Imaging biomarkers extraction and classification for Prion disease

Liane dos Santos Canas

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Medical Physics and Biomedical Engineering
University College London

April 28, 2020

I, Liane dos Santos Canas, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

To the ones that dare to dream. But, mainly, to the ones that have the strength to support dreamers.

Pai, Mãe e Inês

Abstract

Prion diseases are a group of rare neurodegenerative conditions characterised by a high rate of progression and highly heterogeneous phenotypes. Whilst the most common form of prion disease occurs sporadically (sporadic Creutzfeldt-Jakob disease, sCJD), other forms are caused by inheritance of prion protein gene mutations or exposure to prions. To date, there are no accurate imaging biomarkers that can be used to predict the future diagnosis of a subject or to quantify the progression of symptoms over time. Besides, CJD is commonly mistaken for other forms of dementia. Due to the large heterogeneity of phenotypes of prion disease and the lack of a consistent spatial pattern of disease progression, the approaches used to study other types of neurodegenerative diseases are not satisfactory to capture the progression of human form of prion disease.

Using a tailored framework, I extracted quantitative imaging biomarkers for characterisation of patients with Prion diseases. Following the extraction of patient-specific imaging biomarkers from multiple images, I implemented a Gaussian Process approach to correlated symptoms with disease types and stages. The model was used on three different tasks: diagnosis, differential diagnosis and stratification, addressing an unmet need of automatically identify patients with or at risk of developing Prion disease.

The work presented in this thesis has been extensively validated in an unique Prion disease cohort, comprising both the inherited and sporadic forms of the disease. The model has shown to be effective in the prediction of this illness. Furthermore, this approach may have used in other disorders with heterogeneous imaging features, being an added value for the understanding of neurodegenerative diseases.

Lastly, given the rarity of this disease, I also addressed the issue of missing data and the limitations raised by it.

Overall, this work presents progress towards modelling of Prion diseases and which computational methodologies are potentially suitable for its characterisation.

Impact

The methods reported in this thesis have the potential of improving the diagnosis and understanding of prion diseases, in particular, the inherited form of this illness. The diagnosis of prion is challenging while the patient is alive due to the absence of specific imaging and non-imaging biomarkers to characterise prion disease. Consequently, this disease is often mistaken for other neurodegenerative diseases, which leads to a high misdiagnosis rate, hampering the collection of relevant data.

The extraction of imaging and non-imaging biomarkers to identify prion disease is complex. The heterogeneity of the symptoms, as well as the limited sample sizes highly impact the search for spatial and temporal brain patterns that can anticipate the clinical onset. The work presented in this thesis has shown potential to be used in the clinical environment. Specifically, this work presents a tool to automatically identify the prion disease patients using the imaging biomarkers validated in this document. Such a tool could be used to guide prion disease patients, and their families, to the National Prion Clinic to provide the best assistance during the course of symptoms. Furthermore, these approaches could lead to a better understanding of the disease progress and the anticipation of the prodromal phase of the disease, where the treatment to delay symptoms can be more efficient. The correct identification of the prodromal stage would allow more timely and aggressive clinical trials.

The proposed methods, namely the approaches developed to deal with heterogeneous patterns and missing data, can be transferred to the study other rare or acute diseases. These results have also promoted the creation of new projects focused on the sporadic form of prion disease, in collaboration with the clinicians in the National Prion Clinic in UK. This work has shown an impact on the scientific community through the dissemination via journals and presented in conferences of the field. Lastly, this work also has benefits outside of academia, namely as a tool for clinical use, possibly also benefiting the quality of life of patients and families of prion disease patients.

Acknowledgements

I would like to give thanks Dr Marc Modat for his contributions of time, ideas, and all the energy that he focuses on my work. He had shown me how to define which researches topics are worthy to pursue and how to efficiently multi-task. I also would like to thank him for giving me the opportunity to do this PhD, during which I undoubtedly have grown as a person and researcher.

I also would like to express my thanks to Prof. Simon Mead. He always has believed in the relevance of my research to the field, which has given me the motivation to keep working and pursuing better results in a topic so difficult as Prion diseases. Furthermore, I am thankful to the people in the National Prion Clinic, who were really attentive and spend their time to discuss results, present suggestions and guide me in this project. It is also remarkable how people, patients and their families, have been contributing to the understanding of Prion disease. Without their availability to participate and share their clinical information, this disease would be even more difficult to study and understand.

During the time of my PhD, I had the opportunity to encounter many people that despite not being directly related to my project they offer me valuable insights and discussions. Two of these people were Dr Carole Sudre and Dr Jorge Cardoso, who read my work and answer to my questions (many questions!) during the course of my PhD. I also appreciate all the support given by the people in the CMIC and TIG groups, who received me and gave me the conditions to develop my research. In particular, I would like to thank to my research group, *CoolKids*, that have made my experience in both UCL and KCL funnier and lighter. I also want to thank to Rodrigo for all the motivation and belief he gave me to start this PhD.

I would like to thanks to my closest friends in London: François, Stefano and João. These three people, being as different as possible, have been able to support and give me the personal and emotional tools that made this work possible.

Finalmente, gostaria de agradecer à minha família, Pai, Mãe e Inês, por todo o

apoio durante os últimos 4 anos. Apesar das dificuldades, eles são os responsáveis pelo meu percurso na faculdade que me trouxe até aqui. Permitiram-me emigrar e perseguir o meu sonho de investigação, ainda que isso lhes tenha custado a todos os níveis. Depois disso, foram compreensivos pela minha ausência, física e emocional e, entenderam que apesar da falta de tempo e ocasionais maus humores, os amava e estaria lá para eles, de qualquer forma. Mesmo quando quis desistir e não acreditava no sucesso deste trabalho, eles nunca duvidaram, demonstrando que os laços familiares são, sem dúvida, os mais fortes e os únicos que nos motivam a continuar, sempre.

Contents

List of Figures

List of Tables

Abbreviations

AD Alzheimer's Disease. 49, 52, 56–59, 62, 63, 67–70, 74–76, 80, 82, 84–86, 88, 89, 122, 148

ADAS-Cog Alzheimer's Disease Assessment Scale-Cognitive subscale. 68

ADNI Alzheimer's disease neuroimaging initiative. 75, 80, 82, 84, 85, 87, 88

ARD Automatic Relevance Determination. 111, 173, 179, 185

BBSI Brain Boundary Shift Integral. 52

BIC Bayesian Information Criterion. 181, 182, 184

BSE Bovine spongiform encephalopathy. 28, 34

CJD Creutzfeldt-Jakob disease. 17, 19, 28–34, 36–40, 42–46, 48, 63, 89, 91–93, 98–100, 103, 105, 108–110, 112, 113, 115, 118, 121, 124, 129, 130, 133, 134, 151, 165, 166, 168–171

CT Computed Tomography. 38, 43

DTI Diffusion tensor imaging. 52, 119

DWI Diffusion-weighted imaging. 17, 40, 43, 52, 93, 98, 111, 113, 126, 127

EEG Electroencephalogram. 31, 38

EP Expectation Propagation approximation. 74, 135, 136, 157

ERP Event-related potentials. 75

FFI Fatal familial insomnia. 28, 31, 32

FLAIR Fluid-attenuated inversion recovery imaging. 17, 39, 40, 42, 93, 100, 103, 106, 110, 113, 119, 127

fMRI Functional magnetic resonance imaging. 67

GENFI Genetic Frontotemporal Dementia Initiative. 81

GIF Geodesical Information Flows algorithm. 185

GM Grey Matter. 52, 68, 110, 193–196

GP Gaussian Process. 66, 71, 73–76, 85–88, 115, 119, 134, 151, 157, 158

GPC Gaussian Process Classification. 133, 135

GPLVM Gaussian process latent variable model. 158

GSS Gerstmann-Straussler-Scheinker syndrome. 28, 31, 32

IPD Inherited Prion Disease. 30, 31, 38, 39, 43, 48, 91, 92, 98, 100, 109, 113, 119, 121–124, 126, 127, 129, 133, 160

KL Kullback-Leibler divergence. 58

LASSO Least Absolute Shrinkage and Selection Operator. 100–102

LDA Linear Discriminant Analysis. 62

MCI Mild Cognitive Impairment. 56, 58, 62, 63, 67, 68, 70, 74, 75, 85

MD mean diffusivity. 42, 43

MEM Mixed-effects model. 76, 79, 80, 82–85

MI Mutual Information. 58

MKL Multi-Kernel Learning algorithm. 122, 126, 130

MMSE Mini Mental State Examination. 68

MRC Scale Medical Research Council Prion disease rating scale. 42, 44, 45, 92, 112

MRI Magnetic Resonance Imaging. 38, 93

mRMR Minimum Redundancy and Maximum Relevance. 56, 57

MTR Magnetization ratio transfer. 38, 39

NPMC National Prion Monitoring Cohort. 48, 92, 93, 115, 121, 160, 167

OASIS Open Access Series of Imaging Studies. 80

OPRI Octapeptide repeat insertion. 32, 94

PCA Principal Component Analysis. 57, 63

PET Positron emission tomography. 38, 43, 69

RF Random Forest. 63

ROI Region of interest. 56, 57

RVM Relevance Vectors Machines. 64, 66, 67, 70

sCJD Sporadic form of Creutzfeldt-Jakob disease. 18, 29–32, 34, 35, 39, 43, 45, 62, 63, 91, 92, 98, 100, 106, 109, 111, 113, 119, 121–124, 126, 127, 129, 144, 160, 168

SE Squared exponential covariance function. 119, 122, 154

SPECT Single-photon emission computed tomography. 38, 69

SVM Support Vector Machine. 17, 63–70, 121, 122

T1w T1 weighted MRI sequence. 38, 93, 119

T2w T2 weighted MRI sequence. 17, 38, 40

TSE Transmissible spongiform encephalopathie. 27, 28

VBM Voxel-Based Morphometry. 52

vCJD Variant form of Creutzfeldt-Jakob disease. 34, 43

WM White Matter. 52, 57, 110, 193–196

YOAD Young Onset Alzheimer's Disease. 19, 48, 94, 134, 142, 143, 149, 166

Chapter 1

Introduction

Contents

Prion diseases, also known as Transmissible Spongiform Encephalopathies (TSEs)s, are a group of progressive neurodegenerative conditions, which cause cognitive impairment and neurological deficits [2].

The infectious agent of prion diseases is critically comprised of abnormal isoforms of a protein (PrP^C) encoded by *PRNP*. The normal prion protein (PrP^C) is protease sensitive, soluble, and has a high α-helix content. The function of this protein is still uncertain. The abnormal form of the prion protein (PrP^{Sc}) is a beta-sheet rich and partially protease resistant isoform of the cellular prion protein, which accumulates mainly in the nervous system, representing the hallmark of the disease (Figure 1.1). The conversion of PrP^C to PrP^{Sc} is a post-translational event and it involves a conformational change of the protein, as detailed in Figure 1.2. Its transmission can occur by an autocatalytic mechanism [3, 4].

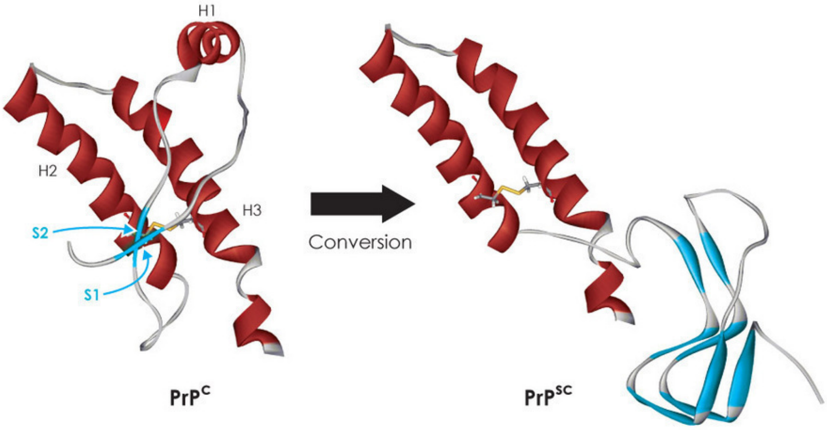


Figure 1.1: Conversion of PrP^C to PrP^{Sc}. H and S indicate α-helix and β-strand, respectively. The two β-strands S1 and S2 are proposed to "seed" β-sheet elongation as the short α-helix H1 unfolds and is converted to the PrP^{Sc} conformation. H2 and H3 remain stabilised via linkage of a disulfide bond. Image adapted from [5].

These illnesses exist in mammals, both humans and non-humans. Scrapie, a disease affecting sheep and goats, was the first prion disease to be identified in the 1730s. In more recent years other prion diseases have been seen in animals, the most common of which is bovine spongiform encephalopathy (BSE). Furthermore, TSE can also affect other mammals like minks (transmissible mink encephalopathy), deer and elk, captive wild ruminants and felines (spongiform encephalopathy) [6]. Various forms of the disease have been identified since Creutzfeldt and Jakob first described the illness later known as CJD in the 1920s. Human prion diseases are classified as CJD, Gerstmann-Straussler-Scheinker syndrome (GSS), fatal familial insomnia (FFI) and Kuru [7].

1.1 Human Prion Disease

The most common human form of Prion disease is the CJD. Prion disease presents a wide spectrum of phenotypes in part due to the different prion strains that can exist. The different phenotypes show heterogeneity in the disease duration, clinical onset, symptomatology and on its distribution of brain microstructural changes, namely the spongiosis, neuronal loss, gliosis, reactive astrocytosis and deposition of prion protein [2, 8]. Most of the CJD phenotypes are characterised by the high rate of progression, and the reduced expected time of survival after diagnosis, which varies between six weeks and three years [7]. The human forms of

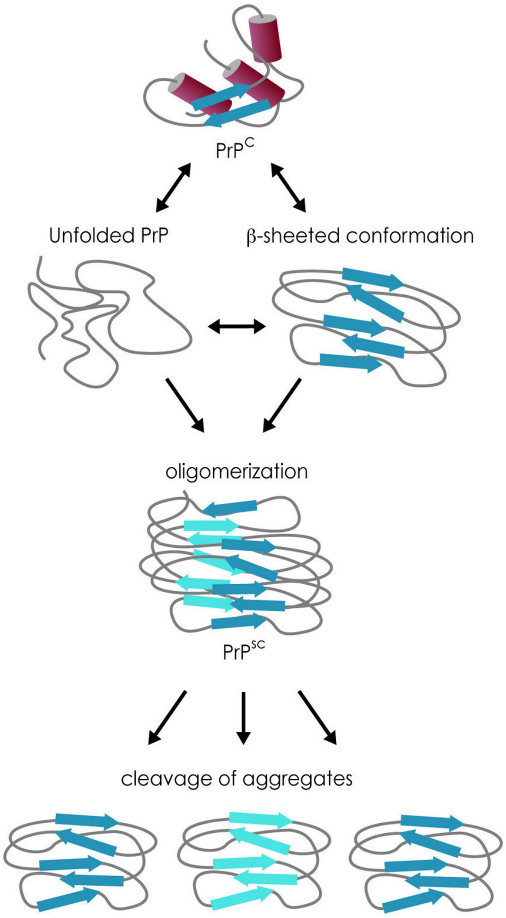


Figure 1.2: Representation of prion misfolding and transmission. β-rich *PrP* has an increased propensity to oligomerize, recruiting other β-rich monomers or unfolded *PrP*, which results in the irreversible formation of PrP^{Sc}. Subsequent cleavage of elongating fibrils leads to the propagation of infectious PrP^{Sc} "seeds". Image adapted from [5].

prion diseases may also be grouped together according to whether they are sporadic (unknown cause), inherited, or acquired (from humans or other mammals).

1.1.1 Sporadic CJD

The sporadic form is the most common among the subjects with CJD, which accounts for about 85% of the cases. This form of CJD is characterised by the occurrence of neuronal loss and the vacuolisation within cell bodies and dendrites, which gives a spongiform appearance to the cortex and deep nuclei (Figure 1.3).

The sporadic form of CJD affects both genders equally, and the average age at the clinical onset is sixty years old, being rare in people under forty years or over eighty years. sCJD has not a particular geographic incidence or a seasonal

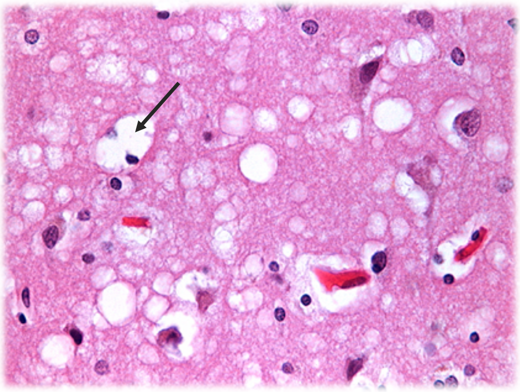


Figure 1.3: Histologic study of sporadic form of human prion disease. The brain sections reveal vacuolation that has been referred to as spongiform degeneration. Most sections also reveal significant neuronal loss [10].

clustering. The causes of infection in patients with sCJD are still unknown. The exposure to infected people does not seem to increase the risk of infection [9], hence it is hypothesised that sCJD follows exogenous infection. Tonsilar and gastrointestinal tissues containing the abnormal prion protein were found, proposing ingestion as route of infection. This hypothesis is supported by the fact that some cases have been found in case-control studies, which those seem to be related with subjects diet namely by eating brains. However, in other cases, diet can be discarded as a major source of sCJD, because lifelong vegetarians have also developed sCJD.

Finally, it has been reported that cases of sCJD originated by transmission among humans by medical procedures. Nevertheless, it is also accepted that the sporadic form of CJD may result from endogenous generation of prion, caused by a random misfolding of the prion protein, which might lead to a cascade of misfolding of normal prion protein into the pathogenic isoform. The sCJD is characterised by the rapid progression of the symptoms with prominent cognitive decline. The median time of life after clinical onset is only five months, and 90% of the patients die within one year [7, 9]. A third of the patients affected by sCJD present with fatigue, disordered sleep and decreased appetite. The second third of the cases has shown behavioural or cognitive changes; whereas the last suffers visual loss, cerebellar ataxia, aphasia, or motor deficits [6, 9].

1.1.2 Inherited Prion Disease

Inherited Prion Disease (IPD), also designated as familial prion disease, is caused by autosomal dominant inheritance of mutations in the *PRNP* gene, which in total are responsible for 10-15% of the incidence of human prion disease [7]. Currently, over thirty different mutations in *PRNP* have been found in patients presenting IPD (Figure 1.4). Although, about 95% of familial cases are caused by four point mutations (at codons 102, 178, 200 and 210) and insertions of five or six octapeptide repeats [6]. A genetic factor influences the susceptibility of an individual to develop prion disease, namely a common variation in the prion protein gene itself. A polymorphism at the codon 129 may influence the susceptibility to some IPD phenotypes. There are two possible genetic types, which in turn specify the body to produce different amino acids at this position. These amino acids are called Methionine and Valine, or M and V for short. MM and MV frequencies in the population are roughly equal (40-50%). It has been known for some years that individuals, who are MV, show a lower risk of developing prion disease than the subjects who are MM or VV [11]. The IPD has an earlier age of clinical onset when compared with sCJD, even if the range of ages of clinical onset is longer and may vary between 20 years old to 85 years old. Note also that the clinical course of IPD is longer than sCJD, which may last 5-11 years, for some mutations.

Given the heterogeneity of clinical phenotypes, IPD can be further divided in three groups: GSS, FFI and CJD. These clinical categories of IPD may be seen as extremes of phenotype, in reality the syndromes overlap considerably. The GSS syndrome is characterised by a clinical onset between 20 and 70 years old and a progressive cerebellar ataxia followed by dementia. The FFI origins a refractory insomnia, hallucinations, automatic dysfunction and dementia. This syndrome is also characterised by neuropathological changes in the thalamus, namely in the anterior ventral and mediodorsal nuclei, and the olivary nuclei [6, 7]. Lastly, some subjects affected by IPD have shown rapidly progressive dementia, with myoclonus and pseudoperiodic discharges on electroencephalogram (EEG). The most common worldwide *PRNP* mutations are E200K, D178N, P102L and OPRI, whereas in the UK the 6-OPRI mutation is the most frequently detected in *PRNP*. Table 1.1 describes the neuropsychologic profiles of these insertional mutations.

Briefly, E200K is the most common cause of IPD. The phenotype associated to E200K is highly heterogenic. Note also that, by the examination of unaffected relatives, asymptomatic mutation carriers were detected in old age, which supports the hypothesis that the penetrance is incomplete [12]. The D178N was initially reported by Medori *et al.* [13], who described this illness as a large case series of untreatable insomnia, dysautonomia and myoclonus. These symptoms are not specific to this mutation, since other mutations, such as V210I, FFI and sCJD may also present these clinical manifestations. In fact, Goldfarb *et al.* [14], established a haplotypic relationship between codon 129 and 178, whereby the mutation on a 129M chromosome leads to FFI, and the mutation on a 129V chromosome leads to familial CJD. The P102L mutation is the classic example of GSS syndrome, in which the patients have been manifested the aforementioned phenotype [15]. Other mutations are also associated with this syndrome, namely F198S, A117V, P105L, G131V, Y145X, H187R and some D178N mutations [16]. The genetic susceptibility factor has shown a small influence in the P102L phenotype, which causes an earlier clinical onset for codon homozygous cases. Finally, the octapeptide repeat insertion (OPRI) mutations assembles the insertions of more than three additional octapeptide repeats in the N-terminal region of PrP. The polymorphism at *PRNP* codon 129 is responsible for the phenotype heterogeneity, namely the different ages of onset and rate of progression of the clinical manifestations (Table 1.1). Note that the heterozygosity at codon 129 have a delayed age of onset by around a decade compared with patients homozygous at codon 129. Moreover, the disease phenotype is also influenced by the number of times that the octapeptide repeats. The degree of spongiosis and astrocytosis is higher in the cerebellum of patients with 8- or 9-OPRI mutations. On the other hand, PrP deposition was visualised by immunocytochemistry as elongated deposits in the molecular layer of the cerebellum for smaller number of replications of the octapeptide. Small OPRI mutations has also a later age of onset and shorter duration. The OPRI mutations have also shown significant behavioural changes, namely an existence of pre-morbid personality disorder characterised by criminality, aggression, delinquency and hypersexuality [6, 15].

