Discussion and Conclusion

As stated in the Introduction chapter, this thesis had three main goals, which were pursued in phases:

1. Illustrating and analysing the extent of scanner variability and subject-variability effects by using voxel-based morphometry and machine learning based on Gaussian process classification. Two different datasets comprising only healthy subjects scanned twice with different coils and different scanner parameters, respectively, were used to illustrate the scanner variability effect.
2. Data from the ADNI project database was used to analyse the impact of scanner and subject variability in automated classification. Healthy subjects, MCI and Alzheimer's patients were involved. The dataset was acquired in 6 different scanners, but only data from 4 scanners were used in machine learning analysis. The protocols were similar in each scanner.
3. Two different methods for correcting data were used in order to study their reliability to remove out confound effect. Regression methods were: Gaussian process regression and Ridge ordinary least squares regression.

7.1 VBM Univariate Analyses and scanner variability effect

In a first phase, a univariate VBM analyses was performed, using SPM8, in order to illustrate the extent of scanner variability effect. VBM analyses largely confirmed the presence of substantial inter-scanner differences at the group level. For the first two datasets, comprising only healthy subjects, the scanner variability effect is spread all over the gray matter with strong activations in localized regions (Figures 15 and 16).

For the ADNI Project dataset, the scanner-variability effect is strong in localized regions and contrary to what happened in the previous dataset, the effect is not dispersed in the gray matter (Figure 21). The effect of disease (Figure 22) was substantially larger than the effect of scanner and failed to find a significant interaction of disease with scanner (Figure 23).

7.2 Multivariate Analysis – Classification and Scanner variability effect

In a first approach, two different datasets with only healthy subjects were used. These subjects were scanned twice with different hardware and scanner parameters. In the first dataset 14 healthy subjects were scanned twice with different coils, and in the second dataset 41 healthy subjects were scanned twice with different scanner parameters. The goal was to perform classification of scanner and demonstrate how well the two scans from the same subjects were distinguishable. Since we are comparing scans from the same subject, inter-subject variability is negligible.

However, a small amount of noise due to scanner noise, physiological noise, and other potential confounds can be still present. In this scenario, classification of scanner was possible with very high accuracy and area under the curve performances (Figure 17 and 19).

For the ADNI project dataset, classification of scanner was also conducted using only healthy subjects in order to investigate if the classifier could distinguish subjects only taking into account differences in scanners. Classification of scanner was also possible with very high performance accuracy and area under the curve, especially when comparing 1.5GE scanner with 3 SIEMENS scanner. Classification of 1.5 GE and 1.5SIEMENS were also high and classification performance of 1.5 SIEMENS with 3SIEMENS was not significant (Table 6). Therefore, this supports the evidence found in other studies that differences between data from scanners of different manufacturers were larger than between data from the same manufacturer [8][15].

Multivariate discrimination weight maps (Figure 18 and 20 and 24) represent the voxels that contribute most strongly and reliably to the classifier's success. In the voxel space, the weight vector normal to the hyperplane will be the direction along which images of the two groups differ most. Hence, it can be used to generate a map with the most discrimination regions. Since the classifier has a multivariate nature, the combination of all voxels as a whole is identified as a global spatial pattern by which the groups differ. Therefore, multivariate pattern recognition analyses reveal discriminating activation patterns and while these maps often show correspondence with the direction of the magnitude of changes but as they are influenced by variance and covariance there is not a direct statistical correspondence.

However, when comparing the weight maps with the univariate analysis, several brain areas that have shown a strong activation in VBM group analysis, overlapped with regions that show higher weights. This is quite noticeable when comparing VBM analysis of main effect of scanner (Figure 21) with the multivariate discrimination weight maps for scanner classification using ADNI project healthy subjects (Figure 24).

7.3 Multivariate analysis – Clinical Scenarios

Scenario 1 – Single site training: On-site application

In the first clinical scenario, controls and the respective patients group were used for classification separately for each of the four sites. If clinically applied, this setting would represent a situation at a large imaging centre, where a sufficiently large training set of both diagnostic groups is available.

The results for this scenario show that both scanners 1.5 GE and 1.5 Philips performed poorly when classifying the MCI disease status (Table 7). In fact, the classifier could not discriminate between healthy controls and MCI patients. For the AD classification, all scanners performed well. Therefore, one possible explanation for these results is that the MCI group selected for these scanners was not homogenous. Mild cognitive impairment is an intermediate stage between the expected cognitive decline of normal aging and the more serious decline of dementia

(Alzheimer's disease). There are several stages of mild cognitive impairment and it can involve problems with memory, language, thinking and judgment and therefore brain structural changes can be quite heterogeneous in a group of MCI patients. For classification it is important that diagnostic groups are homogeneous.

Scenario 2 – Single site training: Transfer to another site

The second scenario is representative of the setting where a specialized imaging centre provides the trained classifier for a clinic without specialized equipment. In this clinical setting, it is paramount that the performance does not drop due to the transfer of the model. In this study, the prediction accuracy did not drop significantly, and this scenario is comparable with Scenario 1 where controls and patients groups were classified separately for each scanner. However, when correcting for scanner and subject-specific covariates, accuracy decreased in the majority of the cases for the MCI classification, particularly when using Gaussian Process correction (Table 8).

This drop in accuracy only occurred when scanners 1.5 GE and 1.5 Philips were involved in the training or testing. Although the accuracy has decreased, the area under the curve improved in some cases after correction. Figure 25 shows the GPC predictions for MCI classification before and after correcting with GPR, when training the model with MCI and healthy subjects from 1.5GE and testing with MCI and healthy subjects from 3 SIEMENS. Apparently, the threshold previously defined as 0.50, is no longer suitable for the classification after correction. Moreover, area under the curve summarizes performances over all possible thresholds and it shows an increase from 0,75 to 0,79 in this case. This drop in accuracy and decalibration of the threshold can be due to the problem of MCI subjects from 1.5 GE and 1.5 Philips. Therefore, future work will involve studying a way to calibrate the GPC threshold without overfitting.

Overall, correction in MCI and AD classification did not improve the result significantly in all cases. Therefore, these results are consistent with Ahmed Abdulkadir *et al*. study [8] where he states that when training on a single scanner, generalization to other scanners is satisfactory.

Scenario 3 – Group-specific sites for training: On-site application

In the third scenario, two different scanners provided different diagnostic groups for training and the reciprocal data sets were used in the testing phase. Classification accuracy decreased substantially for MCI and AD classification when compared to the previous scenarios (Table 9). Since each scanner provided one diagnostic group for training, the classifier primarily detected scanner differences, systematically leading to incorrect classifications. This was evident in this case where the test data were acquired at one of the sites involved in training. The ROLS regression and GPR generally improved the performance in both classifications.

For the MCI classification, there are two important observations. First, several classification performances increased substantially but remain not statistically significant. This could be due to the previous problem with MCI subjects from scanners 1.5 GE and 1.5 Philips. Second, there is a smaller decrease in classification

when confounding the scanners 1.5 SIEMENS and 3 SIEMENS without correction. As stated before, the scanner classification for this two scanners were not significant and the classifier could not distinguish each scanner. However, correction significantly improved classification in this case, showing that even when scanners from the same manufacturer are confounded with disease, impairment in classification may still exist. Yet, this did not happen for the AD classification, where classification improved after correction but not significantly. Structural changes in the stage of AD are much larger than MCI stage. Therefore, differences due to AD can overcome smaller scanner differences in this case.

Although only results from correction of scanner and subject specific covariates are shown, the same analyses were done taking only scanner differences as covariate and only subject specific covariates. In case of improvement, correcting for scanner differences showed more improvement than correcting only for subject specific covariates.

Scenario 4 – Group-specific site for training: Transfer to another site

The fourth and last scenario consisted in transferring a model learned with data from group-specific scanners to data from a third scanner for testing. In this case, the drop in accuracy is not so evident as the previous scenario (Table 10 and 11). Moreover, for the MCI classification (Table 10), correction for scanner and subject-specific covariates did not improve accuracy but there was an overall increase in area under the curve after GPR correction. Furthermore, there was an increase in classification when MCI subjects from 1.5GE and 1.5Philips were not taken into account. This supports the hypothesis that MCI patients from these two scanners are not homogeneous and may lead to misleading results when combined with other MCI patients from other scanners.