1.1.3 Acquired CJD

The acquired CJD is caused by the transmission of infection from mammals to humans or from human to human. The acquired form of human prion disease is rare, however the transmission of these syndromes are untreatable and fatal. Several measures were introduced to decrease the risk of transmission; however, considering the prolonged incubation periods, the absence of tests to identify the infection during the incubation period and the high resistance of prions to disinfection, many cases of acquired prion disease were not anticipated. Thus, it is still crucial to developed metrics of vigilance to identify novel mechanisms of prion transmission and new

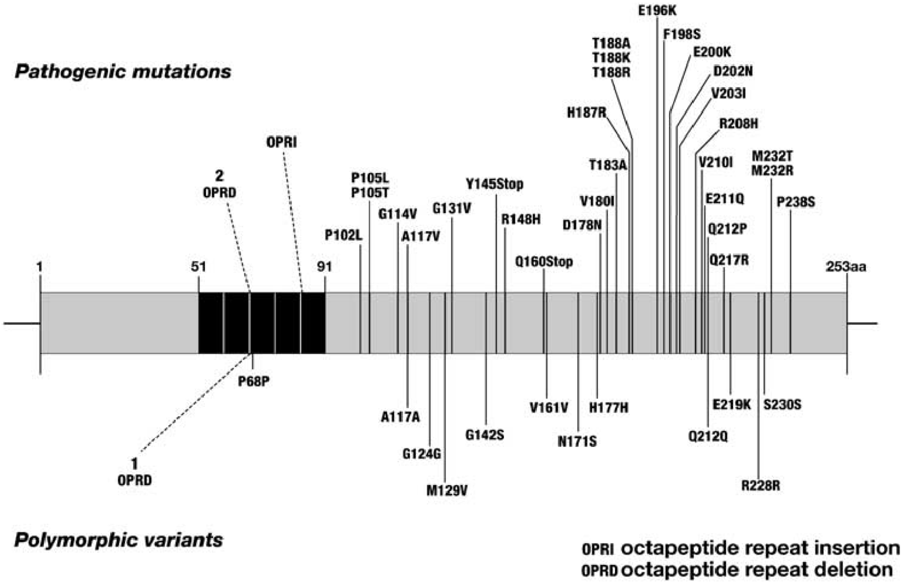


Figure 1.4: Representation of prion disease pathogenic mutations. The grey bar represents the prion protein gene, the definite or suspected pathogenic mutations are shown above this representation. Neutral or prion disease susceptibility/modifying polymorphisms are shown below. Image adapted from [7].

Table 1.1: Description of neuropsychological profile of inherited form of prion disease, caused by insertional mutation. Mean age at clinical onset in years.

Mutations	Age at clinical onset	Progression	Clinical manifestations	Duration (in months)
E200K	61 [31 - 78]	Fast	Peripheral neuropathy, supranuclear gaze palsy and sleep disturbance. Rapidly progressive dementia.	5
E196K	69 [66 - 80]	Fast	Myoclonus and pyramidal, cerebellar or extrapyramidal signs.	—
D178N	50 [20 - 72]	Fast	Untreatable insomnia, dysautonomia and myoclonus. Progressive ataxia with later dementia.	5 - 48
P102L	49 [25 - 70]	Slow	Progressive ataxia with later dementia.	48
5-OPRI	45 [26 - 61]	Slow	Cortical dementia, often with apraxia, cerebellar ataxia. Pyramidal and extrapyramidal, myoclonus, chorea, seizures.	3 - 84†
6-OPRI	34 [20 - 53]	Slow	Cortical dementia, often with apraxia, cerebellar ataxia. Pyramidal and extrapyramidal, myoclonus, chorea, seizures.	3 - 84†
A117V	39 [20 - 64]	Slow	Progressive ataxia with later dementia.	49
Y163X	30 [42 - 70]	Slow	Progressive ataxia with later dementia.	—

† The median time of survival after clinical onset is 7 years, and may vary between3 months and 21 years.

cases of infection [1]. The acquired form of CJD may be classified according to the transmission pathway and the disease phenotype.

Iatrogenic CJD

The acquired form of CJD may be denominated as iatrogenic CJD when the transmission occurs due to mechanisms of iatrogenic transmission, such as by the contact with biological material in, or adjacent to, a contaminated brain, or material used during surgical procedures. The first evidence of iatrogenic transmission of CJD took place in 1974, via a corneal transplant. Will *et al* [1], identified the total number of cases of iatrogenic CJD (Table 1.2), in which the route of inoculation has been parenteral, either by surgery or by intramuscular injection.

Variant CJD

The first article describing a new variant of CJD in the UK was published in 1996, entitled *"A new variant of Creutzfeldt-Jakob disease in the UK"*, in which it was suggested a relation between 10 cases, with unusual clinical phenotype for CJD, to the epidemic of BSE in UK [17].

The variant Creutzfeldt-Jakob disease (vCJD) presents a distinctive phenotype from the sCJD. The vCJD affects younger people (median age of onset is 29 years old), and the disease has a longer course, about 14 months. The patients suffer early psychiatric symptoms and the existence of painful sensory symptoms is common. Neuropathological studies have shown a high level of PrP deposition with many plaques of abnormal prion protein (Figure 1.5, panel B). Until recently, all the variant cases of human prion disease had been homozygous at codon 129 for methionine; however, more recently a patient, who had received blood from a donor diagnosed with vCJD, was diagnosed with vCJD, presenting a codon 129 heterozygous [17].

The hypothesis that the vCJD is a result of a BSE transmissions to humans is sustained by the common signature of this syndromes in the brain: the vCJD creates alterations in MRI in pulvinar region of thalamus, whilst the sporadic form of CJD shows alterations predominantly in the basal ganglion and putamen. The risk of transmission of vCJD by iatrogenic mechanisms is not discarded. However, there is currently no evidence of transmission of vCJD through these routes, but this does not preclude such a possibility because the incubation period could be long and the

Table 1.2: Total cases of iatrogenic CJD world-wide at 2003. Data from [1].

Mode	Cases	Incubation (years)	Clinical Manifestations
Neurosurgery	4	1.6	Visual, cerebellar and dementia
Dura matter	136	6	Visual, cerebellar and dementia
Corneal transplant	3	15.5	Dementia
Depth electrodes	2	1.5	Dementia
Human growth hormone	162	12	Cerebellar
Human gonadotrophin	5	13	Cerebellar

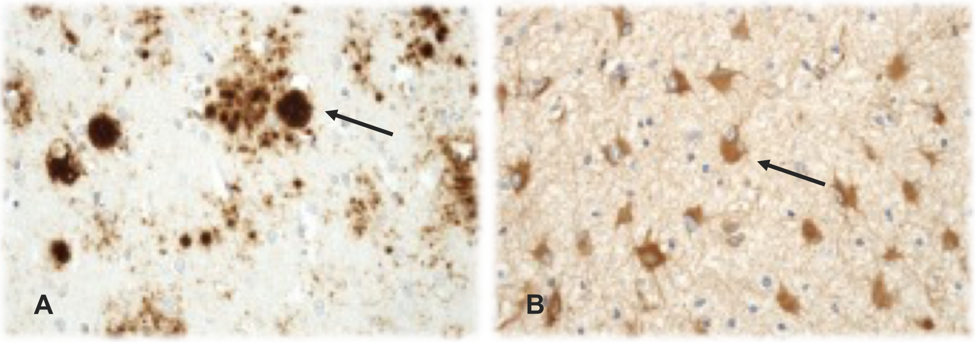


Figure 1.5: Histologic study of the variant form of human prion disease. A - Reactive astrocytosis in the thalamus, glial fibrillary acidic protein antibody; B- PrP deposition in the cerebral cortex. Image adapted from [18]

.

period of current observation is short.

Kuru

Kuru was first identified in 1957 in Okapa, Papua New Guinea. This syndrome is geographically restricted to Okapa area, and between 1957 and 2003 over 2700 cases among 30 000 people were identified. Initially, kuru affected predominantly women and children of either gender. However, the disease epidemic declined and the proportion of women and men affected became similar. In fact, the children born after 1959 are not affected by the disease and there is no evidence of vertical transmission. Endocannibalism[1] was pointed as transmission path, since this is consistent with the other aforementioned ways of transmission - ingestion of contaminated food - and the epidemic declined with the end of this practice in late 1950s. Kuru is characterised by a progressive cerebellar ataxia and cognitive changes developed only in advanced stages of disease. Conversely to sCJD, myoclonus was not registered in Kuru cases, and in the majority of patients the dementia was absent. The incubation period ranges from 4.5 years to at least 40 years (mean incubation period has been estimated to be about 12 years), and the total illness duration in adults ranged from 6-36 months.

[1] The excess of cases in females and children is consistent with the available descriptions of the rituals, since they, and not the men, ate the internal organs, in particular the brain.

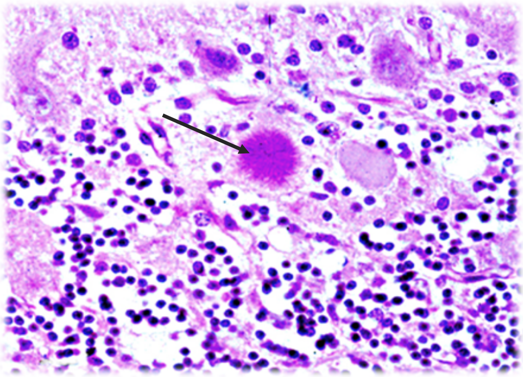


Figure 1.6: Histologic study of the acquired form of human prion disease, namely Kuru syndrome. Section of the patient's cerebellum. The black arrow points a Kuru plaque. Image adapted from [10].

1.2 Clinical Challenges

Currently, the growing interest in the human form of prion diseases is not caused by the increasing incidence, since its incidence is now considered stable at between 0.5 and 2.5 cases per million people per year. The interest on CJD concerns the nature of the transmissible agents, the unpredictable species barriers, the variable distribution of infectivity in tissues and strain variations found, which leads to the broad spectrum of phenotypes and incorrect diagnoses. As aforementioned, Prion disease presents a wide spectrum of phenotypes due to the different prion strains that can exist. The different phenotypes show heterogeneity in the disease duration, clinical onset, symptomatology and on its distribution of brain microstructural changes, namely the spongiosis, neuronal loss, gliosis, reactive astrocytosis and deposition of prion protein [2, 8]. Consequently, the clinical diagnosis of both forms of CJD can be challenging during life, particularly in the earlier stages of the disease as the different phenotypes can mimic other neurodegenerative diseases. An autopsy study found that 40% of cases of neuropathological prion disease were undiagnosed while alive [7]. Whilst the definitive diagnosis is still only possible by brain biopsy, the improved understanding of the pathogenesis of prion diseases have allowed definition of recognisable clinical features and replicable diagnostic criteria *in vivo* [19]. The diagnosis criteria are currently based on a set of neurological, cognitive and psychiatric observations [20, 21]. Recently, an updated diagnostic algorithm was suggested by Manix *et al.* [22], in which the procedures mentioned above provide useful biomarkers that support the need for histopathological tests and the subjects diagnosis.

To date, there is no proven cure for CJD, but clinical studies are underway to investigate possible treatments. Even without a cure, it is crucial to address the current rate of misdiagnosis of CJD cases, in order to increase the sample size and conveniently perform studies to better understand the mechanisms of the disease and possible treatments. By identifying the patients with CJD or at risk of developing symptoms, the recruitment to clinical trials will become easier and timely appropriate to test the *in-silico* treatments. Furthermore, several drugs have delayed the onset of disease in laboratory animals if given before symptoms start, even if none have halted or cured the disease once animals are unwell [11]. Therefore, two main research pathways are aroused by the current clinical challenges of dealing with CJD:

1. Investigation of possible ways to delay and maintain the symptoms of CJD through drug trials;

2. Investigation of new diagnostic criteria that allow the current misdiagnosis rate to be addressed.

Note however that the efficacy of the clinical trials is intimately linked with a timely diagnosis of CJD, specially the identification of prion infection before the clinical onset. Sandberg *et al.* [23], has performed a study in animals that actually have proven that prion propagation is preceded by a clinically silent exponential phase. The propagation phase is not rate-limited by prion protein concentration, which rapidly reaches a maximal prion titre, immediately followed by a distinct switch to a plateau phase after which the clinical symptoms start [23]. It is hypothesised that the treatments can be more effective during this period of incubation, hence the need for new ways to identify microstructural brain changes to recognise the prion infection before clinical onset. Therefore, most of the research developed in the context of CJD aims to identify useful biomarkers to diagnose and characterise this disease. The search for biomarkers is currently moving towards brain imaging data, given their sensitivity and non-invasive nature, but it is also focused in other clinical metrics such as functional and behavioural scales. The following section highlights the most recent attempts to effectively identify novel biomarkers to diagnose the CJD.

1.3 Diagnosis of Prion disease

Several measures have been studied in order to improve the sensitivity and specificity of CJD diagnosis during life; namely noninvasive techniques such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET) or single-photon emission computed tomography (SPECT); furthermore, the detection of the 14-3-3 protein in cerebrospinal fluid (CSF) and periodic sharp wave complexes on the EEG provides important corroborative evidence for the clinical diagnosis [22, 24–26]. Although these metrics may not diagnose CJD with a high degree of certainty, in particular for IPD patients, these procedures nonetheless provide insight into the pathophysiologic aspects of the disease.

1.3.1 Imaging Data

The use of neuroimaging techniques to investigate the CJD symptoms is being applied to achieve higher accuracy in its diagnosis, even in early stages, by evidencing the main brain changes, such as neuronal loss, spongiform change and reactive astrocytosis in the absence of an inflammatory reaction. [27].

Structural Magnetic Resonance Imaging

Three dimensional T1 weighted (T1w) and T2w are acquired to evaluate the structural changes in the brain. Either MRI sequences allows qualitative and quantitative analysis to be performed in order to identify longitudinal and cross-sectional anatomical changes due to CJD. The disease progression can be evaluated through a qualitative assessment of T2w images, specifically by identifying the hyperintensities in the cortex; nevertheless, the T2w allows also a quantitative assessment of the illness severity by T2 relaxometry maps. Using T2w images, Barboriak *et al.*, found high signal in the basal ganglia, thalamus and in the cortex.[28]

Uemura *et al.* [29], also found lesions in the pallidum and white matter, whereas Schroter *et al.* [24], found hyperintensities in the cerebellum (Figure 1.7). In 2004, Matsusue and collaborators [30], demonstrated the presence of lesions in cerebellar grey matter, followed by lesions in the cerebellar white matter around lateral ventricles and brainstem [24, 29, 30]. Siddique and collaborators [8], investigated the cross-sectional, longitudinal and post-mortem cerebral magnetization ratio transfer (MTR) as a surrogate for prion disease pathology. They found highly significant correlations between MTR and prion disease ($p < 0.01$).

Vita *et al.* [31], using quantitative T2w images, have shown gray and white matter atrophy in patients with the inherited form of CJD, namely in the parahippocampal gyrus, mid-orbital gyrus, superior temporal gyrus, insular cortex, middle cingulate, supramarginal gyrus and post-central gyrus. In the same study, Vita *et al.* [32], also detected gray matter atrophy in sCJD patients. However, these alterations were not overlaping with the structures detected in patients with IPD, using the same measure. The sCJD patients have shown alterations in the putamen, thalamus and anterior limb of the internal capsule, as well as in the innumerous cortical areas, such as the anterior cingulate cortex, rectal gyrus, Heschl gyrus, superior temporal gyrus, middle frontal gyrus, middle cingulate cortex and fusiform gyrus [31, 32].

Alner *et al.* [15], explored the potential to use the cortical thickness as a biomarker to characterise inherited form of prion disease, specifically 6-OPRI and P102L. The results of this study have shown significant differences in the mean cortical thickness between 6-OPRI patients and controls in temporal, cingulate, frontal, parietal and occipital lobes. On the other hand, only the mean cortical thickness of parietal lobe was relevant to distinguish controls and P102L patients [15].

Later, Vita and collaborators [33], also detected significant differences between controls and symptomatic subjects using MTR, namely in the caudate, hippocampus, putamen and cortex. Brain progressive structural changes were also identified by applying longitudinal voxel-based morphometry (VBM): significantly greater rates of grey matter decline were observed, predominantly in the pons, the corpus callosum, the thalamus and the putamen, when comparing controls and symptomatic subjects [33–35].

Fluid-Attenuated Inversion Recovery Imaging (FLAIR)

Several studies have explored the potential of FLAIR in the detection and evaluation of human prion disease. Murata *et al.* [36], studied 13 patients diagnosed with CJD and detected signal abnormalities in the basal ganglia, thalamus and cortex, using FLAIR images. Similarly, Collie and collaborators [27], found hyperintensities in the pulvinar, denominated as pulvinar sign, caudate, periacqueductal gray matter and mediodorsal thalami nuclei [27, 36].

Kallenberg and collaborators [37] (Figure 1.8, panel A), as well as Young *et*

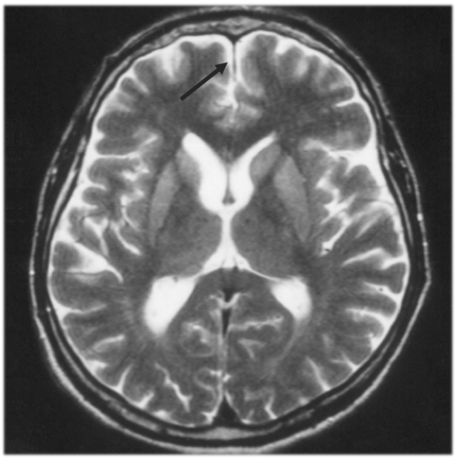


Figure 1.7: T2w magnetic image of a male subject with CJD. Hyperintensities in the basal ganglia and frontal cortex (black arrow). Image adapted from [24]

.

al. [38] (Figure 1.8, panels B and C), diagnosed CJD using FLAIR images, identifying hyperintensities in the cingulate cortex [37, 38]. The FLAIR imaging technique was used in numerous studies of human prion disease (Table 1.3), in which DWI and T2w images were also acquired, once the combined use of these techniques has proven to be an added value in the diagnostic of human prion disease. However, comparing the rate of correct diagnostic obtained using FLAIR, T2w and DWI, it is possible to infer that the FLAIR is more accurate to detect the earlier stages of the disease; nonetheless, FLAIR is less accurate in detecting lesions when compared with DWI, once in same cases the hyperintensities become less prominent during the course of the disease, as it was shown by Shiga *et al.* [38, 39].

Diffusion-Weighted Imaging

DWI is a medical imaging technique that exploits the exquisite sensitivity of magnetic resonance imaging to diffusion processes to measure microscopic tissue orientation characteristics *in vivo*. DWI characterises the three-dimensional diffusion of water as a function of spatial localisation. This technique is highly sensitive to the changes in the diffusion pattern allowing the assessment of the microstructural architecture of brain tissues. Consequently, DWI has became the most relevant and sensitive sequence to detect and characterise human prion disease, namely the sporadic form of this illness. In fact, the combined use of DWI and FLAIR sequences has increased the diagnostic specificity and sensitivity up to 91%, compared with previous studies only using T1w and T2w MRI [53]. The high-intensity lesions, in

Table 1.3: Prospect of the studies analysed by Caobelli and colaborators until 2014. For each studied are presented the main regions affected. BG: basal ganglia, TH: thalamus, CO: cortex, CE: cerebellum, P: pallidus, WM: white matter, PU: pulvinar, BS: brainstem, H: hippocampus. Table adapted from[26].

Authors	**Year**	**Patients**	**Modality**	**MR images**	**Brain regions**
Murata T *et al.* [36]	2002	13	MRI	DWI, FLAIR	BG, TH, CO
Demaerel P *et al.* [40]	2003	5	MRI	DWI, FLAIR	BG, TH, CO
Tschampa HJ *et al.* [41]	2003	6	MRI	T2, DWI, FLAIR	BG, TH, PU
Collie DA *et al.* [42]	2003	86	MRI	FLAIR	BG,PU
Shiga Y *et al.* [39]	2004	36	MRI	T2, DWI, FLAIR	BG, TH, CO
Young GS *et al.* [38]	2005	40	MRI	T2, DWI, FLAIR	BG, TH, CO, CE, WM
Ukisu R *et al.* [43]	2005	9	MRI	T2, DWI, FLAIR	BG, TH, CO
Kallenberg *et al.* [37]	2006	157	MRI	T2, DWI, FLAIR	BG, TH, CO, P
Tschampa HJ *et al.* [44]	2007	39	MRI	DWI, FLAIR	BG, TH, CO
Meissner B *et al.* [45]	2008	55	MRI	DWI, FLAIR	H, CE, BG, TH, CO
Fulbright RK *et al.* [46]	2008	15	MRI	DWI, FLAIR	BG, CO
Shimono T *et al.* [47]	2008	7	MRI	DWI, FLAIR	BG, TH, CO
Kransnianki A *et al.* [48]	2008	63	MRI	DWI, FLAIR	BG, CO, PU, CE
Manners DN *et al.* [49]	2009	10	MRI	T2, DWI, FLAIR	BG, TH, CO
Hyare H *et al.* [50]	2010	17	MRI	DWI, FLAIR	BG, TH, CO
Talbott SD *et al.* [51]	2011	3	MRI	DWI, FLAIR	BG, CO
Vitali P *et al.* [52]	2011	54	MRI	DWI, FLAIR	BG, TH, CO, P, H

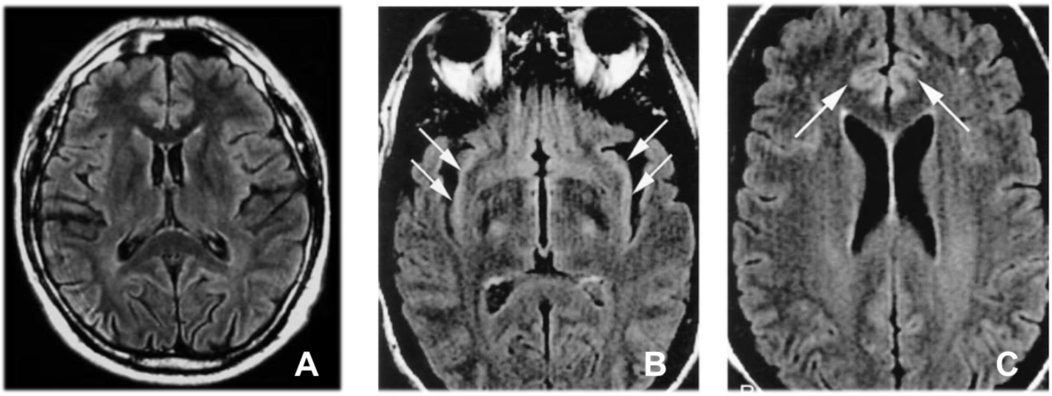


Figure 1.8: CJD detection using FLAIR images. CJD detection using FLAIR images. The image A shows a FLAIR image with abnormalities in the left hemisphere and the dorsal part of the cingulate gyrus [37]. The image B, an axial FLAIR, shows the insular cortex slightly hyperintense (arrows) to neocortex. The image C, axial FLAIR image, shows relative hyperintensity in cingulate cortex [38].

either DWI and FLAIR, appear before any sign of brain atrophy. Kallenberg and collaborators [37], identified typical lesions, caused by CJD, visible in DWI images (Figure 1.9).