For the AD classification (Table 11), the correction of covariates was not significantly beneficial. Though, there are some cases where ROLS regression was beneficial for this classification and other cases where the classification has dropped. Overall there were no significantly differences after correction mainly because structural differences due to AD are strong enough and can overcome scanner and subject variability, as stated before.

Similar to the last scenario, in case of improvement, correcting for scanner differences showed more improvement than correcting for subject specific covariates.

7.4 Limitations and Future work

A number of limitations have to be pointed out when interpreting the results. The first and strongest limitation of this work is the MCI classification performance of scanners 1.5GE and 1.5Philips. The problem with the MCI subjects from these scanners has compromised the study of the benefit of correction in several cases, specifically in the last scenario. Care should be taken when using data from different scanners and balanced diagnostic groups must be ensured. Future work could involve

assessing the clinical scores of these subjects in terms of severity of the cognitive impairment.

A second limitation is the significantly differences in gender when comparing diagnostic groups from 1.5 Philips to other scanners. Including age, sex and intracranial volume into the analyses can capture some of the sample variance but can be incomplete. Scanning of the same population at all sites would be more reliable but it is impractical with a clinical population and such effort would not directly answer the question of whether data from different subjects on different scanners can be pooled.

Both correction methods performed similar in correction, contrary to what happened in Ahmed Abdulkadir *et al*. study where OLS performed poorly. In this case, Ridge OLS was applied meaning that a regularisation parameter was estimated for each scenario in order to prevent multicollinearity. This is a biased estimation that is well known and widely used for being computationally very fast. However GPR also improved the performance and compared to ROLS has two main advantages: (1) the model includes hyperparameters that could adapt to the data, particularly the regularization term for the noise which acts as regulizer of the solution and thereby reducing overvitting and (2) the Gaussian process covariance allows non-linear correction, thereby modelling interactions between design variables.

The ADNI project dataset is a highly standardized and controlled which is not representative for a clinical study. Moreover, the structural changes caused by Alzheimer's disease can be quite large as well as changes caused by Mild cognitive impairment. Future work would be to translate these findings to other neurodegenerative diseases and psychiatric disorders, such as schizophrenia.

Schizophrenia structural changes are much smaller than Mild cognitive impairment and Alzheimer's disease. The next step will consist in accessing data from different patients with schizophrenia acquired in different sites and study the effectiveness of these correction methods in classification. This will also be important since the effect of scanner variability can be differently localized in some regions in the brain. Thus, this may affect classification performance differently, depending the brain disease and scanner differences.

REFERENCES

[1] C. J. Aine, "A conceptual overview and critique of functional neuroimaging techniques in humans: I. MRI/FMRI and PET," Crit. Rev. Neurobiol., vol. 9, no. 2–3, pp.229–309,1995.

[2] W.L. Luo and T. E. Nichols, "Diagnosis and exploration of massively univariate neuroimaging models," NeuroImage, vol. 19, no. 3, pp. 1014–1032, Jul. 2003

[3] N. K. Focke, G. Helms, S. Kaspar, C. Diederich, V. Tóth, P. Dechent, A. Mohr, and W. Paulus, "Multi-site voxel-based morphometry--not quite there yet," NeuroImage, vol.56, no.3, pp.1164–1170, Jun.2011.

[4] J. Chen, J. Liu, V. D. Calhoun, A. Arias-Vasquez, M. P. Zwiers, C. N. Gupta, B. Franke, and J. A. Turner, "Exploration of scanning effects in multi-site structural MRI studies," J. Neurosci. Methods, vol. 230, pp. 37–50, Jun. 2014.

[5] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," NeuroImage, vol. 45, no. 1, Supplement 1, pp. S199–S209, Mar. 2009.

[6] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, and Alzheimer's Disease Neuroimaging Initiative, "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters," *NeuroImage*, vol. 50, no. 3, pp. 883–892, Apr. 2010.

[7] M. Dyrba, M. Ewers, M. Wegrzyn, I. Kilimann, C. Plant, A. Oswald, T. Meindl, M. Pievani, A. L. W. Bokde, A. Fellgiebel, M. Filippi, H. Hampel, S. Klöppel, K. Hauenstein, T. Kirste, S. J. Teipel, and the EDSD study group, "Robust Automated Detection of Microstructural White Matter Degeneration in Alzheimer's Disease Using Machine Learning Classification of Multicenter DTI Data," *PLoS ONE*, vol. 8, no. 5, p. e64925, May 2013.

[8] D. Kostro, A. Abdulkadir, A. Durr, R. Roos, B. R. Leavitt, H. Johnson, D. Cash, S. J. Tabrizi, R. I. Scahill, O. Ronneberger, S. Klöppel, and Track-HD Investigators, "Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing," *NeuroImage*, vol. 98, pp. 405–415, Sep. 2014.

[9] L. Csató and M. Opper, "Sparse representation for Gaussian process models," in *Advances in Neural Information Processing Systems*, 2001, pp. 444–450.

[10] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Mass: MIT Press, 2006.

[11] C. M. Stonnington, G. Tan, S. Klöppel, C. Chu, B. Draganski, C. R. Jack, K. Chen, J. Ashburner, and R. S. J. Frackowiak, "Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease," *NeuroImage*, vol. 39, no. 3, pp. 1180–1185, Feb. 2008.

[12] J. L. Whitwell, W. R. Crum, H. C. Watt, and N. C. Fox, “Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging,” *AJNR Am. J. Neuroradiol.*, vol. 22, no. 8, pp. 1483–1489, Sep. 2001.

[13] H.J. Huppertz, J. Kröll-Seger, S. Klöppel, R. E. Ganz, and J. Kassubek, “Intra- and interscanner variability of automated voxel-based volumetry based on a 3D probabilistic atlas of human cerebral structures,” *NeuroImage*, vol. 49, no. 3, pp. 2216–2224, Feb. 2010.

[14] T. W. J. Moorhead, V.-E. Gountouna, D. E. Job, A. M. McIntosh, L. Romaniuk, G. K. S. Lymer, H. C. Whalley, G. D. Waiter, D. Brennan, T. S. Ahearn, J. Cavanagh, B. Condon, J. D. Steele, J. M. Wardlaw, and S. M. Lawrie, “Prospective multi-centre Voxel Based Morphometry study employing scanner specific segmentations: Procedure development using CaliBrain structural MRI data,” *BMC Med. Imaging*, vol. 9, no. 1, p. 8, May 2009.

[15] A. Abdulkadir, B. Mortamet, P. Vemuri, C. R. Jack, G. Krueger, S. Klöppel, and Alzheimer’s Disease Neuroimaging Initiative, “Effects of hardware heterogeneity on the performance of SVM Alzheimer’s disease classifier,” *NeuroImage*, vol. 58, no. 3, pp. 785–792, Oct. 2011.

[16] J. Dukart, M. L. Schroeter, K. Mueller, and The Alzheimer’s Disease Neuroimaging Initiative, “Age Correction in Dementia – Matching to a Healthy Brain,” *PLoS ONE*, vol. 6, no. 7, p. e22193, Jul. 2011.

[17] E. J., M. J., and B. A., “Using Magnetic Resonance Imaging in the Early Detection of Alzheimer’s Disease,” in *Understanding Alzheimer’s Disease*, I. Zerr, Ed. InTech, 2013.

[18] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, “The clinical use of structural MRI in Alzheimer disease,” *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, Feb. 2010.

[19] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, J. L. Cummings, M. de Leon, H. Feldman, M. Ganguli, H. Hampel, P. Scheltens, M. C. Tierney, P. Whitehouse, and B. Winblad, “Mild cognitive impairment,” *The Lancet*, vol. 367, no. 9518, pp. 1262–1270, Apr. 2006.

[20] Alzheimer’s Disease Neuroimaging Initiative, "ADNI", [online] available at: http://adni.loni.usc.edu/, accessed: 10-Jun-2015

[21] A. L. Scherzinger and W. R. Hendee, “Basic Principles of Magnetic Resonance Imaging—An Update,” *West. J. Med.*, vol. 143, no. 6, pp. 782–792, Dec. 1985.

[22] J. A. Pope, *Medical Physics: Imaging*. Heinemann, 1999.

[23] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*, 2 edition. Cambridge, UK ; New York: Cambridge University Press, 2007.