Aside from the qualitative assessment of DWI performed to detect signal abnormalities due to the presence of CJD, it is also useful to use measures derived from DWI, such as diffusion tensor imaging (DTI), to describe the progress of the disease and infer its severity. Hyare *et al.* [50], calculated the mean apparent diffusion coefficient (ADC) for the caudate, putamen and pulvinar nuclei to evaluate the possibility of using DWI MRI as an imaging biomarker of disease severity. The results seem to suggest that the brain volume loss in inherited prion diseases is followed by the increased cerebral ADC, correlating with the increased disease severity [50, 54]. Hyare *et al.* [55], extracted DTI measures, such mean diffusivity (MD) to evaluate the relevance of the diffusion patterns in the putamen as useful biomarker to diagnose CJD. Their results suggest that the putamen radial diffusivity has potential as a secondary outcome measure in future therapeutic trials in human prion diseases, since it provides useful information regarding brain changes caused by CJD. Moreover, a step-wise linear regression analysis, with dependent variable decline in clinical rating scale and covariates age, gender and disease duration, showed the decline in putamen radial diffusivity was the strongest predictor of decline in Medical Research Council Scale (MRC) ($p < 0.001$) [55]. Nevertheless, the DTI measures, specifically MD measurements, can show either increased or decreased values depending on the brain region and the micro-structural changes happening at a specific stage of the

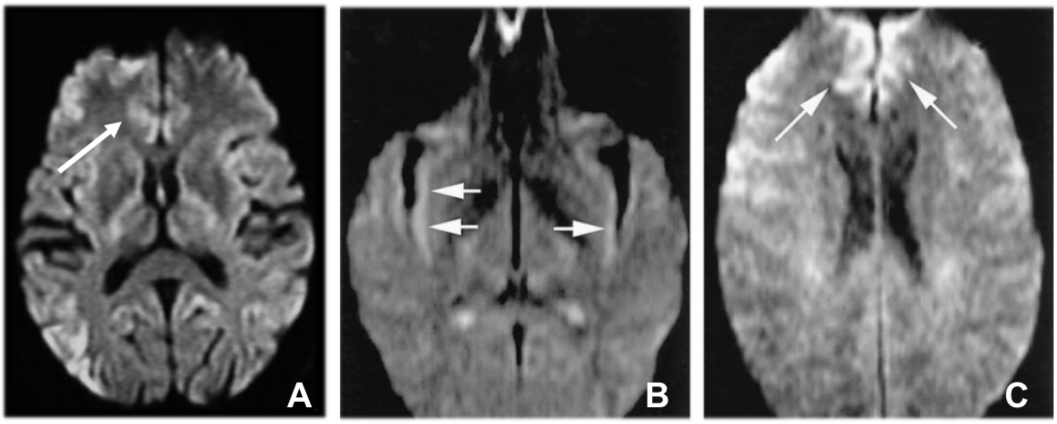


Figure 1.9: CJD detection using DWI images. The image A shows a DWI image with abnormalities in the left hemisphere and the dorsal part of the cingulate gyrus [37]. The image B, an axial DWI, shows the insular cortex slightly hyperintense (arrows) to neocortex. The image C shows hyperintensity in cingulate cortex [38].

disease. These biphasic behaviour is independent of the CJD subtype or the *PRPN* mutation. As an example, Vita *et al.* [56], reported increased cerebral MD in patients with the 6-OPRI mutation also previously found in patients with other forms of CJD, namely in cerebral cortex of patients with the E200K mutation and in the thalamus of vCJD patients. In this study, it was hypothesised that the increased of signal reflected increased gliosis. Conversely, findings of decreased MD have been reported in both sCJD and IPD patients, specifically within the basal ganglia and thalamus, which was assumed to be a result of spongiform changes [56, 57]. Furthermore, these changes in the diffusivity patterns can also be explained by macroscopic brain changes as atrophy. In detail, the increased diffusivity has been associated with loss of neuronal cell bodies, synapses, and dendrites, causing an expansion of the extracellular space and, consequently, a more evident diffusivity. In prion diseases, this increasing in diffusivity can also be explained by gliosis processes followed by neuronal loss, which become dominant over spongiform changes [56].

Other imaging modalities

Whilst MR imaging presents hyperintensities, as an expression of vacuolitic process happening in the brain, and/or structural changes such as volume loss, the 18F-fluorodeoxyglucose (FDG) PET–CT shows the hypometabolism, which may either represent a consequence of neuronal damage, or neuronal loss, or even a result of dedifferentiation in certain brain areas [26]. 18F-FDG PET–CT has proved to be useful to detect CJD in its early stages; although, the brain alterations shown by this technique in the initial stage of CJD are also found in a large variety of neurodegenerative diseases [58]. Conversely, the hyperintensities spotted in MR images are highly supportive for a diagnosis of CJD, specially in MR modalities such as fluid-attenuated inversion recovery imaging and diffusion weighted imaging. For this reason, the data available in the current study is mainly MRI data.

1.3.2 Non-imaging data

The clinical diagnosis of CJD is not exclusively reached using imaging data. In fact, rating scales designed to probe neurological, cognitive, psychiatric and general functional status have been used to identify the clinical onset and to track the evolution of clinical symptoms. Nowadays, a clinical scale, MRC Scale, is used by the National Prion Clinic to characterise the different stages of CJD condition and to evaluate the effectiveness of trials and future treatments. The scale was developed by Thompson and collaborators to tackle the lack of a validation measure of clinical progression during the PRION-1 trial [59].

Analysis of the performance of eight scales, including Rating scales designed to probe neurological, cognitive, psychiatric and general functional status, in PRION-1 in terms of validity, practicality and statistical power in simulated clinical trials supported the need of orientated measures relative to global, neurological, cognitive or psychiatric scales. However, concerning the pathological features of CJD – namely the rapid progression of the disease which may not be well described due to the high decline of the patients between visits, floor effects, such as large numbers of patients with very low score, in most of the scales, except Glasgow Coma Score –, and the size of the sample available to produce reproducible analysis of the different forms of CJD, the scales available, and previously applied to other types of dementia, were not completely successful when used in prion disease [59]. The MRC Scale is a single, functionally-orientated and validated outcome measure, designed especially for the demands of the prion disease clinical trial. It combines elements of three rating scales, which had shown to be useful in PRION-1 analysis: the Modified Barthel Activities of Daily Living Index (Barthel), the Clinical Dementia Rating Sum of Boxes, and the Glasgow Coma Score [59–61]. The MRC Scale is defined from 20 to 0, where 20 denotes the clinically asymptomatic subjects and 0 corresponds to the most severe symptoms.

The final scale evaluates physiological functions, the autonomy of the patient, and its cognitive decline and it has been shown to be useful to characterise the progression of the disease for the most forms of CJD. In fact, the MRC Scale was able to summarise into a single outcome measure the patients symptoms and to capture rapid global decline in a subject with sCJD. The MRC Scale is not able to capture changes in patients with slowly progressive forms, namely with the inherited form of human prion disease. The MRC Scale does not capture the cognitive decline during the disease progression. To overcome this limitation, which compromises the accurate definition of the onset of disease in inherited prion disease, it is essential the joint analysis of neuropsychological tests [59].

Finally, both spinal fluid tests (cerebrospinal fluid), and blood tests are used as diagnostic tests. They look for a wide range of biochemical and metabolic disorders, vitamin deficiencies, signs of viral infection, specific antibody tests and thyroid problems, that can recognise the presence of prion disease [11]. Genetic tests may also be performed to investigate changes in the DNA causing inherited disorders. The identification of genetic abnormalities leading to CJD can be used to identify subjects at risk of developing prion disease. The recruitment of these subjects' into clinical trials is critical to understand the mechanisms of the disease and the evolution of the biomarkers before and after onset. Furthermore, by tracking the brain changes that anticipate the clinical onset, new metrics for more effective diagnosis can be considered.

1.4 Research Objectives

As detailed before, identifying the specific time of onset for each patient might lead to delay many of the symptoms associated with this disease. Also, the accurate knowledge of the disease progression pattern and the anticipation of clinical symptoms might lead to more aggressive clinical trials, which can contribute to the surveillance of symptoms and improvement of the life quality of prion diseases patients. To overcome the current misdiagnosis rate and to address the research limitations that hamper the understanding of human prion diseases, this project aims to identify and extract quantitative imaging biomarkers that can be used to characterise the evolution of the disease over time. In addition, I also investigated the potential of using the extracted imaging biomarkers to identify CJD among other types of dementia, and how to differentiate its subtypes.

The following sections describe the main contributions of this thesis that address these research objectives.

1.4.1 Thesis Contributions

The main contributions of this thesis are organised in three main parts, as following:

1. Extraction and selection of imaging biomarkers - I developed a tailored framework to extract quantitative imaging biomarker to identify and characterised CJD. To my knowledge, this is the first study that takes advantage of complementary information, extracted from three different MRI sequences, to identify subject-specific imaging biomarkers. As a result, two research outputs were presented, demonstrating the viability of the extracted biomarkers.

   (a) **Liane S. Canas**, Benjamin Yvernault, Carole Sudre, Enrico De Vita, M. Jorge Cardoso, John Thornton, Frederik Barkhof, Sébastien Ourselin, Simon Mead, Marc Modat, "Imaging biomarkers for the diagnosis of Prion disease," Proc. SPIE 10574, Medical Imaging 2018: Image Processing, 1057405 (2 March 2018);

   (b) Harpreet Hyare, Enrico De Vita, Marie-Claire Porter, Ivor Simpson, Gerard Ridgway, Jessica Lowe, Andrew Thompson, Chris Carswell, Sébastien Ourselin, Marc Modat, **Liane Dos Santos Canas**, Diana Caine, Zoe Fox, Peter Rudge, John Collinge, Simon Mead, John S Thornton, "Putaminal diffusion tensor imaging measures predict disease severity across human prion diseases", Brain Comunications, *Published April 2020*

2. Modelling of imaging biomarkers through Gaussian Processes - In order to characterise the disease status of each subject using the available multi-source features, I designed a Bayesian framework to find the function that best fits the relationship between imaging features and the subjects diagnosis. The model was used as a proxy to binary diagnosis, differential diagnosis and subjects stratification. The results of this model have been submitted as a journal contribution, demonstrating the relevance and novelty of using machine learning models to diagnose CJD. Lastly, the methodological advances obtained from the design of the model were presented in two peer-reviewed conferences.

(a) **Liane S Canas**, Benjamin Yvernault, Carole H Sudre, Jorge Cardoso, John Thornton, Frederik Barkhof, Sébastien Ourselin, Simon Mead, Marc Modat. "Multikernel Gaussian Processes for patient stratification from imaging biomarkers with heterogeneous patterns". In Learning from Limited Labeled Data: Weak Supervision and Beyond, NIPS, Long Beach, 2017

(b) **Liane S. Canas**, Benjamin Yvernault, David M. Cash, Erika Molteni, Tom Veale, Tammie Benzinger, Sébastien Ourselin, Simon Mead, Marc Modat, "Gaussian processes with optimal kernel construction for neurodegenerative clinical onset prediction," Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, 105750G (27 February 2018);

(c) **Liane S Canas**, Carole H Sudre, Enrico De Vita, Akin Nihat, Tze How Mok, Catherine F Slattery, Ross W Paterson, Alexander J M Foulkes, Harpreet Hyare, M Jorge Cardoso, John Thornton, Jonathan M Schott, Frederik Barkhof, John Collinge, Sébastien Ourselin, Simon Mead, Marc Modat, "Prion disease diagnosis using subject-specific imaging biomarkers within a multi-kernel Gaussian process", Neuroimage: Clinical, *Published, October 2019*

3. Dealing with missing data - Due to the fast progression of prion disease, there are often missing samples caused by the impossibility of acquiring all the data. To avoid dropout that would reduce the sample size, I investigated ways of dealing with missing data, without compromising the sample. Hence, I developed a novel framework for classification, regression or stratification, where Gaussian Processes are conditioned by the uncertainty of the imputed values. Firstly, an imputation scheme is used to account for data heterogeneity through a robust observation model. Second, when training using the real and imputed values, I considered the uncertainty of each individual imputation in the optimisation of the overall model. This method not only estimates the missing samples, avoiding dropout, as well as identifies possible sources of bias created by the imputation methods. The results suggest that the imputation method efficiently estimates the missing samples.

(a) **Liane S. Canas**, ... , Simon Mead, Marc Modat, "Uncertainty Embedding for Partial Data in Gaussian Processes", *Paper in Preparation*

1.4.2 Thesis Outline

This thesis starts with an introduction focusing on the clinical and scientific motivation, the research objectives of this work and the contributions of this project to the field of study (Chapter 1). In Chapter 2, I introduce the relevant background to the development of this project, as well as the relevant literature related to the proposed methods. Chapter 3 presents the work regarding the extraction and selection of imaging biomarkers to characterise both forms of CJD: IPD and CJD. A comparison with current state-of-the-art methods is presented, motivating the need for a novel framework to extract and select subject-specific biomarkers. In Chapter 4, I describe a Bayesian model for the diagnosis of CJD, tailored to tackle the limitations of the current models when used in the context of prion diseases, such as small dataset and heterogeneous features. Chapter 5 introduces a multi-class classification model, where the assumptions considered in Chapter 4 are extended to account to multiple classes used for stratification and differential diagnosis. The model is evaluated using both the National Prion Monitoring Cohort (NPMC) and YOAD datasets. The robustness of the stratification model, when used to predict the clinical onset, is also assessed in this chapter. In chapter 6, I address the limitations raised by incomplete data and how to deal with this issue when in presence of limited data. A two-step framework is introduce in this chapter to estimate the missing samples, while reducing the statistical impact of the imputation techniques in the original sample. Lastly, Chapter 7 presents the main conclusions of this project, limitations of the proposed methods and the future work that could improve the obtained results.

Chapter 2

Machine Learning applied to Neuroimaging data - Background

Contents

Neuroimaging has made it possible to measure macro- and micro-structural changes over the course of brain pathologies, such as Alzheimer's disease (AD). During the past decades, the biomarkers extracted from neuroimaging data have been continuously integrated in machine learning algorithms in order to extract specific brain patterns during the course of these diseases. These models have proven to be a useful tool for subjects diagnosis and prognosis [62].

In this chapter, I present an overview of the machine learning methods used in the context of neuroimaging studies, in particular in the study of AD. A summary about dimensionality reduction techniques adopted in these frameworks, including feature extraction and feature selection methods, is presented in sections 2.1 and 2.2. The following sections, section 2.3 and 2.4, introduce the models used for the subjects diagnosis and prognosis, respectively. Finally, in section 2.5, the applicability of these methods to prion disease characterisation is presented. A standard notation for data and variables for machine learning models is adopted here. The input data is defined as $[x_{id}, y_{ic}]$, where $\mathbf{X} = [x_{id}] \in \mathcal{X}$ is the feature space and $\mathbf{Y} = [y_{ic}] \in \mathcal{Y}$ is the response variable, for a subject $i = \{1, \ldots, N\}$ given a set of $d = \{1, \ldots, D\}$ features and set of labels $\mathcal{C} = \{1, \ldots, C\}$. Without losing generality, both classification and regression models are defined as machine learning models. Lastly, the response variable is discrete in classification task, whilst for regression problems the response variable is continuous.

2.1 Feature Extraction and Feature Embedding

The main goal of a machine learning algorithm is to estimate the underlying function between the input and the output data. Machine learning algorithms must rely on a large number of samples and features to do it.

In neuroimaging studies, the term "*features*" typically refers to the informative measures derived from the post-processing step applied to the raw medical data, namely to the imaging data as MRI scans. This post-processing step includes feature methods designed to encode the relevant features to explain specific biological processes, either pathological or healthy processes. The features extracted from various imaging modalities can be in isolation or combined to make use of the complementary information provided by several modalities [63].

The methods adopted to extract meaningful features depend on (1) the type of information that is relevant to the machine learning task and (2) the imaging modality. Figure 2.1, illustrates the feature extraction methods used to derive informative measures from specific imaging modalities.

Structural MRI techniques (Figure 2.1, top scheme), give relevant information about the brain structure, namely about structural changes, such as thinning of the cortical surface, structural variation in several brain regions and regional tissue densities caused by cerebral neurodegeneration. These features serve as markers of the stage and aggressiveness of the neurodegenerative aspect of illnesses, such as AD [62]. The three main feature extraction methods for assessing structural variation con-

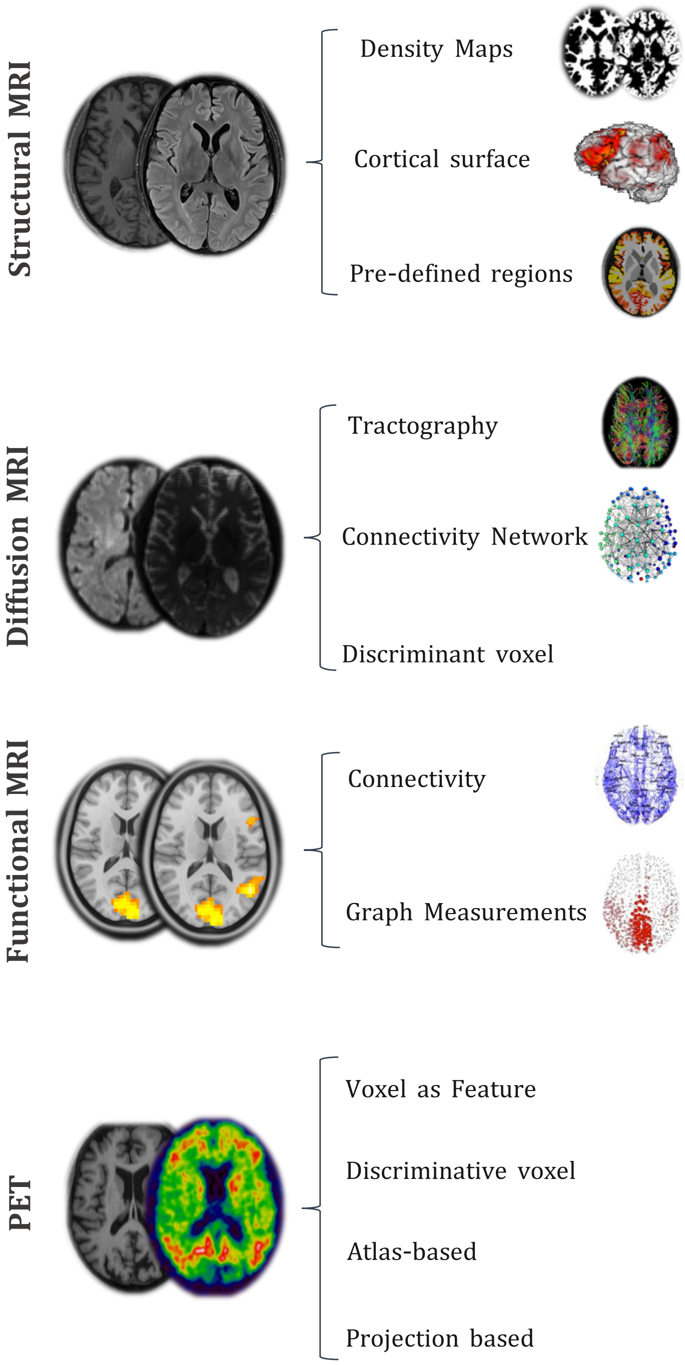


Figure 2.1: Feature extraction methods for imaging data. Main methods used to process raw imaging data in order to derive informative measures to be used as input in machine learning algorithms.

sidered are: (1) density maps, (2) cortical surface and (3) pre-defined regions-based methods. In detail, the density map-based methods quantify patterns of atrophy by computing the density map of white matter (WM), grey matter (GM) and CSF brain tissues. These maps can be generated by methods such as voxel-based morphometry (VBM) [64] or regional analysis of volumes examined in normalised space maps [65]. Cortical surface-based methods can also be used to extract measures of atrophy of the brain. Nick Fox *et al.* [66], use brain boundary shift integral (BBSI) as a measure of brain atrophy and demonstrate its application to study AD. Similarly, Dickerson *et al.* [67], analysed measurements of the cortical surface to identify signs of brain atrophy in AD patients. Lastly, predefined regions-based methods are also use to extract relevant features to characterise neurodegenerative diseases. These methods are based on the prior knowledge of the magnitude and spatial pattern of this illness that were acquired by studies previously conducted on histological or imaging data [62]. The analysis of specific brain regions requires the segmentation of the brain tissues and its parcellation in to regions of interest. Several segmentation and parcellation methods have been successfully used to extract structural features to characterise AD [68–70].

Most neurodegenerative diseases are associated with loss of myelin, thereby compromising the integrity of WM and leading to abnormal diffusivity patterns. Therefore, DWI and Diffusion tensor imaging (DTI) are used to analyse water diffusion at the microstructural level of the brain for determining the abnormal diffusion pattern of AD [62]. The DTI-based features can be grouped according to the way that they are extracted as (1) tractography [71], (2) connectivity network measurements [72] and (3) discriminative voxel analysis [73] (Figure 2.1, second scheme).

Functional connectivity between various brain regions are often compromised due to the neurodegenerative process induced by AD [62]. The changes in the functional connectivity are generally measured using fMRI. The evidence of disrupted functional connectivity and its association with AD led researchers to develop measures to properly quantify the functional connectivity and to capture the global distribution of its abnormalities for AD patients [74]. The quantification of this measure involves spatial parcellation of fMRI data according to a structural brain template and the calculation of pair-wise connectivity between the activation in all pairs of regions, as illustrated in Figure 2.1.

Finally, the characteristic patterns of glucose metabolism on brain FDG-PET and the amyloid deposition on amyloid PET can help differentiating AD from healthy controls. Four main methods that use the cerebral glucose metabolism rate are detailed in Figure 2.1: voxels as feature-based, discriminative voxel selection-based, atlas based and projection based methods.

Irrespective of the feature extraction algorithms, the number of features often supplants the number of samples available, requiring the reduction of the dimensionality of the feature space. Some feature embedding methods can both extract the features from raw data while transforming the original feature space into a lower-dimensional subspace. Feature embedding methods can be divided into linear and nonlinear techniques [75]. The linear methods include Principal Component Analysis (PCA) and linear discriminant analysis (LDA) [76, 77]. PCA is one of the most classical linear dimensionality reduction methods, widely used in neuroimaging studies. In brief, this approach finds the optimal subspace that represents the data distribution. Thus, a mapping matrix consisting of the first d feature vectors are used. Note that these vectors correspond to largest feature values from the covariance matrix and consequently are sufficient to explain the data variance to capture the patterns to explain the output via a machine learning model [76]. There are also many nonlinear techniques, such as Kernel PCA, Multidimensional scaling (MDS) and Isometric Feature Mapping (Isomap), that are used to encode non-linear dependencies between the input and the output [77–80].

The feature extraction transforms the original feature space in to a lower dimensional subspace [77], in which features encode the patterns used to estimate the function that better explains the output.

More recently, deep learning models are used for both feature extraction and classification models [77]. These techniques have became the state-of-art in many fields, including the medical imaging field [81]. In particular, convolutional neural networks (CNNs) have proven to be powerful tools for a broad range of computer vision tasks [81, 82], since deep CNNs automatically learn mid-level and high-level patterns obtained directly from raw data, such as images, hence no prior assumptions regarding the data are required. These methods have been widely used for lesion detection, segmentation and shape models and subjects classification [81]. Furthermore, these techniques can deal with different sources of data, including MRI, CT, ultrasound etc., without any specific pre-processing step. However, they required a large sample size to be conveniently trained, which is not often possible in the medical imaging context. Consequently, they are often inappropriate to study very heterogeneous and rare diseases.

In neuroimaging studies, even if the number of samples is reduced, hampering the use of deep learning techniques, after feature extraction the number of features often supplants the number of samples available. When compared with the size available samples, large features spaces cause many issues, such as the need for more storage and greater computational complexity during the training of the model. Additionally, by having a higher number of features without increasing the number of training samples, the dimensionality of the feature space would increase as well as its sparsity. Due to this sparsity, the optimisation of the machine learning model would likely overfit, thus hampering the generalisation ability and reducing the predictive power of the model. Therefore, it is reasonable and important to ignore the input features with reduced effect on the output, since by including irrelevant input features the computational cost of the algorithm increases, whereas the performance of the model might be compromised. To tackle these issues, feature selection methods can be used in statistical analysis and machine learning frameworks, as they allow the relevant features to be chosen to explain the output. The following section introduces some of the most used feature selection algorithms used to handcraft the best subset of features, to improve the performance of the machine learning model.

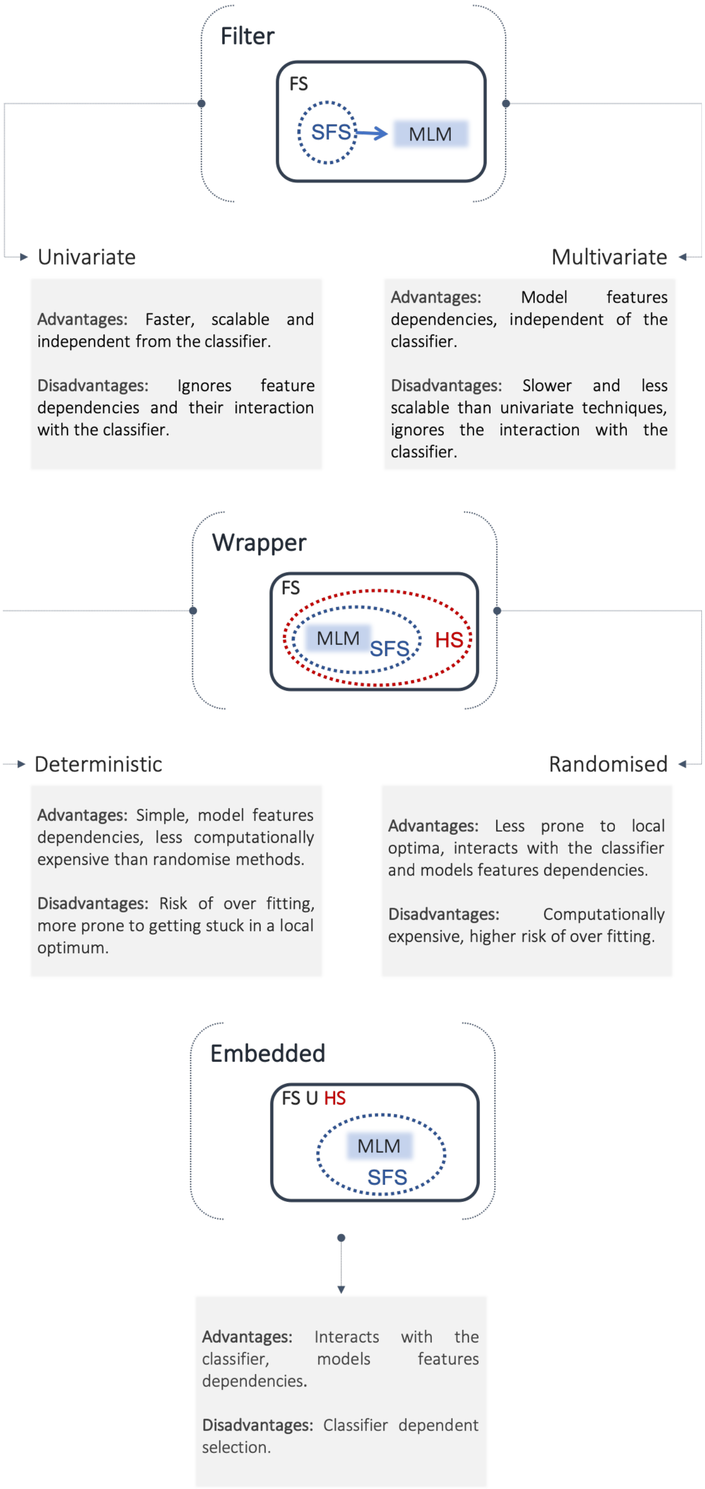
2.2 Feature Selection

By using feature selection approaches, the performance of the machine learning models can be improved. In fact, the obtained subset of feature will lead to more robust models, namely by (1) avoiding over-fitting and improving model performance, (2) providing faster and more cost-effective models training and by (3) gaining a deeper insight into underlying processes that generated the data [62, 77, 83, 84].

Feature selection methods can be organised into three categories: filter, wrapper and ensemble methods, as presented in Figure 2.2, where feature space (FS), hypothesis space (HS) and the subset of features selected (SFS) are considered differently by the machine learning model (MLM), depending on the type of implemented method [83].

2.2.1 Filter methods

Feature selection based on filter techniques assess the relevance of the features based only in the intrinsic properties of the data [83]. These methods apply a ranking



Advantages: Faster, scalable and independent from the classifier.

Disadvantages: Ignores feature dependencies and their interaction with the classifier.

Advantages: Model features dependencies, independent of the classifier.

Disadvantages: Slower and less scalable than univariate techniques, ignores the interaction with the classifier.

Advantages: Simple, model features dependencies, less computationally expensive than randomise methods.

Disadvantages: Risk of over fitting, more prone to getting stuck in a local optimum.

Advantages: Less prone to local optima, interacts with the classifier and models features dependencies.

Disadvantages: Computationally expensive, higher risk of over fitting.

Advantages: Interacts with the classifier, models features dependencies.

Disadvantages: Classifier dependent selection.

Figure 2.2: Taxonomy of feature selection techniques. For each main group of feature selection techniques it is presented their main advantages and disadvantages, when compared with other feature selection methods. The schemes present the workflow of the different techniques where FS: Feature Space; SFS: Selected Feature Space; HS: Hypothesis space; MLM: a generic machine learning model, either classifier or regression model, that uses the feature selected within the learning task.

system to order the variables according to a **scoring criteria** and a threshold is then used to remove the variables below it. Essentially, a feature is discarded if it has lower or no influence on the response variable [84]. Note also that most of the proposed filter methods are univariate: each feature is considered separately, ignoring features dependencies, which may lead to worse performance of the machine learning algorithm due to redundancy [83].

Filter methods, contrarily to other feature selection method, are applied prior to any machine learning model; therefore, only the selected subset of variables is used as input during the training phase (Figure 2.2) [77, 83]. As a consequence, the filter methods ignore the interaction with the model; i.e., the search in the feature subset space is separated from the search in the hypothesis space in which the model selection is done [83]. The main challenge of these methods is how to measure the relevance of the variables to the data or to the response variable. Several metrics are used to establish the relevance of the features, such as Maximum Variance, Laplacian score and Fisher score [77]. In the context of medical imaging, the most common metrics used as scoring criteria to rank features are the correlation criteria and the mutual information [84, 85], as detailed below.

Statistical tests: t-test

Among the various ranking methods, the t-tests have been successfully used as feature selection method [86]. Wee *et al.* [87], used a filter-based approach to select the most relevant region of interest (ROI) in the brain for AD and mild cognitive impairment (MCI) predictions. In this study, only the features with p-values smaller than the predefined threshold, measured via between t-test, were used for subsequent analysis. Due to the limitations of these techniques, the features retained by this approach may still inevitably be inter-correlated. Therefore, another filter-based approach, **minimum redundancy and maximum relevance (mRMR)** [88], is used to further reduce the feature dimensionality. The mRMR algorithm was initially developed for feature selection of microarray data and genetic information. It tends to select a subset of features having the most correlation with a class – relevance – and the least correlation between themselves – redundancy. In this algorithm, the features are ranked according to the mRMR criteria. Relevance can be calculated by using the F-statistic when features are continuous, or mutual information for discrete features; whereas, the redundancy can be calculated by using

Pearson correlation coefficient (for continuous features) or mutual information (for discrete features). The efficiency of mRMR as feature selection method was assessed by Wee *et al.*, that proved that the features selected by this technique improved the classification performance [89].

Tang *et al.* [90], also used a t-test approach to further select the relevant brain features obtained from principal component analysis (PCA). The resulting features were used to differentiate AD patients from healthy controls, boosting the classifier performance.

Correlation Criteria

The Pearson's correlation, Equation 2.1, is also used to detect dependencies between the features and the response variable [84, 85]. This ranking metric finds only linear dependencies between the j_{th} feature of $x_{ij} \in X = [\boldsymbol{x}_{.1}, \ldots, \boldsymbol{x}_{.D}]$, and the response variable y_i, for the sample i.

$$\rho_j(x_j) = \frac{\sum_{i=1}^{N}((x_{ij} - \bar{\boldsymbol{x}}_j)(y_i - \bar{\boldsymbol{y}}))}{\sqrt{\sum_{i=1}^{N}(x_{ij} - \bar{\boldsymbol{x}}_j)^2 \sum_{i=1}^{N}(y_i - \bar{\boldsymbol{y}})^2}} \tag{2.1}$$

Davatzikos and collaborators use the Pearson coefficient, as defined by Equation 2.1, to identify brain ROI in which the tissue density is well correlated with the clinical score [91]. Wee *et al.* [72], also quantified the discriminative power of a feature to the classification, as well as its generalizability, by measuring its correlation with the labels. The Pearson's coefficient was used to rank the features, as proposed by Fan *et al.* [92]. The larger the absolute value of Pearson's correlation coefficient is, the more relevant to the classification task the feature is. The generalizability of the features when used in different samples was then evaluated via leave-one-out cross-validation when measuring the correlation of the feature with respect to the clinical labels via Pearson correlation coefficient [92]. This feature selection method was effective in the identification of brain ROI with WM alterations, even though the wrapper methods, also analysed in this study, have lead to an improvement of the accuracy of the classifier [72].

Mutual information

Mutual information (MI) can be used as filter feature selection method, where the dependency between each feature and the labels is used as a scoring rank [84]. MI can be used to compare discrete variables, where the concept of Shannon's entropy (Equation 2.2, upper row) is used to estimate the uncertainty in the response variable $\boldsymbol{y}$, and the conditional entropy (Equation 2.2, bottom row) implies that by observing variable $\boldsymbol{x_j}$, the uncertainty in the response variable is reduced. Simply, the MI measures how much an input variable (feature) can explain the response variable. As a result, Equation 2.3 gives the MI between the response variable and each feature.

$$\begin{aligned} \mathrm{H}(\boldsymbol{y}) &= -\sum_{y\in\mathcal{Y}} p(y)\log p(y) \\ \mathrm{H}(\boldsymbol{y}|\boldsymbol{x}) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \end{aligned} \tag{2.2}$$

$$\mathrm{MI}(\boldsymbol{y},\boldsymbol{x}) = \mathrm{H}(\boldsymbol{y}) - \mathrm{H}(\boldsymbol{y}|\boldsymbol{x}) \tag{2.3}$$

In the case of continuous variables, MI can also be interpreted as the Kullback-Leibler divergence (KL), Equation 2.4, where $D_{\mathrm{KL}}(\boldsymbol{y} \parallel \boldsymbol{x})$ gives the mutual information between each feature and the labels in terms of their probabilistic density functions $p(\boldsymbol{y}), p(\boldsymbol{x})$ and $p(\boldsymbol{y},\boldsymbol{x})$ [88, 93].

$$\mathrm{MI}(\boldsymbol{y},\boldsymbol{x}) \triangleq D_{\mathrm{KL}}(\boldsymbol{y} \parallel \boldsymbol{x}) = \int\int p(\boldsymbol{y},\boldsymbol{x})\log\left(\frac{p(\boldsymbol{y},\boldsymbol{x})}{p(\boldsymbol{y})p(\boldsymbol{x})}\right) d\boldsymbol{x}d\boldsymbol{x} \tag{2.4}$$

For both cases, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, the MI will be zero, and greater than zero if they are dependent. The features are ranked using the MI and a threshold is set to select $d < D$ features.

Peng *et al.* [88], had improved the classification accuracy by selecting the subset of features via MI, proving the effectiveness of this method to select the relevant variables. Korolev *et al.* [94], also implemented a MI approach to select the relevant biomarkers in the context of neuroimaging data. By adopting this feature selection method, Korolev and collaborators were able to identify the increasing risk of MCI subjects to developed AD [94].

The main advantages of the filter methods are their computational efficiency and their ability to avoid over-fitting, proving to work well for certain dataset. Furthermore, filter methods do not rely on the machine learning algorithms used for further analysis using the selected subset. Thus, they do avoid introducing a bias on the models' estimation, caused by changing the data to fit the learning algorithm. Nevertheless, the filter methods show some limitations, since they might select a subset of features that is not the optimal, but rather redundant, because the correlation between the features is not taken into account. Some of these limitations are addressed by wrapper and embedded methods, as detailed below.

2.2.2 Wrapper methods

Unlike filter methods that use a feature relevance criteria, wrapper methods rely on the learning model to obtain a subset of relevant feature. Therefore, these methods embed the model hypothesis search within the feature subset search (Figure 2.2) [83]. Wrapper methods are black-box systems that use the current prediction information as the objective function to evaluate the variable subset [77, 84].

Wrapper methods evaluate the optimal subset of features heuristically. Heuristic search methods and sequential selection algorithm are used to guide the search for an optimal subset. The widely used heuristic approaches are mainly evolutionary algorithms, including genetic algorithms (GA) and Particle Swarm Optimisation (PSO), and others. However, note that when the number of dimensions is high, the computational time and complexity of these methods also increase [77, 95].

These methods, even if not used as often as filter methods, have already been used to select relevant features to characterise AD patients. Beheshti and collaborators used a hierarchical feature selection method to reduce the high-dimensional dataset, which combines feature ranking with a GA to reduce the dimensionality, and to select optimal features for the high performance MCI conversion prediction and AD classification. The performance of this method was evaluated using PCA data reduction and raw-feature vectors. The results showed the potential of wrapper methods to further reduce the feature space, for neurodegenerative studies [96].

2.2.3 Embedded methods

Lastly, feature selection can also be achieved through embedded methods, as presented in Figure 2.2. Embedded methods include feature selection as part of the training process of the model. These methods do not require splitting the data into training and testing sets, for both feature selection and training of the model, avoiding also double-dipping issues. Therefore, embedded feature selection methods are particularly useful for small data sets [77].

Since the selection of the features is based on the ranking of the features during the training stage of the model, the weights (rank) of the features can be used as classifier weights. By conducting sensitivity analysis of the weights, feature selection can be achieved; i.e., the change in weight can be viewed as changing the relevance of a feature. Some studies have suggested to use the change in the objective function to select the meaningful features [84]. This concept of using the weights as the ranking and the search using the change in the objective function is applied to the SVM classifier to perform recursive feature elimination, also defined as the SVM-RFE method [86, 97]. In the SVM-RFE method, the $L2$-norm is used in the SVM minimisation problem. It is shown in the literature that other functions can be used, which help in feature selection. Similar to optimising the SVM and assigning weight to features, the same can be done using Neural Networks. Multilayer perceptron networks are trained and feature weights are calculated using a saliency measure calculated from the trained network [98]. In this study a penalty is applied for features with small magnitude at the node and the nodes connecting to these input features are excluded. This type of node removal, also called Network Pruning, is commonly used to obtain the optimum network architecture for Neural Networks [98]. These methods are mainly used in the context of classification tasks, and the feature selection is a step embedded in the full framework. Therefore, these methods and examples of their applications to medical imaging problems will be detailed in section 2.3.

2.3 Subjects Diagnosis

The diagnosis of symptomatic subjects among healthy controls can be achieved through classification models, using the features extracted and selected from neuroimaging data. Similarly, the differential diagnosis of these patients, as well as their stratification according to symptoms severity, can be modelled in a multi-class fashion.

The modelling of imaging biomarkers can be performed either via parametric or non-parametric models, depending on the initial assumptions and understanding of the disease to be analysed. In the parametric models, the parameters of the function that models the features are fixed, and most of them are known except for a few parameters, which are estimated using training data [99, 100]. When the pattern of the features is close to the assumed parametric model, parametric classifiers are expected to perform very well. However, these classifiers often lead to poor performance when one or more model assumptions are violated. Since it is challenging to validate the assumptions *a priori*, these models tend to perform poorly. On the other hand, non-parametric models are more flexible, since they do not rely on initial assumptions regarding the samples pattern. Nevertheless, these models are also not ideal and can perform worse than parametric models, especially in the presence of small training samples and statistical instability [100, 101]. Therefore, their major limitation is their inability to include additional information, such as *a priori* assumptions that can condition the problem leading to a better estimation of the model [99, 100].

In the study of neurodegenerative diseases, both types of models have been implemented, showing their flexibility to be used to model both imaging and non-imaging biomarkers, aiming at subjects diagnosis. The following sections present an overview of some of the examples of parametric and non-parametric models, as well as their application to this field.

2.3.1 Parametric Models

Logistic Regression

Logistic regression, Equation 2.5, is a discriminative parametric model used to predict the odds of a given label, ϕ_i, based on the values of the independent variables (covariates), $\mathbf{x}_i$, and the vector of the model's parameters $\boldsymbol{\beta}$ [102].

$$\operatorname{logit} \phi_i = \log \frac{\phi_i}{1 - \phi_i} = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2.5}$$

This model is widely used in the analysis of either binary or binomial responses and several explanatory variables. The parameters of the logistic regression model can be determined via maximum likelihood [102, 103]. Given a dataset $\{\mathbf{x}_i, y_i\}$, where $y_i \in [0, 1]$, for $i = 1, \ldots, N$ subjects, the likelihood function is given by:

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} \phi_i^{y_i} \{1 - \phi_i\}^{1 - y_i} \tag{2.6}$$

where $\mathbf{y} = (y_1, \ldots, y_N)^T$ and $\phi_i = p(\mathcal{C}_1|\mathbf{x}_i)$ for a given class $\mathcal{C}_1$. Note that the probability of the other possible class is given by $p(\mathcal{C}_2|\mathbf{x}_i) = 1 - p(\mathcal{C}_1|\mathbf{x}_i)$. The parameters of the model are then optimised using the cross-entropy error function [103].

Logistic regression can also take the form of a multi-class paradigm, in which the probability of the response variable y_i is given by the product of the probabilities of each class $\mathcal{C}_k, k \in \{1, \ldots, C\}$ [103]. Therefore, the likelihood function becomes:

$$p(\mathbf{y}|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) = \prod_{i=1}^{N} \prod_{k=1}^{K} \phi_{i,k}^{y_{i,k}} \tag{2.7}$$

where it is used a 1-of-K coding scheme in which the target vector, y_i, for a feature vector $\mathbf{x}_i$ belonging to a class $\mathcal{C}_k$ is a design vector with all values to 0, except k, which has value 1. This formulation can be used either for differential diagnosis or subjects stratification.

The logistic regression is widely used for the diagnosis of AD. Rao Lee *et al.* [104], applied a sparse logistic regression to classify AD patients among healthy controls. In this study, they used a high dimensional feature space composed of voxel-wise grey matter volumes derived from structural MRI, hence the need for an adaption of the conventional logistic regression to a sparse logistic regression. The methods implemented were able to automatically select clinically relevant regions for AD, while simultaneously performing the classification with better accuracies than other methods such as linear discriminant analysis (LDA) [104]. Note that the sparse logistic regression is implemented by means of a prior on the weight vector, which penalises the log-likelihood and regularises the estimation of weights of the different covariates. The penalty term, which incorporates both an *L1* and *L2* penalty on the weight vector, is the elastic net penalty, also used in regression and classification problems when aiming the reduction of the feature space.

Moradi *et al.* [105], presented a novel method for predicting the MCI-to-AD conversion from one to three years before clinical diagnosis, using MR data. Logistic regression was used as feature selection method. The features that maximise the performance of the logistic regression in the classification of AD, were selected as the most relevant features to identify MCI conversion. The results had shown the versatility of this method to both classify AD and to select the most significant features to identify early signs of dementia [105].

More recently, logistic regression was also used to study prion diseases. Forner *et al.* [106], implemented a statistical analysis, including logistic regression, to assess the ability of CSF biomarker to diagnose sCJD subjects. Logistic regression was used also to compare the predictive value of CSF biomarkers against MRI biomarkers for the sCJD diagnosis. This study shows the potential of logistic regression to be also used for the diagnosis of CJD, being a convenient model with simple interpretation, when the biomarkers are already identified.

Finally, Wang *et al.* [107], combined multimodal data, including MRI, phenotyping-morphometry and structural connectomics in an AD diagnosis tool. Three machine learning algorithms were tested, including logistic regression. The best performance regarding the classification of MCI *versus* AD was achieved by the combination of PCA and logistic regression, for the morphometry and connectome features, only supplanted by random forest (RF) when using CSF biomarkers.

These studies have shown the potential of logistic regression for subjects diagnosis. However, note that for high dimensional feature spaces, sparse forms of logistic regression need to be considered, to avoid the complexity of the model and expensive computations. Furthermore, this method highly relies on previous assumptions regarding the biomarkers selected, given its parametric nature.

Support Vector Machines

SVM's can be described as linear models of the form of:

$$y(\mathbf{x}) = \boldsymbol{\beta}^T \psi(\mathbf{x}) + b \tag{2.8}$$

where $\psi(\mathbf{x})$ denotes a fixed feature-space transformation and b is a bias parameter. Assuming that the training dataset is linearly separable in the feature space, the choice of the parameters $\boldsymbol{\beta}$ and b satisfies $y(x_i) > 0$ for samples defined with target values +1 and $y(x_i) < 0$ for target values -1 [103]. There can be many solutions that separate the classes as mentioned. Therefore, it is considered as an optimal solution the model that gives the smallest generalisation error. The SVM approach tackles this issue through the concept of the margin (Figure 2.3), which is defined to be the smallest distance between the decision boundary and any of the samples. Additionally, the decision boundary, also defined as hyperplane, is chosen to be the one that maximises the margin [103, 108].

The perpendicular distance of a point x_i from a hyperplane defined by $y(x_i) > 0$ is given by $\frac{|y(\mathbf{x})|}{||\boldsymbol{\beta}||}$. Besides, the optimisation strategy only considers the solutions for which the samples are correctly classified, such as given Equation 2.8 and the real targets t_i, $t_i y(\mathbf{x}_i) > 0$. Thus, the distance of a point x_i to the hyperplane is given by

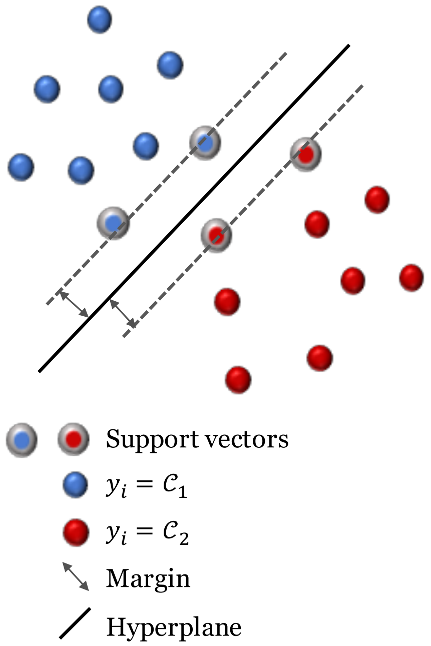


Figure 2.3: Schematic representation of a linear SVM. The margin is defined as the perpendicular distance between the decision boundary – hyperplane, represented by the full black line – and the closest of the data points, defined as support vectors. Maximising the margin leads to a particular choice of decision boundary. The location of this boundary is determined by the support vectors represented here with the grey outline.

Equation 2.9. Finally, the maximum margin solution is found by solving Equation 2.10.

$$\frac{t_i y(\mathbf{x}_i)}{||\boldsymbol{\beta}||} = \frac{t_i(\boldsymbol{\beta}^T \psi(\mathbf{x}_i) + b)}{||\boldsymbol{\beta}||} \tag{2.9}$$

$$\arg\max_{\boldsymbol{\beta},b} \left\{ \frac{1}{||\boldsymbol{\beta}||} \min_i \left[t_i(\boldsymbol{\beta}^T \psi(\mathbf{x}_i) + b) \right] \right\} \tag{2.10}$$

Note that, in the scenarios where the dataset is not linearly separable in the feature space $\mathcal{X}$, the feature space becomes linearly separable by using a nonlinear feature space transformation, defined implicitly by a non-linear kernel function. This approach will be further detailed in section 2.3.2. Furthermore, other adaptations of the linear SVM have been used to improved classification performance, such as relevance vectors machines (RVM). RVM can be used to provide posterior probabilistic outputs due to its Bayesian formulation [109]. A brief overview of these methods will be also discussed in the section 2.3.2.