[24] E. J., M. J., and B. A., “Using Magnetic Resonance Imaging in the Early Detection of Alzheimer’s Disease,” in *Understanding Alzheimer’s Disease*, I. Zerr, Ed. InTech, 2013.

[25] J. L. Whitwell, “Voxel-Based Morphometry: An Automated Technique for Assessing Structural Changes in the Brain,” *J. Neurosci.*, vol. 29, no. 31, pp. 9661–9664, Aug. 2009.

[26] DN Greve, " An absolute beginner's guide to surface- and voxel-based morphometric analysis", Proc Intl Soc Mag Reson Med 2011

[27] G. F. Busatto, B. S. Diniz, and M. V. Zanetti, “Voxel-based morphometry in Alzheimer’s disease,” *Expert Rev. Neurother.*, vol. 8, no. 11, pp. 1691–1702, Nov. 2008.

[28] J. Ashburner and K. J. Friston, “Voxel-Based Morphometry—The Methods,” *NeuroImage*, vol. 11, no. 6, pp. 805–821, Jun. 2000.

[29] A. Mechelli, C. J. Price, K. J. Friston, and J. Ashburner, “Voxel-Based Morphometry of the Human Brain: Methods and Applications,” *Curr. Med. Imaging Rev.*, vol. 1, no. 2, pp. 105–113, Jun. 2005.

[30] C. E. Rasmussen , H Nickisch, "The GPML toolbox version 3.1", 2011, Available: www.gaussianprocess.org, accessed on: 10-March-2015

[31] J. Ashburner, “VBM tutorial,” *Tech RepWellcome Trust Cent. Neuroimaging Lond. UK*, 2010.

[32] J. Ashburner and K. J. Friston, “Image segmentation,” *Hum. Brain Funct.*, pp. 695–706, 2003.

[33] H. K.-F. Mak, Z. Zhang, K. K.-W. Yau, L. Zhang, Q. Chan, and L.-W. Chu, “Efficacy of Voxel-Based Morphometry with DARTEL and Standard Registration as Imaging Biomarkers in Alzheimer’s Disease Patients and Cognitively Normal Older Adults at 3.0 Tesla MR Imaging,” *J. Alzheimers Dis.*, vol. 23, no. 4, pp. 655–664, Jan. 2011.

[34] R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis*. Cambridge University Press, 2011.

[35] W.-L. Luo and T. E. Nichols, “Diagnosis and exploration of massively univariate neuroimaging models,” *NeuroImage*, vol. 19, no. 3, pp. 1014–1032, Jul. 2003.

[36] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak, “Statistical parametric maps in functional imaging: A general linear approach,” *Hum. Brain Mapp.*, vol. 2, no. 4, pp. 189–210, Jan. 1994.

[37] N. H. Timm, Ed., *Applied Multivariate Analysis*. New York, NY: Springer New York, 2004.

[38] L. Moutinho and G. D. Hutcheson, *The SAGE Dictionary of Quantitative Management Research*. SAGE, 2011.

[39] X. Zhu, "Linear Regression", CS731 Spring 2011 Advance Artificial Intelligence, University of Wiconsin-Madison

[40] M. M. Kamel and S. F. Aboud, "Ridge Regression Estimators with the Problem," *Appl. Math. Sci.*, vol. 7, no. 50, pp. 2469–2480, 2013.

[41] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[42] G. R. Pasha and M. A. Shah, "Application of ridge regression to multicollinear data," *J. Res. Sci.*, vol. 15, no. 1, pp. 97–106, 2004.

[43] J. Glascher , D. Gitelman (2008) Contrast weights in flexible factorial design with multiple groups of subjects, Div. of Humanities and Social Science, Caltech, Dept. of Neurology and Radiology, Northwestern, 2008.

[44] T. Mengesha, "Gram-Schmidt Orthogonalization", Pennsylvania State University, 2010

[45] C. H. Mason and W. D. Perreault Jr, "Collinearity, power, and interpretation of multiple regression analysis," *J. Mark. Res.*, pp. 268–280, 1991.

[46] J.B. Poline, F. Kherif, C. Pallier, and W. Penny, "Contrasts and classical inference," *Stat. Parametr. Mapp. Anal. Funct. Brain Images Anal. Funct. Brain Images*, 2011.