SVM's became popular in various fields of research for solving classification and regression problems. The main advantage of SVM is that the determination of the model parameters correspond to a convex optimisation problem, consequently, any local solution is also a global optimum [103]. In neurodegenerative studies, SVM have shown good performance for the diagnosis of patients, dealing with multiple sources of data, such as MRI data, functional data, genetics, etc., [110]. However, in most of the these studies the feature space $\mathcal{X}$ is not linearly separable, hence the need to use a non-linear, non-parametric SVM. Examples of applications of SVM to study neurodegenerative diseases will be explored in section 2.3.2, given their non-parametric nature.

In summary, the two parametric models widely used for subjects' diagnosis with neurodegenerative diseases are logistic regression and SVM. SVM is essentially a decision machine; thus, it does not provide posterior probabilities, whereas logistic regression estimates the posterior probabilities. The logistic regression is also more sensitive to outliers than SVM, since the cost function of logistic regression diverges faster than the hinge loss used in the optimisation of SVM. Besides, the logistic loss does not go to zero even if the sample is classified confidently. This might lead to minor degradation in accuracy, when compared with SVM. On the other hand, SVM tries to maximise the margin between the closest support vectors, while logistic regression maximises the posterior class probability.

Nevertheless the good performance of these models in neuroimaging data analysis, the non-parametric models are more flexible and do not require strong prior assumptions. The following section will present an overview of the non-parametric models used in the context of neurodegenerative diseases and their main advantages when compared with parametric approaches.

2.3.2 Non-parametric Models

Kernel Machines

Kernel machines, also denominated as kernel methods, consist in a set of methods in which the feature space $\mathcal{X}$ is transformed in a new different feature space by means of a kernel function $\psi(\mathbf{X})$ [103]. In other words, kernel methods embed the data in some Hilbert spaces, $\mathcal{H}$, and search for linear relations in these spaces [111]. Formally, considering the feature space $\mathcal{X}$, and an embedding space $\mathcal{H}$, it applies the mapping $\psi : \mathcal{X} \rightarrow \mathcal{H}$ [111]. Given two samples, $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, a kernel function

k returns the inner product between the two embedded samples in the space $\mathcal{H}$ (Equation 2.11). The kernel function returns a kernel matrix $\mathbf{K} \in \mathrm{R}^{N \times N}$, which also corresponds to the Gram matrix $\mathbf{K_{ij}}$ computed using a covariance function.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \psi(\mathbf{x}_i), \psi(\mathbf{x}_j) \rangle \tag{2.11}$$

The kernel is a symmetric function and it produces a positive, (semi)definitive matrix [112]. The choice of kernel is crucial for the success of the kernel method, even if the selection of a proper kernel is not trivial [111, 113]. There are several forms of kernel functions commonly used in different kernel methods, such as kernel SVM and Gaussian Process (GP). Many kernel functions have the property of being a function only of the difference between the arguments, which are known as *stationary kernels* because they are invariant to translations in input space. Additionally, specific cases of these functions, which depend only on the magnitude of the distance between the arguments, are defined as radial basis functions. A set of basis kernel functions is introduced in Appendix B [103, 112]. Some studies have addressed the difficulty of kernel selection for a specific dataset [113–115]. However, the most practical and often used way to select and design the appropriate kernel still remains the empirical approach.

Several classification models can be reformulated in terms of a dual representation in which the kernel function arises naturally [103]. The kernel SVM's are an example of that, in which the feature space is transformed using a $\psi(\mathbf{X})$ kernel function. Figure 2.4 demonstrates the transformation of the feature space, using the a kernel function.

As previously mentioned, the RVM models are a good alternative to SVM, given their probabilistic nature. These models use Bayesian inference to obtain parsimonious solutions for regression and probabilistic classification [116]. The RVM has an identical functional form to the support vector machine, but given that provide a probabilistic estimation of the label, the likelihood function is given by:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{K}\boldsymbol{\beta}||^2\right\} \tag{2.12}$$

where $\mathbf{y}$ refers to the response variable, $\boldsymbol{\beta}$ is the vector of the weights and $\mathbf{K}$ is the kernel matrix obtained by a given covariance function (Appendix B). Note that RVM is actually equivalent to a Gaussian process model with Gaussian covariance function. Compared to SVM, the Bayesian formulation of the RVM avoids the set

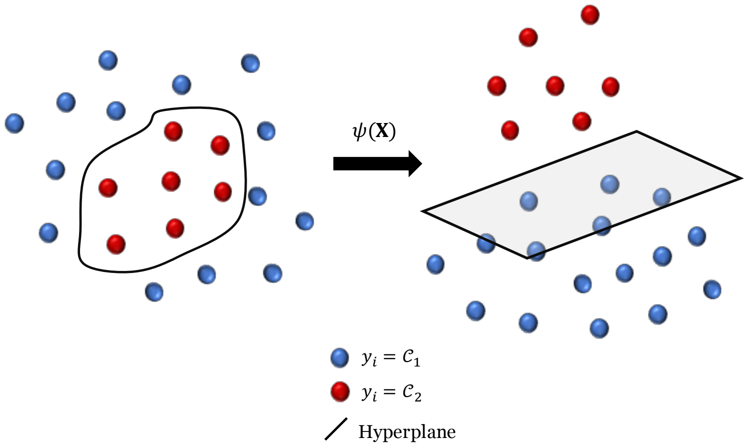


Figure 2.4: Schematic representation of a kernel SVM. The decision boundary is represented by the full black line, and the shadow area defined in the left side of the image. The function $\psi(\mathbf{X})$ represents the feature space transformation.

of free parameters that usually requires the use of a cross-validation optimisations scheme. Conversely, RVM uses an expectation maximisation learning method, hence it is at risk of local *minima* convergence. This is unlikely result from the standard sequential minimal optimisation (SMO)-based algorithms employed by SVM models, which are guaranteed to find a global optimum (of the convex problem). Nevertheless, they still give more reliable posterior estimations when compared with methods such as logistic regression.

In neuroimaging studies, both the linear and Gaussian kernels are commonly used to transform the feature space, for subjects diagnosis. In several studies, kernel methods are used to reduce the high dimensionality of the feature space, usually comprising the number of voxels of a 3D MR image, and to encode the brain patterns. Davatzikos *et al.* [91], used a non-linear SVM to detect patterns of brain structure characterising MCI, the prodromal phase of AD. In this work, from each support vector, the path of fastest change was constructed; in detail, for each sample that was close to the interface between MCI and controls the path of fastest change was extracted from the SVM gradients. From this information, they built an average spatial map to evaluate the brain's changes per region. The best classification rate was achieved for the hippocampus using the non-linear SVM [91].

Mourão-Miranda *et al.* [117], also used kernel SVM to analyse functional magnetic resonance imaging (fMRI). The study introduces a new framework, a spatio-temporal SVM, where the imaging features are encoded using a linear kernel SVM. The new method was compared with traditional SVM [117]. They show that by using the spatio-temporal approach, they can perform a dynamic discrimination analysis, showing how the regions discriminating between two cognitive states change over time. The results show that the spatio-temporal SVM discloses relevant transient responses in distributed brain systems that would be ignored by other models, such as SVM and generalised linear models. The model was validated only in healthy subjects, but it is flexible to be used for different purposes such as subjects' diagnosis [117].

The relevance of SVM in clinical context was assessed by Klöppel *et al.* [118]. Using a linear SVM, they classified GM segment of MR scans from pathologically proven AD subjects and cognitively normal elderly subjects. The AD subjects were correctly classified using whole brain images, with an curracy up to 96%. The intermediate cases, MCI, were correctly identified among AD and healthy controls, with an accuracy of 86%. Therefore, the results sustained the hypothesis that SVM correctly identify AD patients among healthy controls. Furthermore, these methods are also effective in the differential diagnosis of two forms of dementia: AD and frontotemporal lobar degeneration [118].

Zhang and collaborators also used a linear SVM to encode features from two imaging modalities and CSF information, as detailed in Figure 2.5 [119]. In this framework, the kernel parameters, for each modality, are jointly optimised with the SVM parameters, in an iterative way. After the kernel parameters are estimated, those parameters are used to combine the multiple kernel into a mixed kernel, as detailed in Figure 2.5, and then the SVM estimates the subjects status according to the mixed kernel. The proposed multi-kernel approach provides an efficient way to combine data from different modalities. The results showed that the combined approach outperforms the models using only the information from one modality, for both AD *versus* healthy controls, and MCI *versus* healthy controls classification. Furthermore, this study also proves the flexibility of the kernel-SVM to be used to combine the information from both imaging and non-imaging data [119].

Later, Zhang *et al.* [120], used the aforementioned multimodal approach to estimate continuous variables, namely the mini mental state examination (MMSE) and the Alzheimer's disease assessment scale-cognitive subscale (ADAS-Cog), via a support vector regression framework, as well as categorical variables via multimodal SVM. From a subset of features, previously selected using a multi-task feature selection, they used the multi-modal SVM, including multi-modal support vector classification and multi-modal support vector regression, to train the final support

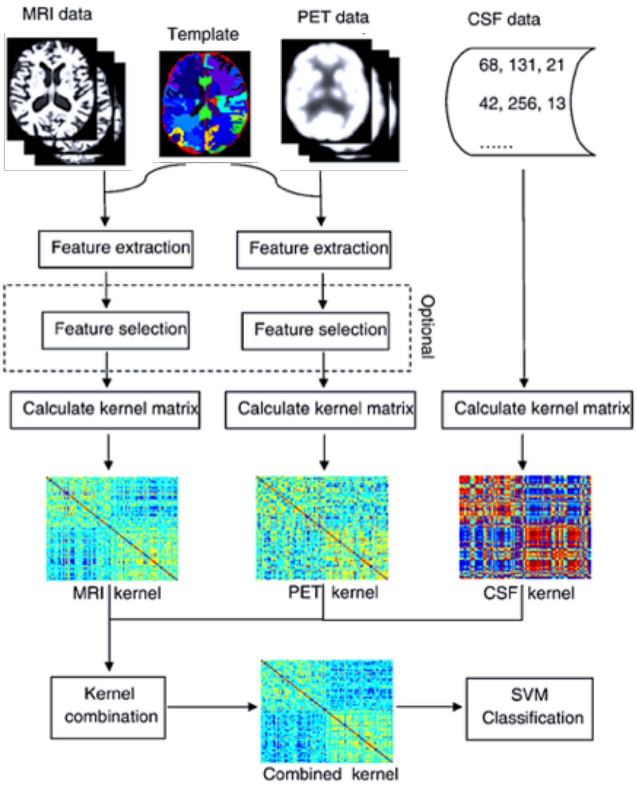


Figure 2.5: Representation of the multimodal data fusion and classification proposed by Zhang *et al.*. Image adapted from [119]

vector classification and regression models, respectively. The authors also consider that since a common subset of features is used to train both the classification and regression task, the models are actually performing a multi-modal multi-task learning [120]. The results showed that the proposed multi-modal multi-task method can effectively perform multiple-tasks learning from multi-modal data.

Ramírez *et al.* [121], proposed a fully automatic computer-aided diagnosis tool to improve the early detection of AD. The proposed approach uses a non-linear SVM to classify the subjects, based on imaging features extracted from imaging modalities, such as PET and SPECT. Their approach also reduces the dimensionality of the initial feature space, in order to find the most informative ROIs and the most discriminant image parameters with the aim of improving the accuracy of the system. Among all the features evaluated, coronal standard deviation and sagittal correlation parameters were found to be the most effective ones for reducing the dimensionality of the feature space and improving the diagnosis accuracy when a radial (RBF) basis function kernel SVM is used. The proposed model yielded to 90.38% accuracy in the diagnosis of the early stages of AD and outperformed existing techniques including the voxel-as-features (VAF) approach. Note that different kernel functions were considered in this study, highlighting the flexibility of the kernel machines to deal with different functions [121].

Dyrba and collaborators explored the used of a multi-kernel learning model to encode the information from different imaging modalities, aiming the diagnosis of symptomatic AD subjects [73]. Multi-kernel SVM enables the contribution of each modality to the classification results to be controlled more closely and to use complementary information provided by the modalities within the model. The results did not show significant improvement when using multimodal information, rather than using a single modality. However, this method shows the possibility of using multimodal information to better characterise the evolution of the disease [73].

Similarly, Ahmed *et al.* [122], also proposed the multi-kernel learning for multimodal signatures fusion to recognise MCI patients against AD subjects and healthy controls. Here, a global fusion framework was used to combine the signatures computed from structural MRI, DWI and CSF to distinguish between healthy controls, MCI and AD subjects. The multi-kernel algorithm combines the multiple kernel as a weighted linear combination of the kernels. During the training stage both SVM parameters and the weights are estimated withing the same optimisation scheme. Conversely to the results presents by Dyrba *et al.*, [73], the concatenation of the information from the different modalities benefit the classification performance [122].

Future studies need to confirm whether or not multimodal imaging provides additional diagnostic accuracy in prodromal stages of AD or in differential diagnosis between different types of dementia [73].

Finally, other kernel methods, such as RVM, were also used aiming the diagnosis of AD, even if used as a surrogate approach for diagnosis. Franke *et al.*, used a RVM to estimate the subjects brain age from T1w-MRI [123]. The estimation of the brain age was then used as a surrogate measure of the brain damage for both healthy and AD subjects.

Kernel methods are very powerfull to encode information from different modalities to diagnose neurodegenerative diseases. Most of these methods are discriminative approaches, which leads to simple models with a direct estimation of the likelihood of the predictions. Besides, the discriminative approaches are very appealing to deal with classification problems since their solution is directly modelled, as $p(y|\mathbf{X})$. However, to deal with missing input values, outliers and unlabelled data points in a principled fashion it is very helpful to have access to $p(\mathbf{X})$, and this can be obtained from marginalizing out the class label y from the joint as $p(\mathbf{X}) = \sum_y p(y)p(\mathbf{X}|y)$ in the generative approach. Furthermore, by using a generative approach is also possible to incorporate any prior information that can help to model the data in case of very noisy samples. To tackle the limitations of discriminative models, specially when in presence of noisy labels and missing samples, generative models can be used, even if considering the added complexity of these models. The following section introduces a generative non-parametric model, as well as some examples of its application for neuroimaging studies.

Gaussian Process

Machine learning studies, such as [124], have shown that for small datasets, without strong assumptions about the behaviour of input features, generative models are the most appropriate choice for classification, as they perform better. Nonetheless, discriminative models can also perform well in presence of small samples or missing data. Particularly, GP based approaches allow robust modelling even in the circumstances of highly uncertain or incomplete datasets. GP models learn a fit of the probability distribution of the response variable y given the input observation s$\mathbf{x}$ through the estimation of the posterior distribution, $p(\mathbf{x}|y)$, selecting the the most likely label y.

Ramsmussen and Williams have defined formally a Gaussian process as a *collection of random variables, any finite number of which have a joint Gaussian distribution* [112]. In other words, a Gaussian process is completely defined by its mean function, $m(\mathbf{X})$, and covariance function, $k(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j \in [1, \ldots, N]$, as described in equations 2.13 and 2.14 respectively.

$$m(\mathbf{X}) = \mathbb{E}\left[f(\mathbf{X})\right] \tag{2.13}$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}\left[(f(\mathbf{x}_i) - m(\mathbf{x}_j))(f(\mathbf{x}_i) - m(\mathbf{x}_j))\right] \tag{2.14}$$

Therefore, GP is defined as:

$$f(\mathbf{X}) \sim \mathcal{GP}(m(\mathbf{X}), k(\mathbf{x}_i, \mathbf{x}_j)) \tag{2.15}$$

In realistic scenarios, such as the study of neurodegenerative diseases, it is typical to not have access to function values themselves, but only noisy versions thereof $y =$

$f(\mathbf{X}) + \varepsilon$ [112]. By assuming additive independent identically distributed Gaussian noise ε with variance σ_N^2 , the prior on the noisy observations becomes:

$$\mathrm{cov}(y_i, y_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_N^2 \delta_{ij}$$

where δ_{ij} is a Kronecker delta which is one iff $i = j$ and zero otherwise, and i and j are two different samples. Note that $\sigma_N^2 \delta_{ij}$ is equivalent to $\sigma_N^2 I$, where $\boldsymbol{I}$ is an identity matrix. For noisy observations, the joint distribution of the observed target labels, y and the function labels, $f(\mathbf{X}_*)$, at the test[1] locations under the prior is defined as demonstrated in Equation 2.16, where $K(.,.)$ represents the covariance matrix obtained using a specific kernel function[2], with hyperparameters globally defined as $\boldsymbol{\theta}$.

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \left(\mathbf{0}, \quad \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I}, & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}), & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \tag{2.16}$$

Deriving the conditional distribution considering Equation 2.16, and the definition of a Gaussian process presented in Equation 2.15 the predictive functions for a Gaussian process are:

$$\mathbf{f}_* | \mathbf{x}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)), \quad \text{where} \tag{2.17}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*] = K(\mathbf{X}_*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I} \right]^{-1} \mathbf{y} \tag{2.18}$$

$$\mathrm{cov}(\mathbf{f}_*) = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I} \right]^{-1} K(\mathbf{X}, \mathbf{X}_*) \tag{2.19}$$

The aforementioned equations can be further simplified to express a more compact notation, where $K_* = K(\mathbf{X}_*, \mathbf{X})$ to denote the vector of covariances between the test samples, X_*, and the N training samples/subjects, and $K = K(\mathbf{X}, \mathbf{X})$.

$$\bar{f}_* = K_*^T \left[K + \sigma_N^2 \mathbf{I} \right]^{-1} \mathbf{y} \tag{2.20}$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - K_*^T \left[K + \sigma_N^2 \mathbf{I} \right]^{-1} K_* \tag{2.21}$$

[1] The * subscript refers to the unseen samples, also defined here as testing samples. The corresponding training sets are denoted by the same notation, without *.

[2] Similarly to the kernel methods described in section 2.3.2, the kernel function is defined by Equation 2.11, taking different forms as described in Appendix B.

Note also that the mean function (Equation 2.13), and the covariance function (Equation 2.14) of the (Gaussian) posterior process is now given by equations 2.20 and 2.21, respectively. The evidence of the predictions is finally given by marginal likelihood of the predictions $p(\mathbf{y}|\mathbf{X})$, Equation 2.22, which corresponds to the integral of the likelihood times the prior [112].

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f} \tag{2.22}$$

Under the assumptions of a Gaussian process model the prior is Gaussian, such as $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(0, K)$. Consequently, the likelihood is a factorised Gaussian $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_N^2\mathbf{I})$. Using the integration described by Rasmussen *et al.* [112], the logaritmic marginal likelihood is then given by:

$$\begin{aligned} \log p(\mathbf{f}|\mathbf{X}) &= -\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{N}{2}\log 2\pi, \\ \log p(\mathbf{y}|\mathbf{X}) &= -\frac{1}{2}\mathbf{y}^T(K+\sigma_N^2\mathbf{I})^1\mathbf{y} - \frac{1}{2}\log|K+\sigma_N^2\mathbf{I}| - \frac{N}{2}\log 2\pi. \end{aligned} \tag{2.23}$$

where σ_N^2 is the sample noise and N is the number of samples, here also representing the number of subjects. The best parameters of the model can be found through the maximisation of the marginal likelihood, Equation 2.23, where both the sample noise σ_N^2 and the hyperparameters of the kernel function $\boldsymbol{\theta}$ are optimised.

The concepts detailed above are applied in a GP model definition, when intending to build a regression model. However, these concepts can be modified envision a classification task, such as binary classification. The main principle behind binary classification using GP, as diagnostic tool, is to use a prior over the latent function $f(\mathbf{X})$, and ”squash” this through the logistic function to obtain the prior on $\pi(\mathbf{X}) \triangleq p(y = +1|\mathbf{X}) = \sigma(f(\mathbf{X}))$. In other words, this formulation consists in generalisation of the linear logistic logistic regression model, where the linear function $f(\mathbf{X})$ function is replaced by a GP formulation, and correspondingly the prior is replace on the weights by a GP prior [112].

The inference step of a GP classifier model is achieved by two main steps:

1. Initially the distribution of the latent variable corresponding to a new sample is computed, as detailed in Equation 2.24, where $p(\mathbf{f}|\mathbf{X},\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})/p(\mathbf{y}|\mathbf{X})$ is the posterior over the latent variables [112];

2. Secondly, using the distribution over the latent function f_* the probability

prediction is computed, as detailed in Equation 2.25 [112].

$$p(f_*|\mathbf{X},\mathbf{y},\mathbf{X}_*) = \int p(f_*|\mathbf{X},\mathbf{X}_*,\mathbf{f})p(\mathbf{f}|\mathbf{X},\mathbf{y})d\mathbf{f} \tag{2.24}$$

$$\pi(\mathbf{X}) \triangleq p(y = +1|\mathbf{X},\mathbf{y},\mathbf{X}_*) = \int \sigma(f_*)p(f_*|\mathbf{X},\mathbf{y},\mathbf{X}_*)df_* \tag{2.25}$$

However, given the non-Gaussian likelihood considered in Equation 2.24, the integral is analytically intractable. Similarly, Equation 2.25 can be also intractable [112]. Therefore, there are approximations that can be considered to estimate the integral values, such as Laplace and expectation propagation (EP) approximations, or solutions based on Monte Carlo sampling [112, 125]. Some studies have explored the performance of GP classifiers when using different approximation functions [126, 127]. However, the choice of the approximation used is highly dependent on the balance of the accuracy of the estimation o the parameters *versus* the computational time required for that estimation. Therefore, the choice of the approximation used should be based on the aim of the model and the constrains aforementioned.

The concepts of GP models were used for the diagnosis of AD patients. Young and collaborators have introduced GP classifier as a way to identify MCI conversion [128]. The fully Bayesian framework naturally produced probabilistic predictions, which were well correlated with the actual chances of converting to AD within 3 years in a population of 96 MCI-stable and 47 MCI-conversion subjects. Furthermore, this study also showed the flexibility of GP to be used in a similar manner than the kernel methods describe in section 2.3.2, which can integrate multimodal data. Specifically, Young *et al.* [128], included information extracted from volumetric MRI, FDG-PET, CSF, and APOE genotype within the classification process through the use of a mixed kernel. Similarly to the MKL approaches, the GP approach aids the combination of different data sources by learning parameters automatically from training data via type-II maximum likelihood. The proposed method was compared to a more conventional approach based on cross validation and an SVM classifier. The results for predicting MCI conversion based on the combination of all three types of data showed a balanced accuracy of 74%. This is a substantially higher accuracy than could be obtained using any individual modality or using a multikernel SVM, and is competitive with the highest accuracy yet achieved for predicting conversion within three years on the widely used Alzheimer's disease neuroimaging initiative (ADNI) dataset [129].

Similarly, Challis *et al.* [130], investigated the performance of Bayesian Gaussian process logistic regression (GP-LR) models with linear and non-linear covariance functions, when used to classify both AD and MCI patients. The GP-LR models can be interpreted as a Bayesian probabilistic framework analogue to kernel SVM classifiers. In this study, the class probability estimates were considered as measure of the confidence of the model's predictions. Hence, these class probabilities, seen as confidence score, may be extremely useful in the clinical setting. The proposed methods were applied to a sample of 77 subjects; 27 with a diagnosis of probable AD, 50 with a diagnosis of amnesic MCI and a control sample of 39. The input data considered in this study comprises only MRI data. The results support the hypothesis that GP-LR models can be effective at performing patient stratification, since the implemented model achieved 75% accuracy in the identification of subjects with amnesic mild cognitive impairment among healthy subjects, and 97% accuracy disambiguating amnesic mild cognitive impairment subjects from those with Alzheimer's disease [130].

More recently, Fruehwirt *et al.* [131], combined multivariate pattern analysis and GP classification to analyse elctrophysiological data, namely event-related potentials (ERP) to study the neurodegenerative processes in AD patients. The new method integrated interregional synchrony of ERP time signatures to account for the temporal information of ERPs. In this study, Fruehwirt *et al.* [131], showed that the proposed framework is useful to build neurophysiological markers to be used as features in classification tasks for single subjects diagnosis. Furthermore, the study also demonstrated the added-value of using a GP model, hence it outperformed the probabalistic methods used as baseline, with the highest AUC overall (0.802) being achieved using the new spatiotemporal method in the prediction of rapid cognitive decline [131].

These studies demonstrate the feasibility of using non-parametric models, namely kernel methods, to study neurodegenerative diseases. GPs models are particularly interesting to study small samples with noisy labels and/or missing samples. Therefore, these models are specially interesting to be used in clinical context, where the data is often incomplete and the clinical diagnostic, used as label, can be inconsistent. However, the methods described above, including both parametric and non-parametric models, only allow the identification of patients among healthy controls or the differentiation of the several stages of the disease. Therefore, they lack metrics to anticipate the clinical onset of symptoms and the prediction of conversion to illness statuses, such as AD and other neurodegenerative diseases.

To address this issue, some studies have proposed to perform subjects prognosis, in order to not only identify symptomatic subjects in the early stages of the disease, but also to predict the onset of clinical symptoms in a defined period of time for subjects apparently normal at the current time-point. Subjects' prognosis is often achieved by anticipating the clinical onset by modelling the disease evolution based on the population [107, 132–135], whereas other studies proposed to model the biomarkers evolution during the course of the disease to indirectly predict the conversion of asymptomatic subjects to symptomatic status [136–139].

The section 2.4 details both parametric and non-parametric models used aiming subjects' prognosis in context of neurodegenerative diseases. The concept of mixed-effects model (MEM) will be presented in section 2.4.1, as an example of parametric models used to predict the evolution of the biomarkers over the course of the disease, before and after onset. Section 2.4.2 introduces the non-linear MEM and GP regression as examples of non-parametric models used to both model the evolution of biomarkers and the subjects' prognosis.

2.4 Subjects Prognosis

Alternatively to subjects diagnosis approaches presented in section 2.3, modelling the patterns of biomarkers that are used to characterise and understand the disease progression is also an effective way to predict the subject's status in early stages of the disease. The evolution of the clinical manifestations encoded by imaging or non-imaging biomarkers can be studied through disease progression models, which often take the form of regression models. In order to understand and predict the evolution of clinical manifestations, several types of disease progression models have been applied. These models are largely used to accurately stage subjects in clinical trial and to predict their prognosis. Similarly to classification models, disease progression models rely on machine learning techniques, such as statistical pattern recognition, to learn the behaviour of biomarkers over time, using the learnt patterns to estimate the function that best describes the disease progression and consequently leads to subjects' staging [140].

Paradigms with known physiological connotation can be explored with the aforementioned techniques, by using parametric models (section 2.4.1). Nevertheless, novel scenarios can also be explored, in which it is possible to ponder new paradigms without strong initial assumptions (section 2.4.2). Therefore, disease progression models may not only be used as an approach to generically model a neuroscience problem, but may also be a way to explore and validate new neurophysiological hypothesis. The models used to define disease progression can be summarised as in Table 2.1. Note, however, that the aforementioned table does not detail all the models that may be used for subjects prognosis.

Given that the aims of this thesis are focused on the understanding of CJD, I considered two varieties of disease progression models, including both their parametric and non-parametric formulations. The two models, detailed in the sections 2.4.1 and 2.4.2, explore the main advantages of the models introduce in table 2.1: namely the (1) extraction of a common biomarker trajectory from population samples, (2) interaction between the biomarkers and (3) inclusion of an individual rate of progression. Some applications of these models to neuroimaging data are also presented in these sections.

2.4.1 Parametric Models

Mixed-effect models

For the studies of neurodegenerative diseases, longitudinal studies have proven to be useful to characterise the temporal trajectories of disease-related biomarkers. In fact, longitudinal analysis approaches, such as linear and non-linear mixed-effect models have been largely used to explore the evolution of disease-related features over time [141].

Linear mixed-effect model can handle unbalanced data with high inter-subject variability in the time-points acquisition, as well as with missing data points. Therefore, these types of model offer a parsimonious yet effective approach to model the mean and covariance structure of longitudinal data. In other words, linear mixed effect-model allow the specific variance of a subject's biomarkers to be modelled according to the mean values of these biomarkers in the population. Considering a linear model with a response variable $\mathbf{y} \in \mathcal{Y}$, the distribution of y is given by Equation 2.26, where $\mathbf{w}$ is the vector of known prior weights, $\boldsymbol{\beta}$ is the vector of estimated weights, $\mathbf{X}$ is the matrix of input variables belonging to the feature space $\mathcal{X}$ with

Table 2.1: Summary of the main types of models used to characterise the disease progression. These models are organised according to their application and trajectory shape. A comparison among the different types of models is also presented based on their limitations.

Model description	**Trajectory Shape**	**Subjects Staging**	**Main Limitations**
Stages Comparison	Not modelled	Only classification	Biased categories
Event-based Model	Step-functions	Discrete Stages	No include the time notion
Differential Equations Model	Non-parametric	Disease onset; Rate of progression.	The individual trajectories are not aligned
Disease Progression Score	Sigmoid	Disease onset; Rate of progression.	Assume the sigmoidal assumption.
Self modelling Regression	Non-parametric	Disease onset; Rate of progression.	Assumes the same rate of progression for all the subjects.
Manifold Model	Sigmoid	Disease onset; Rate of progression.	Assumes a parametric shape of the biomarker trajectories; Ignore the interaction between biomarkers.

N subjects and D covariates/features, $\mathbf{o}$ is the vector of offset terms and σ is vector with the noise of the population, also defined as the scale parameter [142].

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{o}, \sigma^2\mathbf{w}^{-1}) \tag{2.26}$$

In a linear mixed-effect model, the model expressed in Equation 2.26 takes the form of Equation 2.27, where $\mathbf{Z}$ is the matrix containing the random effects covariates and $\mathbf{b}$ is the vector of the weights related to these covariates.

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2\mathrm{diag}(\mathbf{w}^{-1})) \tag{2.27}$$

with $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}_N^2\mathbf{D})$, where $\boldsymbol{\sigma}_N^2\mathbf{D}$ is the matrix of variance-covariance. Note that this formulation can be simplified in order to express the response variable in terms of the time variable as:

$$y(t) = \sum_{f=1}^{F} \beta_f X_f(t) + \sum_{r=1,}^{R} b_r Z_r(t) + \boldsymbol{\Sigma} \tag{2.28}$$

where y is the time-dependent response variable for a subject, $\mathbf{x}$*(t)* represents the value of $\mathbf{x}$ at time t, corresponding to the fixed-effects, $\mathbf{Z}$*(t)* represents the value of $\mathbf{Z}$ at time t, which denotes the random-effects, $\boldsymbol{\beta}$ and $\mathbf{b}$ are respectively the coefficients associated to the fixed and random effects, and $\boldsymbol{\Sigma}$ is the independent and identically distributed zero-mean Gaussian measurement noise, with $\boldsymbol{\Sigma} = \sigma^2\mathrm{diag}(\mathbf{w}^{-1})$. The parameters in the model, both $\boldsymbol{\beta}$ and $\mathbf{b}$ are estimated either by maximum likelihood, or by restricted maximum likelihood, based on the marginal density of the response variable [143, 144]. If MEM is a non-linear function in $\mathbf{b}$, then the model is defined as non-linear MEM and the estimation of the parameters requires approximations to the log-likelihood, such as Lindstrom and Bates approximation [144].

MEM are mature approaches, well known in the statistics community. Several studies have shown that MEM are a powerful and versatile framework to analyse real-life longitudinal neuroimaging data, envisioning subjects' prognosis. Bernal-Rusiel *et al.* [141], used linear MEM to analyse clinical longitudinal neuroimaging data. The proposed approach provided a quantitative empirical evaluation of the performance of linear MEM, competing with alternatives popularly used in prior longitudinal structural MRI studies, namely repeated measures ANOVA and the analysis of annualised longitudinal change measures (e.g. atrophy rate). In fact, the results suggested that the linear MEM approach offers superior statistical power in detecting longitudinal group differences, when compared with the aforementioned approaches [141]. Following this approach, Bernal-Rusiel *et al.* [145], introduced an extension of the linear MEM modelling approach to be applied to the mass-univariate analysis of longitudinal neuroimaging data. The proposed method, called spatiotemporal linear MEM or ST-LME for short, builds on the flexible linear MEM framework and exploits the spatial structure in image data. The model was used for the analysis of cortical surface measurements (e.g. thickness). The proposed ST-LME method was validated using two brain MRI datasets obtained from the ADNI and Open Access Series of Imaging Studies (OASIS). The experiments showed that ST-LME increased the statistical power and repeatability of findings, while providing good control of the false positive rate [145].

Luo *et al.* [146], introduced a multilevel response model to analyse the multivariate longitudinal data of mixed types, such as continuous and categorical data extracted from clinical studies. This study analyses several hierarchical joint models with the hazard of terminal events, such as death or clinical dropout, dependent on shared random effects from various levels. Luo and collaborators conducted extensive simulation study to evaluate the performance of various models under different scenarios. The proposed hierarchical joint models were applied to the motivating deprenyl and tocopherol antioxidative therapy of Parkinsonism study to investigate the effect of tocopherol in slowing down the progression of Parkinsons disease [146]. The results had shown that the proposed model provides accurate parameter estimates, in addition to subject-specific disease severity estimation. Furthermore, it also provided additional insight into the correlation between the multivariate longitudinal outcomes and the dependent terminal event at both subject and centre levels. Luo *et al.* [146] also aim to develop a nonparametric statistical model, including a multilevel item response model, to define and estimate the time-dependent treatment effect.

Sabuncu and collaborators had also explored the associations between longitudinal neuroimaging measurements and the time of clinical onset of AD [147]. These associations were tested by using a linear mixed-effect model, in order to capture the temporal variation in serial imaging data. The results obtained by Sabuncu *et al.*, using this model, suggested that linear mixed-effect model can offer excellent statistical power to detect associations between longitudinal imaging features and clinical symptoms [147].

More recently, Rohrer *et al.* [148], have also used linear mixed-effects models in the context of the genetic frontotemporal dementia initiative (GENFI). In this case, a linear mixed-effects model was implemented to examine whether the differences existed between non-carriers and mutation carriers in the association between the clinical scores or the neuroimaging features and the time to expected onset of clinical symptoms. Once more, the obtained results suggest that linear mixed-effects modelling is a powerful tool to analyse longitudinal data, since the model detected measurable markers that showed rates of decline before symptom onset in frontotemporal dementia [139]. Nonetheless the promising results obtained by linear mixed-effect models, these models tend to analyse the contribution of each biomarker independently. Furthermore, as suggested by Bilgel *et al.* [135], linear mixed-effects models do not always account for the fact that subjects enter a study at various disease stages and progress at different rates. Taking into account the limitations of linear mixed-effect models, Bilgel *et al.* [135], have proposed a multivariate non-linear mixed effect model, which accounts for such differences across subjects. Bilgel *et al.*, suggested to adapt the disease score principle to study longitudinal neuroimaging data by making substantial innovations to the progression score model and parameter estimation procedure. In this approach, the progression score s is modelled as a linear function of time t, for each subject and allow for the prediction of separate slopes and intercepts to account for the variability across subjects. The progression score s_{ij}, for the subject i at the time-point (visit) j, is assumed to be an affine transformation of the subjects's age at t_{ij}, as described by Equation 2.29 [135]. The subject-specific variable α_i and β_i account to the differences between subjects in terms of rate of progression and baseline disease progression respectively.

$$s_{ij} = \alpha_i t_{ij} + \beta_i \tag{2.29}$$

The complete version of the model assumes a form a mixed-effect model (Equation 2.30), where $\boldsymbol{f}$ and $\boldsymbol{b}$ incorporate the fixed-effects, and s_{ij} corresponds to the random-effects; the biomarker evolution over time is defined by $\mathbf{y}_{ij}$ and ε is the independent and identically distributed zero-mean Gaussian measurement noise, with variance σ^2.

$$\mathbf{y}_{ij} = \boldsymbol{a} s_{ij} + \boldsymbol{b} + \boldsymbol{\varepsilon}_{ij} \tag{2.30}$$

The results obtained by using this model suggest that the method can be extended to analyse several types of imaging data, to extract individualised summary scores indicating the disease progression and to provide trajectories that may be compared between brain regions. However, the current formulation does not capture dynamic processes over longer periods; one way to overcome this issue is by investigating the relationship between progression scores and time, and by selecting an appropriate function to link these variables. Another limitation of the proposed framework is the assumption of a linear trajectory of the biomarkers over time.

On the other hand, Schiratti *et al.* [149], proposed a generative statistical model for longitudinal data, described in a univariate Riemannian manifold setting, which estimates an average disease progression model, subject-specific time shifts and acceleration factors. The model tackled the limitations of linear MEM, which do not take into account the fact that subjects may be at different stages of disease progression. Schiratti *et al.* [149], considered that the time shifts account for variability in age at disease-onset time, whilst the acceleration factors account for variability in speed of disease progression. Lastly, for a given individual, the estimated time shift and acceleration factor define an affine re-parametrization of the average disease progression model. The model was used to analyse ADNI data. The obtained results showed that the proposed framework can distinguish between slow versus fast progressing and early versus late-onset individuals [149].

Guerrero *et al.* [150], also proposed a framework based on non-linear MEM to derive global and individual marker trajectories for a training population. In detail, the framework consists of two main parts: (1) for a new unseen patient, specific models are instantiated using a stratified *marker signature* that defines a sub-population of similar cases within the training database; (2) from this sub-population, personalised models of the expected trajectory of several markers are subsequently estimated for unseen patients. When applied to the prognosis of subjects to AD, the defined patient-specific models of markers were shown to provide better predictions of time-to-conversion to AD than population based models [150].

In summary, due to the substantial inter-subject variability in clinical studies, models that attempt to describe biomarker trajectories for a whole population will likely lack specificity to represent individual patients. Therefore, individualised models provide a more accurate alternative that can be used for tasks such as population stratification and subject-specific prognosis. These models can be built based on MEM approaches, which consider both the information retrieved from the population data while considering individualised trajectories of the disease progression. Despite the good performance of parametric MEM, these models strongly rely on prior assumptions about the disease progression. Therefore, non-parametric models can be helpful when the knowledge regarding the progression of the biomarkers is limited. Additionally, non-parametric models can also be advantageous when in the presence of noisy and small samples. Therefore, in the next section, two examples of non-parametric models are introduced to tackle the limitations of parametric disease progression models, as well as their application in clinical studies.

2.4.2 Non-parametric Models

Non-parametric Mixed-effect models

Non-parametric MEM are an extension of generalised MEM, however they provide tools to model both the mean and covariance structure non-parametrically for Gaussian distributed data [151]. Attending to the definition of MEM introduced in the Equation 2.27, the conditional distribution of $\mathbf{y}$ on the random effects $\mathbf{b}$ is now given by:

$$\begin{aligned}\mathbb{E}[\mathbf{y}|\mathbf{b}] &= \mu_i \\ \mathbb{V}[\mathbf{y}|\mathbf{b}] &= a_i(\phi)\nu(\mu_i)\end{aligned} \tag{2.31}$$

where y_i is the response variable for the subject i, $\nu(\cdot)$ is a known variance function, and $a_i(\phi)$ are known functions of the dispersion parameter ϕ, and the estimated variance is defined by $\mathbb{V}[\mathbf{y}|\mathbf{b}]$ [151]. Note that μ_i depends on $\mathbf{b}$ random effects, while ϕ is independent of $\mathbf{b}$. Lastly, considering a link function g, the conditional predictor is then modelled by Equation 2.32, where a non-parametric function f is also used to model the fixed effects.

$$g(\mu_i) = f(\mathbf{x}_i) + \mathbf{Z}_i^T\mathbf{b} \tag{2.32}$$

Similarly to parametric MEM, non-parametric MEM have been used aiming the subjects'prognosis in context of neurodegenerative diseases. Donohue *et al.* [152], proposed general semiparametric MEM model to estimate simultaneously the pathological stage (timing) and long-term growth curves for the biomarkers considered. The resulting estimates of long-term progression were fine-tuned using cognitive trajectories derived from the short-term. Using a synthetic dataset, Donohue *et al.* [152], demonstrated that the method can recover long-term disease trends from short-term observations. The method was also used for subjects' staging with respect to disease pathology, providing subject-specific prognostic estimations of the time until onset of symptoms. The method was applied to ADNI data to assess its robustness and effectiveness in real data. The estimated growth curves were in general agreement with prevailing theories of the AD cascade [152].

Dalca and collaborators [153], presented a semi-parametric model. It incorporated the population trend and the subject-specific information to predict the subsequent value of a specific biomarker and also the subsequent anatomical image. The model developed considered the change of a phenotype $\Delta y_t = y_b - y_t$ from the baseline to the timepoint t, using a linear regression (Equation 2.33), where β is the subject-specific regression coefficient, $\Delta\psi_t$ is the time interval between scans, and ε is the independent and identically distributed zero-mean Gaussian measurement noise, with variance σ^2 [153].

$$\Delta y_t = \Delta\psi_t \boldsymbol{\beta} + \varepsilon \tag{2.33}$$

Note however that β is defined by a non-linear mixed-effect model (Equation 2.34), where $\bar{\beta}$ is the global regression coefficient, common to the entire population, and H defines the deviation from this coefficient based on the subject's specific features, namely the genetic information (g), clinical information (c) and the image features (f_b). A Gaussian process was implemented to compute H, as suggested previously by Ge *et al.* [154] using a different kernel covariance k to obtain the kernel matrix $\mathbf{K}$, which encode the features pattern. The parameters of the model are estimated via the restricted maximum likelihood approach, as initially proposed by Harville *et al.* and used in MEM studies [143, 151, 154].

$$\begin{aligned} \boldsymbol{\beta} &= \bar{\boldsymbol{\beta}} + H(g, c, f_b) \\ H &\sim \mathcal{GP}(\mathbf{0}, \tau_H^2 \mathbf{K}) \end{aligned} \tag{2.34}$$

Dalca and collaborators have highlighted the potential of using these methods as a prediction method to be used in context of disease progression models [153].

Schmidt-Richberg *et al.* [155], implemented a non-parametric approach to also model the evolution of the biomarkers related to the progression of AD. Rather than MEM, the authors used a quantile regression to learn statistical models describing the evolution of biomarkers. In this study, two separate models were constructed using (1) subjects that progress from a cognitively normal (CN) stage to MCI and (2) subjects that progress from MCI to AD during the observation window of a longitudinal study. These models were then automatically combined to develop a multi-stage disease progression model for the whole disease course. A probabilistic approach was derived to estimate the current disease progress (DP) and the disease progression rate (DPR) of a given individual by fitting any acquired biomarkers to these models. This method is particularly advantageous since it is applicable even if individual biomarker measurements are missing for the subject. Employing cognitive scores and image-based biomarkers, the presented method is used to estimate DP and DPR for subjects from ADNI. The results showed the potential use of these metrics as features for different classification tasks [155]. Later, Schmidt-Richberg *et al.* [156], applied the proposed approach to three possible applications for disease progress estimation. The authors demonstrated the versatility of the proposed approach, by using it for classification, construction of a spatio-temporal disease progression atlas and prediction of future disease progression, using ADNI data [156].

Gaussian Process Regression

The GP models, introduced in section 2.3.2, allow to define the distribution of a response variable $\mathbf{y}$ over functions and hence can model sequential observations as a function of time. Therefore, GP models are very promising when the time series is hard to discretise in time as is the case with clinical time series data in which observations are often missing and spaced irregularly in time [157]. Furthermore, given their properties, GP models are particularly interesting for the prognosis of

AD and other forms of dementia while used as regression models.

For the particular scenario of disease progression models, GP models fit the patients biomarkers and their clinical time series. As mentioned in section 2.3.2, GP are parameterised by their mean functions and covariance function, where the mean function is the function of time. Since patients may be encountered at different age and under different circumstances (e.g., different stages of the disease), there is no good way to align their time origins. Consequently, the only way to feasibly align them is to set their mean functions equal to a constant as $m(t) = M$, which makes the mean function of a GP time invariant [157]. The mean function can be obtained from the average of all the observations from all the patients, which gives a constant mean reflecting the population data. Differently, the covariance kernel function measures the similarity of two biomarker values $f(t)$ and $f(t)'$ based on their input time t and t'. In general, the covariance function should reflect the properties of the modelled time series, such as its smoothness or periodicity [157].

A study developed by Hyun *et al.*, supports that GPs are a good approach to model longitudinal neuroimaging data [158]. Hyun and collaborators proposed a spatial-temporal Gaussian process framework to accurately delineate the development trajectories of brain structures and function, whilst incorporate the spatial and temporal features of longitudinal neuroimaging data to improve the accuracy and sensitivity of the predictions. Considering a longitudinal dataset with N subjects, a response variable $y_i(d, t_i)$ that corresponds to a specific neuroimaging measure at voxel d and a vector of covariates (such as age, gender and diagnostic status) denoted by $x_i(t_i)$, for the subject i, at the time-point t, the measurement model of a spatial-temporal Gaussian process is defined by:

$$y_i(d,t) = \mu(d, \mathbf{x}_i(t)) + \eta_i(d,t) + \varepsilon_i(d,t), \text{for } i \in 1, \dots, N \tag{2.35}$$

where $\mu(d, \mathbf{x}_i(t))$ is the mean structure that characterises the effects of the covariates defined by $\mathbf{x}_i(t)$; the $\eta_i(d,t)$ are the random functions that characterise both individual features variations from $\mu(d, \mathbf{x}_i(t))$ and the long range dependence of longitudinal imaging data; the $\varepsilon_i(d,t)$ denotes the the errors associated to the local spatio-temporal dependence structure of longitudinal data. It is considered that $\eta_i(d,t)$ and $\varepsilon_i(d,t)$ are independently and modelled by Gaussian processes with mean 0 and mean $\mu(d,t)$, respectively; the variances are defined by the following expressions:

$$\Sigma_\eta((d,t),(d',t')) = \sum_{l=1}^{\infty} \lambda_l \psi_l(d,t)\psi_l(d',t') \tag{2.36}$$

$$\begin{cases} \mathrm{cov}(\varepsilon_{i,k(d)(d,t)}, \varepsilon_{i,k(d')(d',t')}) = \Sigma_\varepsilon((d,t),(d',t');\boldsymbol{\theta}_k) & \text{for} \quad k(d) = k(d') \\ 0 \quad \text{otherwise} \end{cases} \tag{2.37}$$

where $\psi_l(d,t)$ denotes the orthogonal eigenfunctions corresponding to the ordered eigenvalues obtained by a functional principal component.