[47] T. Davis, K. F. LaRocque, J. A. Mumford, K. A. Norman, A. D. Wagner, and R. A. Poldrack, "What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis," *NeuroImage*, vol. 97, pp. 271–283, Aug. 2014.

[48] J. Ashburner, C. Chu, A. Marquand, J. Mourao-Miranda, J. M. Monteiro, A. Morgado, C. Phillips, J. Richiardi, J. Rondina, M. J. Rosa, and others, "The PRoNTo Development Group," 2013.

[49] P. Dattalo, *Analysis of Multiple Dependent Variables*. Oxford University Press,2013.

[50] J. Schrouff, M. J. Rosa, J. M. Rondina, A. F. Marquand, C. Chu, J. Ashburner, C. Phillips, J. Richiardi, and J. Mourão-Miranda, "PRoNTo: Pattern Recognition for Neuroimaging Toolbox," *Neuroinformatics*, vol. 11, no. 3, pp. 319–337, Feb. 2013.

[51] K. Kapitanova and S. H. Son, "Machine Learning Basics," *Intell. Sens. Netw. Integr. Sens. Netw. Signal Process. Mach. Learn.*, vol. 13, 2012.

[52] A. Marquand, J. M. Rondina, J- Morão-Miranda, V. Rocha-Rego, V. Giampietro, "Pattern Recognition of Brain Imaging data Toolbox (PROBID)", available at: http://www.kcl.ac.uk/ioppn/depts/neuroimaging/research/imaginganalysis/index.aspx

[53] R.A. Davis, "Gaussian Processes", *Encyclopedia of Environmetrics, Section on Stochastic Modeling and Environmental Change*, (D. Brillinger, Editor), Wiley, New York, 2001

[54] E. L. Snelson, “Flexible and efficient Gaussian process models for machine learning,” Citeseer, 2007.

[55] D. Duvenaud, “Automatic model construction with Gaussian processes,” University of Cambridge, 2014.

[56] M. Ebden, "Gaussian Process for Regression: A quick introduction", University of Oxford, 2008

[57] J. Melo, “Gaussian Processes for regression: a tutorial.”, Faculty of Engineering, University of Porto, Portugal,

[58] H. Nickisch and C. E. Rasmussen, “Approximations for Binary Gaussian Process Classification,” 2008. [Online]. Available: http://eprints.pascal-network.org/archive/00005312/. [Accessed: 21-Apr-2015].

[59] C.E. Rasmussen, “Gaussian Processes Covariance Functions and Classification,” Technical report, 2006.

[60] J. M. Lobo, A. Jiménez-Valverde, and R. Real, “AUC: a misleading measure of the performance of predictive distribution models,” *Glob. Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, 2008.

[61] S. Halligan, D. G. Altman, and S. Mallett, “Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach,” *Eur. Radiol.*, vol. 25, no. 4, pp. 932–939, Apr. 2015.

[62] P. C. O’Brien and M. A. Shampo, “Statistics for clinicians. 5. One sample of paired observations (paired t test),” *Mayo Clin. Proc.*, vol. 56, no. 5, pp. 324–326, May 1981.

[63] M. Ojala and G. C. Garriga, *Permutation Tests for Studying Classifier Performance*. .

[64] M. D. Smucker, J. Allan, and B. Carterette, “A comparison of statistical significance tests for information retrieval evaluation,” in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 623–632.

[65] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[66] B. Bostanci and E. Bostanci, “An evaluation of classification algorithms using Mc Nemar’s test,” in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, 2013, pp. 15–26.

[67] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, Feb. 2014.

[68] J. Schrouff, J. Cremers, G. Garraux, L. Baldassarre, J. Mourao-Miranda, and C. Phillips, "Localizing and Comparing Weight Maps Generated from Linear Kernel Machine Learning Models," in *2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2013, pp. 124–127.

[69] A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, and J. Mourão-Miranda, "Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes," *NeuroImage*, vol. 49, no. 3, pp. 2178–2189, Feb. 2010.

[70] J. E. Peelle, R. Cusack, and R. N. A. Henson, "Adjusting for global effects in voxel-based morphometry: gray matter decline in normal aging," *NeuroImage*, vol. 60, no. 2, pp. 1503–1516, Apr. 2012.