The estimation procedure is separated into three stages:

1. Estimate the parametric (or nonparametric) regression function $\mu(:,:)$;
2. Estimate the covariance function $\Sigma_n((d,t),(d',t'))$;
3. Estimate the unknown parameters in the covariance model using a restricted maximum likelihood estimation (ReML).

The proposed model had shown good results when applied to real data such as the ADNI cohort, predicting the surface of the lateral ventricle surface with a lower uncertainty when compared with other methods (the model achieved a reduction of error between 9 and 11%) [158].

Lorenzi *et al.* [159], used disease progression model to quantify the diagnostic uncertainty of individual disease severity in an hypothetical clinical scenario, with respect to missing measurements, biomarkers, and follow-up information. The model was formulated within a probabilistic setting, via a GP regression model. This study had shown that the subjects staging provided by the model was in agreement with the clinical diagnosis [159]. Further, using follow-up measurements, they were able to largely reduce the prediction uncertainties. This approach had also shown that the transition from healthy to pathological stages is mostly associated with the increase of brain hypo-metabolism, temporal atrophy, and worsening of clinical scores. The results presented by Lorenzi *et al.* [159], suggest that GP regression models provide an accurate probabilistic assessment of the pathological stage of unseen individuals, while representing a valuable instrument for identifying the clinical value of biomarkers across disease stages.

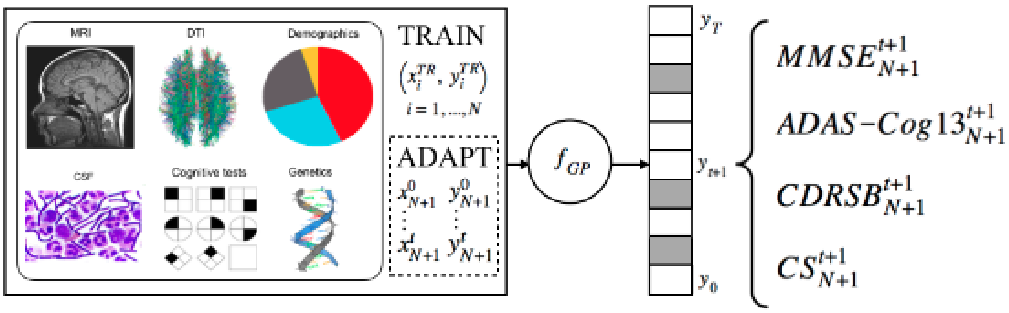


Figure 2.6: Personalised GPs model proposed by Peterson *et al.* The population model is first trained using all past visits data of N patients $(x_{TR}; y_{TR})$, where the time difference between two visits is 6 months. The model personalisation to the target patient *(N+1)* is then achieved by sequentially adapting the model predictions of the future metrics $y_t + 1$ (using the posterior distribution of GPs -fGP), informed by the visits data up to time stept. The shaded fields in the output represent the time-points for which no visit data is available for a given patient. Image adapted from [87].

More recently, Peterson *et al.* [87], introduced a personalised GP (pGP) to predict the main biomarkers of the AD progression (MMSE, ADAS-Cog13, CDRSB and CS) based on each patients previous visits. In this study, a time-point consists in a patients visit, which refers to the data collected at a single time-point sample during the ADNI cohort. The model is initialised by learning a population model using multi-modal data of previously seen patients using the base GP regression approach. Then, this model is adapted sequentially over time to a new patient using the notion of domain adaptive GPs. The main contribution of this approach is the novel adaptation strategy for personalising the GP population model, as detailed in Figure 2.6. The results presented by Peterson *et al.*, leads to significant improvements in the prediction performance of the future clinical status and cognitive scores for target patients when compared to the population [87].

These studies support that GP regression models are interesting to study neurodegenerative diseases, particularly to be used as prognosis tool and anticipation of clinical onset.

Nevertheless, GP models also come with limitations, namely the fact that the mean function of the GP is a function of time and in order to make the GP independent from the time origin, which needs to defined as a constant value. However, this significantly limits the model ability to represent changes or different modes in time series dynamics, hampering its used in context of scenarios where the time normalisation is not straightforward.

In summary, mixed-effect models, both the parametric and non-parametric forms of these models, are still one of the most common approaches to characterise disease progression as a continuous time-dependent function. However, these models have still limitations, since most of them assume a parametric shape of the biomarkers trajectories and a common progression for all the subjects, ignoring the potential diversity of symptoms and rates of progressions presented by the subjects. Moreover, these models tend to define different functions for each biomarker, assuming the independence of the observed features. Conversely, Bayesian frameworks, namely the Gaussian processes formulations, are a practical way to process the longitudinal data, since they are the appropriate choice to model time-series, particularly in the presence of missing time-point or for making long term prediction. However, these models require a more complex formulation in the presence of time inconsistencies. Currently, the models used for subjects' prognosis are still hampered by the need to provide an initial temporal alignment of the samples.

2.5 Summary

Machine learning techniques have been broadly used on neuroimaging data to classify, stratify and predict the outcome of patients. Both parametric and non-parametric model have shown good results in the diagnosis of neurodegenerative diseases, such as AD.

However, the current models, detailed in this chapter, are not appropriate for Prion disease analysis, hence the need of a new model that may be able to overcome the limitations of the current frameworks and to be able to model conveniently this illness. The new formulation will need to account for (1) different rates of progression of the disease, (2) interactions and correlations between biomarkers, (3) heterogeneity of features even among patients at the same stage of the disease, and (4) independent of the time origin in the samples considered. The following chapters tackle the limitations of the existent models, adapting the current formulations to the specificities of CJD.

Chapter 3

Imaging biomarkers for CJD description

Contents

In the clinic environment, the diagnosis of CJD relies on the visual read of MRI scans, as they present signal abnormalities induced by micro-structural changes caused by CJD. The lack of quantitative biomarkers, as well as the difficulty to accurately identify the onset of the disease and the fast rate of progression of CJD, have limited the clinical understanding of the progression of CJD [6, 160] and the development of automated tools for diagnosis and prognosis.

In an attempt to overcome the aforementioned issues, research is directed towards identifying the right biomarkers that characterise and discriminate the illness. In this chapter, I introduce a framework to extract and select relevant imaging biomarkers from MR images. The proposed framework aims the extraction of subject-specific biomarkers, and it is validated on a cohort composed by both the sCJD and IPD forms of prion disease. I also implement conventional approaches for dimensionality reduction, in order to demonstrate the advantages of a subject-specific feature selection framework to characterise CJD.

The proposed framework is validated on the National Prion Monitoring Cohort (NPMC) dataset. The dataset is detailed in section 3.1. Section 3.2 describes the methods used for feature extraction. The results of the feature selection are described in section 3.3. Lastly, the correlation between the features obtained from both the group-wise approaches and the proposed method and the clinical scores is analysed in section 3.4.

3.1 National Prion Monitoring Cohort (NPMC)

The data used in this study were obtained from the NPMC. NPMC (2008-) is a prospective observational interval-cohort study of patients with any form of prion disease in the UK or willing to travel to the UK. It includes regular follow-up clinical and psychological assessments of sCJD patients, patient with IPD and their relatives, who may be known carriers of *PRNP* gene mutations, at-risk but not had a genetic test or healthy controls. The current dataset comprises (a) symptomatic patients with confirmed prion disease, for both the inherited and sporadic forms of the disease; (b) healthy subjects without a clinical diagnosis of IPD who carry *PRNP* gene mutations and are therefore at increased risk of disease in the future, defined in this study as asymptomatic subjects; (c) healthy individuals without a confirmed diagnosis but at increased risk, (d) healthy individuals without either prion disease or increased risk, defined as healthy controls (HC). From the aforementioned sample, I defined a group composed by the subjects at clinical onset (CO): subjects within one year of a clinical diagnosis and an MRC scale of 20, including both symptomatic subjects with no severe neurodisability within this time frame and asymptomatic patients with the diagnosis later confirmed. This new group is used to examine specific brain changes occurring close to the clinical onset. To avoid the overlap of criteria used to defined both the CO and IPD groups, the IPD group is composed by symptomatic subjects with MRC Scale equal or lower than 20, in which the scans were acquired outside the time frame specified; i.e., one year or more after clinical onset.

The data from the 125 subjects include MRI scans, neurological and neuropsycological assessment and scoring on the MRC Scale [21]. MRI was acquired using a Siemens Magneton Trio (Siemens, Erlanger, Germany) 3 Telsa with conventional body coil for transmission and a 32-channel head-only receive coil. Structural imaging used 3D T1-weighted images (T1w) MPRAGE sequence with repetition time 2.2 s, echo time 2.9 ms, inversion time 900 ms, echo spacing 6.7 ms, flip angle 10°, matrix size $256 \times 256 \times 208$, voxel size $1.1 \times 1.1 \times 1.1$ mm. 2D Axial FLAIR were acquired using a standard clinical FLAIR-TSE sequence with a voxel size of $0.9 \times 0.9 \times 5.2$ mm. The diffusion weighted imaging (TR/TE 9500/93ms) were acquired using 64 non-colinear directions at $b = 1000\text{s/mm}^2$ and 8 images with b=0. For all subjects a T1w image was acquired as well as either a FLAIR, a DWI, or both. The sample's demographics are detailed in Table 3.1. The quality of the MR images was assessed visually. None of the 125 scans had shown significant artefacts that would lead to the exclusion of these subjects from this study. Lastly, an independent sample of healthy controls is used for normalisation purposes (as detailed in section 3.3). The sample comprises MRI data acquired at the Dementia Research Centre London, using the MRI machine mentioned above. The scanning protocol includes the acquisition of 3D T1w MRI, FLAIR and DWI. Both T1w and FLAIR were acquired with the same protocol used for the acquisition NPMC dataset, whist the DWI was acquired using multiple shells. For better harmonisation, this work only used the shell that had the most similar b-value (b=700) to the one used for the prion data acquisition (b=1000). The similarity of the MRI acquisition protocols as well as the use of the same scanner ensures the viability of using this sample for data normalisation. However, I acknowledge the limitation raised by the differences in the DWI acquisition as a potential bias in the results of the feature selection. The data demographics of this sample are detailed in Table 3.1.

3.2 Features Extraction

The framework is designed to extract quantitative features from the three MRI pulse-sequences: T1w, FLAIR, DWI for each subject. The different sequences provide complementary information about brain microstructural changes caused by CJD. The framework, Figure 3.1, consists of three sections: (A) data pre-processing that includes artefact correction, bias field correction, correction of the effects of eddy currents in DWI scans and rigid registration; (B) and (C) specific feature extractions according to type of MRI sequence and quantification. In section (A) both DWI and FLAIR scans are rigidly registered to T1w scans using the *NiftyReg* open-source software [161].

Table 3.1: Demographic and imaging information of subjects in the baseline of NPMC, YOAD database[a] and an independent control sample[b], included in this study. The full model is the model trained using only the subjects with all the three MRI sequences available. The number of mutations details the number of different mutations existing among the subjects.

		Groups	Age (years)	Full Model (Male)	T1w (Male)	FLAIR (Male)	DWI (Male)	#Mutations
		Independent healthy controls	47.4 (23 - 67)	91 (40)	91 (40)	91 (40)	91 (40)	—
		Healthy Controls	48.2 (23.3 - 75.2)	29 (16)	31 (16)	29 (16)	26 (16)	—
Diagnosis		IPD	47.7 (24.9 - 61.4)	16 (11)	30 (18)	21 (12)	18 (11)	8^{a}
		Sporadic CJD	63.7 (53.3 - 76.7)	17 (10)	28 (15)	20 (11)	17 (10)	—
		YOAD	61.0 (48.0 -74.0)	—	32 (10)	—	32 (10)	—
Stratification		Asympt. IPD	42.7 (19.5 - 72.3)	22 (6)	31 (11)	29 (16)	22 (6)	6^{b}
		Clinical Onset	50.7 (41.6 - 65.2)	4 (3)	5 (3)	5 (3)	4 (3)	3^{b}
	Stage I	IPD	44.5 (24.9 - 61.3)	15 (9)	20 (11)	20 (12)	16 (10)	7^{a}
		sCJD	63.9 (54.3 - 75.7)	5 (3)	5 (3)	7 (4)	5 (3)	—
	Stage II	IPD	26.1	1 (0)	3 (2)	1 (0)	1 (0)	1^{d}
		sCJD	62.5 (53.3 - 71.5)	11 (6)	21 (12)	19 (11)	19 (10)	—

[a] 6-OPRI, 5-OPRI, A117V, D178N, E196K, P102L, Y163X
[b] 6-OPRI, 5-OPRI, A117V, D178N, E200K, P102L
[c] D178N, E200K, E196K
[d] P102L

[a] The YOAD database is not part of the NPMC. The details regarding imaging acquisition and data available are detailed in Chapter 5, where this data is used for clinical validation of the proposed approaches.

[b] The independent control sample is used for normalisation purposes. It was been acquired using the same scanner and imaging protocol than NPMC data. The need for this sample is motivated in section 3.3.

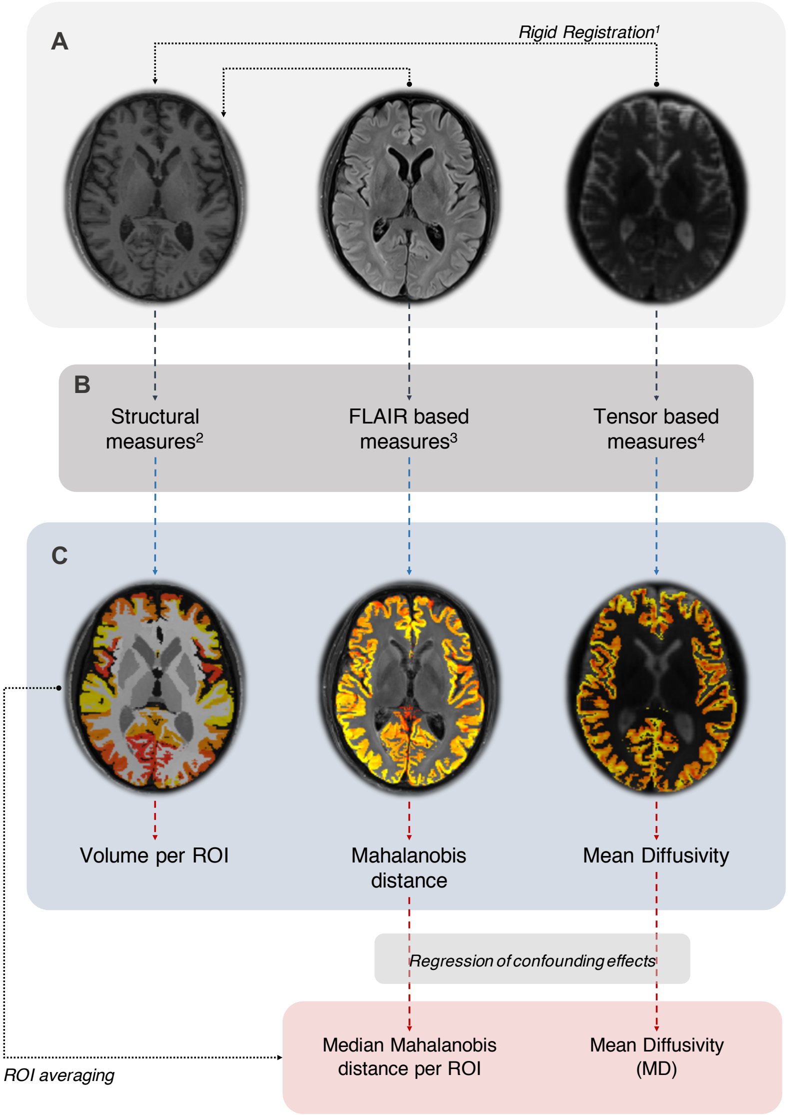


Figure 3.1: A: data pre-processing step, including rigid registration using (1) *NiftyReg* [161]. B: Feature extraction per MRI sequence, applying (2) GIF algorithm [162] to T1, using (3) BaMoS algorithm to extract the intensity distributions of FLAIR [163], and computing the diffusion tensor from DWI using (4) *NiftyFit* [164]. C: The quantitative features were computed from the images obtained in the section B of the framework.

.

To identify nerve cell loss and consequently atrophy of cortical and deep GM areas, I extracted volumetric information from T1w MRI scans using automated region of interest delineation. The Geodesic Information Flows (GIF) [165] algorithm, that relies on multi-atlas segmentation propagation, is used to parcellate the brain into multiple regions. The volume of each 128 individual brain region is then computed. These regions are fully detailed in Appendix D. Hyperintensity abnormalities visible on FLAIR images need to be carefully considered since the degree and distribution of these histological changes tend to vary significantly among the different time of scanning [166, 167]. To characterise the degree of abnormality in each subject's brain, I consider as a feature the distribution of signal intensities in GM tissues in FLAIR images. Using the Bayesian Model Selection (BaMoS) algorithm [163], I automatically segment the normal and abnormal appearing tissue types. Knowing that CJD mainly causes lesions in the GM tissues, I compute the Mahalanobis distance [168], between the normal appearing WM intensity distribution and the GM intensities for each region of interest as defined by T1w derived parcellation. The Mahalanobis distance per region, $\mathrm{d}_{\mathrm{M(GM,WM)}}$ is computed as

$$\sqrt{(\mu_{\mathrm{GM}} - \mu_{\mathrm{WM}})^{T} \cdot \mathrm{S}_{\mathrm{WM}}^{-1} \cdot (\mu_{\mathrm{GM}} - \mu_{\mathrm{WM}})}, \tag{3.1}$$

where μ_{GM} is the mean of intensities in each GM region, μ_{WM} is the mean intensity of WM tissue after excluding the lesions detected as outliers, and S_{WM} corresponds to the covariance of the WM tissue distribution.

The obtained values are a quantitative measure of signal abnormalities in GM and they can be used as a feature with the assumption that the larger the amount of hyper-intensity in a given region of interest, the larger the Mahalanobis distance. The assessment of the hyper-intensities in the brain mimics the clinical practice, in which CJD is typically diagnosed based on the presence of these signal abnormalities.

The most typical brain microstructural change caused by CJD is vacuolation, or spongiosis. Spongiosis can result from abnormal membrane permeability and increased water content within neuronal processes; however, the molecular mechanisms behind vacuolation are still unclear [169]. Spongiosis is visible in DWI scans as an increase in the diffusion signal and it can be quantified using the MD measurements. I initially process the DWI scans using the *NiftyFit* pipeline, described in [164], in which MD measurements are computed according to [170]. The median MD value per ROI is computed and used as a feature.

I regress out the impact of confounding effects, such as age and the total intracranial volume, by comparison with a healthy population. This correction is applied *a priori* to all the features extracted from different sequences.

I evaluate the statistical significance of the features extracted before feature selection. Table 3.2 shows the brain regions and their respective *p*-value computed using a two sample t-test comparing the different groups with the healthy population. Neither the Asymp. nor the CO presented any significant difference in features compared to the healthy controls. The features extracted from FLAIR are insufficient to identify brain regions that are relevant to diagnose CJD.

Table 3.2: The two sample t-test was used to identify which brain regions show significant differences between symptomatic subjects and the healthy population. The p-values indicate the test rejection of the null hypothesis at 5% significance level, considering the Bonferroni correction.

DWI		
	Brain Regions	**P-value**
IPD	Right cuneus	1.74E-5
	Left central operculum	2.48E-5
	Right anterior cingulate gyrus	2.91E-5
	Right inferior frontal gyrus	3.54E-5
	Right angular gyrus	3.61E-5
sCJD	Right frontal operculum	1.03E-6
	Left entorhinal area	1.05E-6
	Cerebellar Vermal Lobules VI-VII	1.99E-6
	Cerebellar Vermal Lobules I-V	2.35E-6
Structural		
	Brain Regions	**P-value**
IPD	Left cuneus	6.01E-8
	Right cuneus	7.55E-5
	Left central operculum	5.03E-6
sCJD	Left cuneus	2.45E-9
	Right cuneus	1.11E-6
	Left central operculum	3.89E-6
	Left hippocampus	5.00E-5
	Right hippocampus	5.49E-5

Figure 3.2 shows the *p*-values of the brain regions significantly different from the healthy population, projected into the MNI152 linear template [171] using the MRIcroGL visualisation software[1]. The brain regions are considered as significantly different from the healthy population for the *p*-value $< 3.80 \times 10^{-04}$, after Bonferroni

[1] https://www.nitrc.org/plugins/mwiki/index.php/mricrogl:MainPage

correction, presented in the Figure 3.2 with light orange.

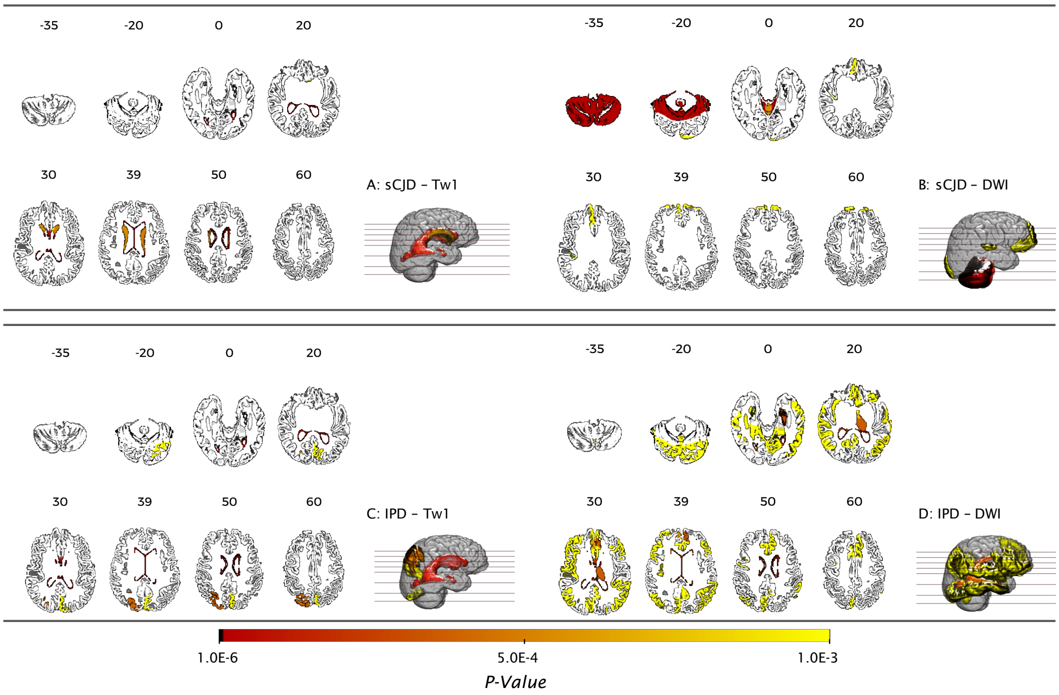


Figure 3.2: The colour map encodes the p-value obtained from the two sample t-test, for each brain region showing . A: structural features extracted from IPD subjects; B: DWI features obtained from IPD scans; C: sCJD structural features; D: DWI features extracted from sCJD data.

3.3 Features Selection

In neuroimaging studies, the samples size is often quite small, when compared with the feature space dimension. As a result, the numbers of potentially available features greatly outnumber the observations.

The study of CJD using MRI data is as well hampered by the reduced number of samples available. The number of observations/subjects available is in fact much lower than the number of feature extracted: 89 subjects (among healthy controls and symptomatic subjects) for 128×3 features. Due to the reduced number of samples, the problem of finding correlation between subjects at the same stage of the disease by using the set of features extracted is ill-posed.

To prevent the aforementioned issues, I reduce the dimensionality of the feature space, before using any machine learning model. A set of different feature selection approaches are compared in the following sections. Given the conflicting signals from different MRI sequences, the feature selection is performed independently for each sequence. The resulting features will give complementary information about the undergoing physiological processes in the brain. Two different hypothesis are considered to select the most relevant features: (1) group-wise approaches and (2) subject-specific feature selection.

3.3.1 Group-wise biomarkers selection

For most neurodegenerative diseases, group-wise approaches are used to select the relevant features. This can be done as the features are consistent and homogeneous across subjects and throughout the disease stages. Even though CJD had shown in clinical practice to be highly heterogeneous across subjects, I compare three approaches – an unsupervised method and two supervised approaches – commonly used to select relevant features from neuroimaging data.

Stepwise regression

Stepwise regression can be used as a feature selection method as it selects the best subset of models to explain the response variable. It is a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression. The method begins with an initial model and then compares the explanatory power of incrementally larger and smaller models. At each step, the p-value of an F-statistic is computed to test models with and without a potential term. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model.

The assessment of the subset of models can be performed either by forward stepwise, backward selection or bidirectional elimination. In brief, the forward selection of variables chooses the subset models by adding one variable at a time to the previously chosen subset. Forward selection starts by choosing as the one-variable subset the independent variable that accounts for the largest amount of variation in the dependent variable. This will be the variable having the highest simple correlation with $\mathbf{y}$ response variable. At each successive step, the variable in the subset of variables not already in the model that causes the largest decrease in the residual sum of squares is added to the subset. In contrast, backward elimination of variables chooses the subset models by starting with the full model and then eliminating at each step the one variable whose deletion will cause the residual sum of squares to increase the least. This will be the variable in the current subset model that has the smallest partial sum of squares.

I implemente a bidirectional elimination scheme to identify the most relevant brain structures in the brain to identify CJD. In this experiment, both IPD and sCJD are considered. The asymptomatic and subjects at clinical onset were excluded avoiding noisy labels during the feature selection step. The feature selection was performed for each MRI pulse-sequence independently. The initial feature space, composed by 128 features, was reduced to 15 features.

The results of the stepwise selection, Figure 3.3, show that only the DWI data has relevant information to identify symptomatic subjects among healthy controls. These results are in agreement with the clinical assumptions, since in clinical environment DWI is taken as the most reliable imaging data. Both T1w and FLAIR data seem to not have significant information to identify CJD.

These results sustain the assumption that a group-wise approach is not sensitive enough to select meaningful features to identify the heterogeneous features of CJD. However, since the subset of models produce by stepwise regression can be oversimplifications of the real models of the data [172], I apply also a more conservative model that considers not only the correlation between the features and the response variable, but also the correlation between the features in order to validate the assumption that group-wise approaches are not appropriate to select the features to characterise CJD, as detailed below.

Least absolute shrinkage and selection operator and Elastic net

Both least absolute shrinkage and selection operator (LASSO) and Elastic Net are feature selection techniques that combine both machine learning and feature selection steps by enlisting a regularisation framework. Their formulation includes a penalty term that constrains the size of the estimated coefficients [173]. Therefore, they resemble ridge regression except in the fact that both techniques set more coefficients to zero when the penalty term increases, which results in a model with fewer predictors. As such, they are a good alternative to stepwise regression or other model selection and dimensionality reduction techniques [174, 175].

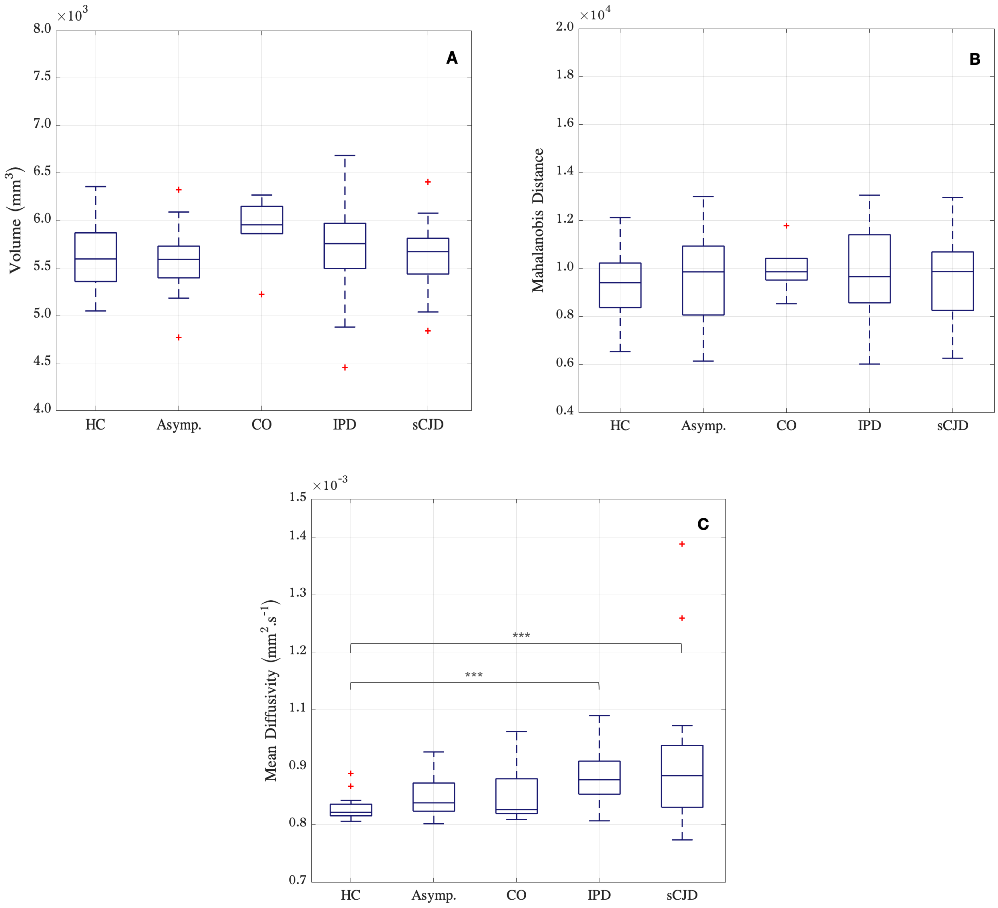


Figure 3.3: Dimensionality reduction using a stepwise regression. The initial feature space was reduce to 15 dimensions. The mean of the 15 features is computed across subjects per group. A: Structural features obtained from T1w MRI; B: Features extracted from FLAIR scans; C: biomarkers computed using the DWI scans. The red crosses represent outliers, whilst the grey asterisks represent a statistical significance of $p - value < 0.001$, after Bonferroni correction. HC – healthy controls; Asymp. – asymptomatic subjects; CO – clinical onset; IPD – inherited prion disease and sCJD - sporadic CJD.

LASSO technique solves the problem expressed by Equation 3.2, where N is the number of observations, y_i is the response at observation i, $\mathbf{x}_i$ is data, a vector of p predictors at observation i; λ is a positive regularisation parameter, which controls the sparsity of the model and consequently the generalisation ability of the model. The parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ are scalar and p-vector size respectively, and as λ increases, the number of nonzero components of $\boldsymbol{\beta}$ decreases.

$$min_{\boldsymbol{\beta}_0,\boldsymbol{\beta}}\left(\frac{1}{2N}\sum_{i=1}^{N}(y_i-\boldsymbol{\beta}_0-\mathbf{x}_i^T\boldsymbol{\beta})^2+\lambda\sum_{j=1}^{p}|\,\boldsymbol{\beta}_j\,|\right) \tag{3.2}$$

LASSO has proven to be very successfully in neuroimaging studies, since it is able to cope with a large number of predictors and fewer observations, as it yields a small set of model coefficients with the majority of coefficients set to zero [104, 176]. However, LASSO tends to select only one variable when a group of predictors are highly correlated, whilst the elastic net encourages grouping effect in presence of highly correlated predictors and it has no limit of the number of variables selected.

Elastic net (Equation 3.3) is an hybrid of ridge regression and LASSO regularisation. Like LASSO, elastic net can generate reduced models by generating zero-valued coefficients. The *L1* penalty promotes the sparsity, whilst *L2* enforces the stability of the solution and acts as a bound on the number of features selected [174, 175].

$$min_{\boldsymbol{\beta}_0,\boldsymbol{\beta}}\left(\frac{1}{2N}\sum_{i=1}^{N}(y_i-\boldsymbol{\beta}_0-x_i^T\boldsymbol{\beta})^2+\lambda P_\alpha(\boldsymbol{\beta})\right) \tag{3.3}$$

where

$$P_\alpha(\boldsymbol{\beta})=\frac{(1-\alpha)}{2}\parallel\boldsymbol{\beta}\parallel_2^2+\alpha\parallel\boldsymbol{\beta}\parallel_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\boldsymbol{\beta}_j^2+\alpha\mid\boldsymbol{\beta}_j\mid\right)$$

By using the elastic net in place of LASSO, I am not only concerned with the selection of the most significant biomarkers to explain the data but also to find the ones with more relevance in the disease characterisation and their interaction and mutual influence. In fact, based on the formulation of elastic net algorithm, I can define the interaction between the biomarkers (covariates) through the definition of the parameter α. I define an α equal to 0.75 and the λ value is optimised in each step of the framework using a nested cross-validation algorithm to avoid overfitting[2]. The experimental design adopted here is equivalent to the experimental design used for stepwise selection approach.

The elastic net only identified relevant features from DWI data (Figure 3.4). Among the identified brain regions, the basal ganglia (caudate, insula, amygdala and frontal opercullum) was the most significant region to distinguish the symptomatic

[2] The nested cross-validation is a convenient approach to train a model in which the hyperparameters also need to be optimised. Note that the nested cross-validation estimates the generalisation error of the underlying elastic net model and its (hyper)parameter search. The model selection, aiming the feature selection, without employing the nested cross-validation, would use the same data to tune the model parameters and evaluate the model performance. This set-up can potentially "leak" relevant information into the model and overfit the data. The magnitude of this effect is primarily dependent on the size of the dataset and the stability of the model [177]. Therefore, to prevent this effect, I employ a nested cross-validation to fit the model hyperparameters during the model selection.

patients from controls, with an associated *p-value* below to 2.9×10^{-5}. These results are in agreement with previous studies, which also identified the basal ganglia as abnormal for CJD patients.

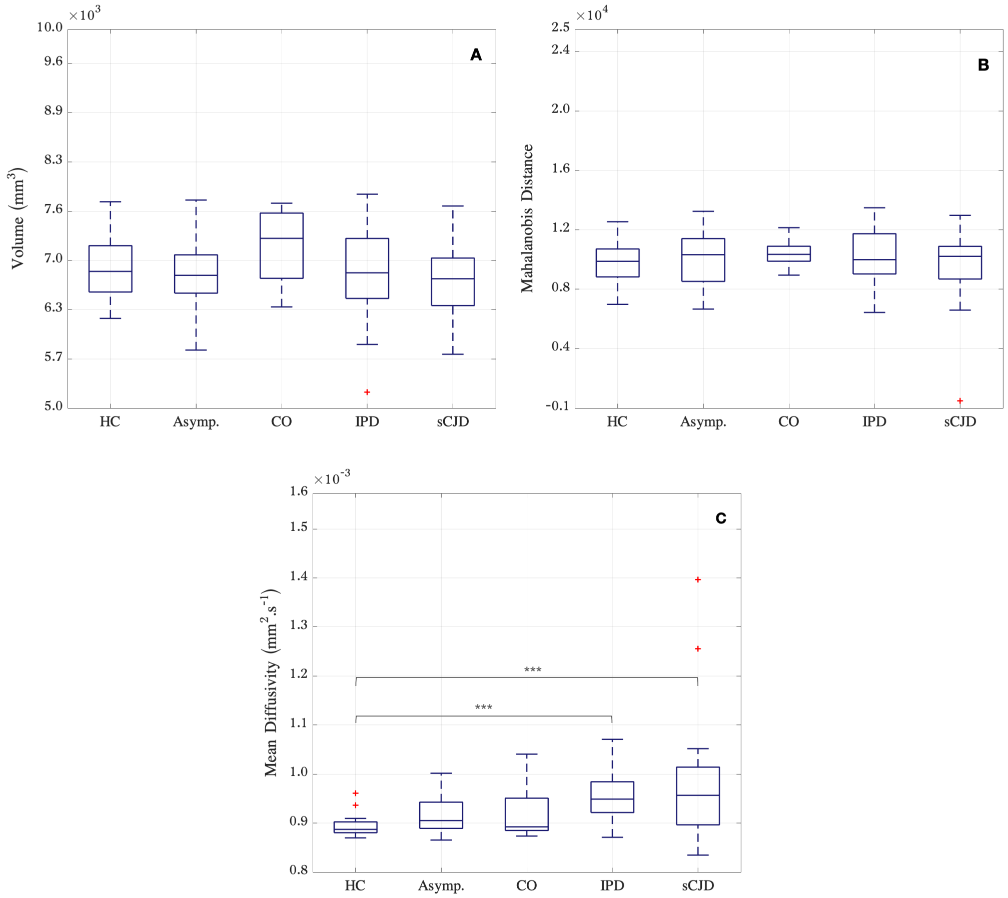


Figure 3.4: Feature selection via LASSO. The initial feature space was reduced to 15 dimensions. The mean of the 15 features is computed across subjects per group. A: Structural features obtained from T1w MRI; B: Features extracted from FLAIR scans; C: biomarkers computed using the DWI scans. The red crosses represent outliers, whilst the grey asterisks represent a statistical significance of $p\text{−}value < 0.001$, after Bonferroni correction. HC – healthy controls; Asymp. – asymptomatic subjects; CO – clinical onset; IPD – inherited prion disease and sCJD - sporadic CJD.

Nevertheless, this approach still excludes all the potential features from T1w and FLAIR images. The potential reasons for that are: (1) the heterogeneity of features discussed above, or (2) the existence of noisy labels due to different stages of the disease of the symptomatic subjects considered.

To analyse the impact of the noisy labels in the selection of features, I implement an unsupervised feature selection method, described in the following section.

Isometric feature mapping

The isometric feature mapping or *isomap*, is a nonlinear dimension reduction procedure. Its underlying principle is to embed a set of observations in an Euclidean feature-space while preserving as close as possible their intrinsic metric structure: the geodesic distances between points on the observation manifold [178].

The algorithm is initialised with the computation of the Euclidean distance $d_{\mathcal{X}}(i,j)$ between all pairs i,j from N data points in the high-dimensional features space $\mathcal{X}$. In this first step, the algorithm determines which points are neighbours on the manifold , based on the distances $d_{\mathcal{X}}(i,j)$ between all pairs i,j. Each point is connected to all points of its K-nearest neighbours. These neighbourhood relations are represented as a weighted graph $\mathcal{G}$ over the data points, with edges of weight $d_{\mathcal{X}}(i,j)$. Secondly, the *isomap* estimates the geodesic distances $d_M(i,j)$ between all pairs of points on the manifold $\mathcal{M}$ by computing their shortest path distanced $d_G(i,j)$ in the graph $\mathcal{G}$. This step is initialised by $d_G(i,j) = d_{\mathcal{X}}(i,j)$ if (i,j) are linked by an edge, whereas $d_G(i,j) = \infty$. Then for each value of $K = 1,2,...,k$ in turn, replace all entries $d_G(i,j)$ by $\min\{d_G(i,j), \quad d_G(i,k) + d_G(k,j)\}$. The matrix of final values $\mathbf{D_G} = \{d_G(i,j)\}$ will contain the shortest path distances between all pairs of points in $\mathcal{G}$. The final step applies classical multi-dimensional scaling to the matrix $\mathbf{D_G}$, constructing an embedding of the data in a d–dimensional Euclidean output space $\mathcal{X}'$ that best preserves the manifold's estimated intrinsic geometry. The coordinate vectors x_i for points in $\mathcal{X}'$ are chosen to minimise the cost function $E = ||\tau(\mathbf{D_G})\tau(\mathbf{D}_{\mathcal{X}'})|L_2$ where $\mathbf{D}_{\mathcal{X}'}$ denotes the matrix of Euclidean distances $\{d_{\mathcal{X}'}(i,j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||\}$ and $||\boldsymbol{A}||L2$ is the $l2$- matrix norm which takes the form of $\sqrt{\sum_{i,j} \mathbf{A}^2_{i,j}}$. The τ operator[3] converts distances to inner products, which leads to a more efficient optimisation [179, 180].

I have created an *isomap* representation of the feature space that better characterises the group of subjects at a specific stage of the disease. The initial feature space $\mathcal{X}_m$, where m is the MRI sequence, $m \in \{\text{T1w}, \text{FLAIR}, \text{DWI}\}$ was reduced to 15-dimensional feature space, similarly to the aforementioned approaches. Figure 3.5 shows the mean value per group of the 15-highest distances, selected as features.

The results show that the *isomap* is able to identify meaningful features from all MRI pulse-sequences. Consequently, the unsupervised model can be more appro-

[3] For two vectors $\mathbf{x}$ and $\mathbf{x'}$, the τ operator transforms a distance into a inner product as $\tau||\mathbf{x}-\mathbf{x'}|| = \sqrt{<\mathbf{x}-\mathbf{x'}, \mathbf{x}-\mathbf{x'}>}$.

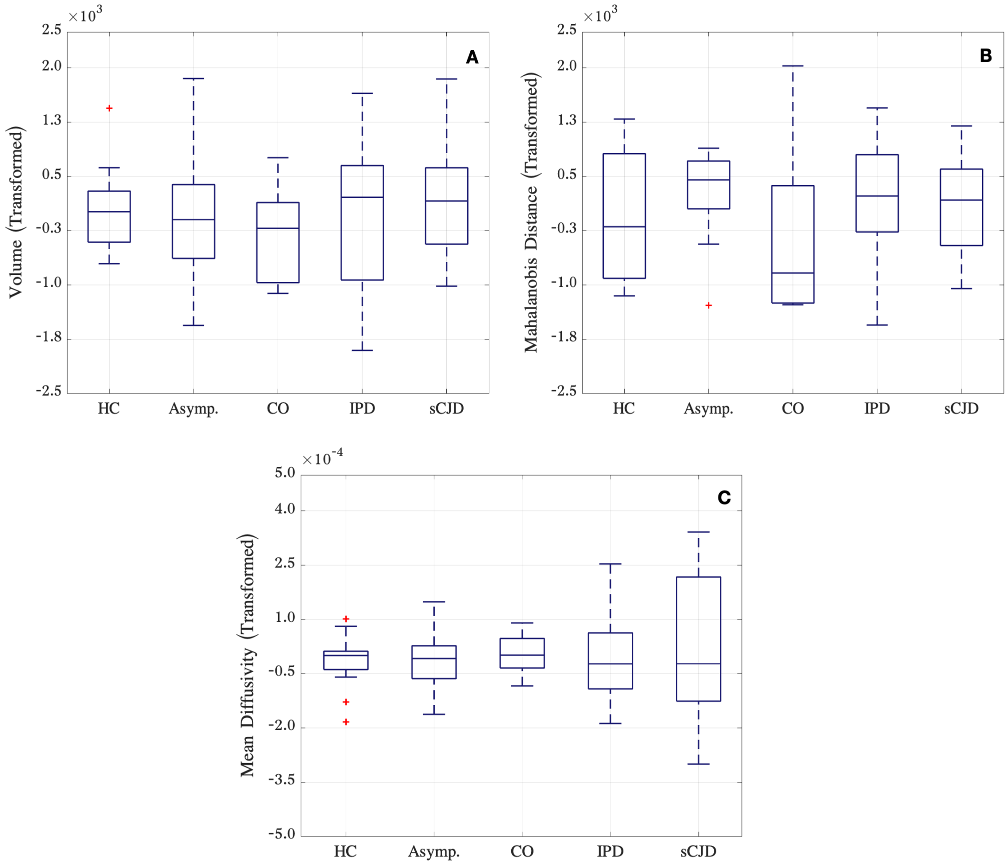


Figure 3.5: Feature selected using the *isomap*, a non-linear dimensionality reduction technique. The initial feature space was reduced to 15 dimensions. A: Structural features obtained from T1w MRI; B: Features extracted from FLAIR scans; C: biomarkers computed using the DWI scans. The red crosses represent outliers, whilst the grey asterisks represent a statistical significance of $p\!-\!value < 0.001$, after Bonferroni correction. HC – healthy controls; Asymp. – asymptomatic subjects; CO – clinical onset; IPD – inherited prion disease and sCJD - sporadic CJD.

priate to select the relevant features to characterise CJD, when compared with the supervised approaches. Furthermore, the results show that the noisy labels hamper the performance of the supervised models, which leads to a bad performance of embedded feature selection methods, namely the methods detailed in section 3.3.1.

Nonetheless, the use of *isomap* as feature selection method would require a more extensive analysis in order to justify the features obtained from MRI data.

3.3.2 Subject-specific biomarkers

Due to the assumption of spatial heterogeneity of brain changes caused by prion disease, I implement a subject-specific features extraction and selection. This approach selects the most significant features to characterise the evolution of symptoms for each subject, neglecting the spatial origin of features in the brain. In detail, given the hypothesis that the disease does not follow a geometrical pattern in the brain, the quantification of abnormality rather than its location is thus used to quantify the progression of the disease[4]. To characterise the amount of abnormality of signal for the different types of feature, I apply a framework previously validated with IPD data [181], in which the different features were converted into z-scores by comparison with measurements obtained from an independent population of healthy subjects, described in Table 3.1. The use of an independent sample ensures that the normalisation of the features via z-scores is not biased towards the control population when the identification of symptomatic patients. The z-scored values are then ranked per type of feature and only the highest values for each MRI sequence are considered for subsequent learning and inference stages. As a consequence, only regions of the brain that most differ from the healthy control sample are kept for each subject, and the resulting sets of feature are subject-specific. Figure 3.6 shows the mean of the 15 most significant features per subject and across groups.

In more detail, Figure 3.7 shows the probability density of the regions selected for each group, smoothed by a kernel density estimator. The Figure 3.7 details also the unbalanced sample size, already mentioned in Table 3.1; e.g., the clinical onset group has a much smaller sample than the remaining groups. For all groups, T1w and FLAIR show a higher density in the first quartile that is caused for smaller z-scores values. Thus, the features extracted from these MRI pulse-sequences are less relevant to characterise symptomatic subjects. On the other hand, DWI features have broader probability density, showing a higher dispersion of the values of the selected features particularly for the symptomatic subjects. These results are explained by the different stages of the disease presented by the patients.

Given that the subject-specific approach is able to find several significant features from T1w and DWI scans, I investigate which set of features allows for the best differentiation of the groups using a multi-comparison test. The resulting p-values are corrected for multiple comparison using the Bonferroni method. Table 3.3 presents the detailed results of the statistical analysis. The mean of the highest ranked features extracted from both T1w and DWI are significantly different across

[4] This assumption is based on previous clinical studies, where it was reported considerable variability in radiologic patterns for sCJD [19]. Given the similarities of the MRI findings for both sCJD and IPD [46, 52], the assumption of variability of radiologic patterns can be extended to IPD subjects.

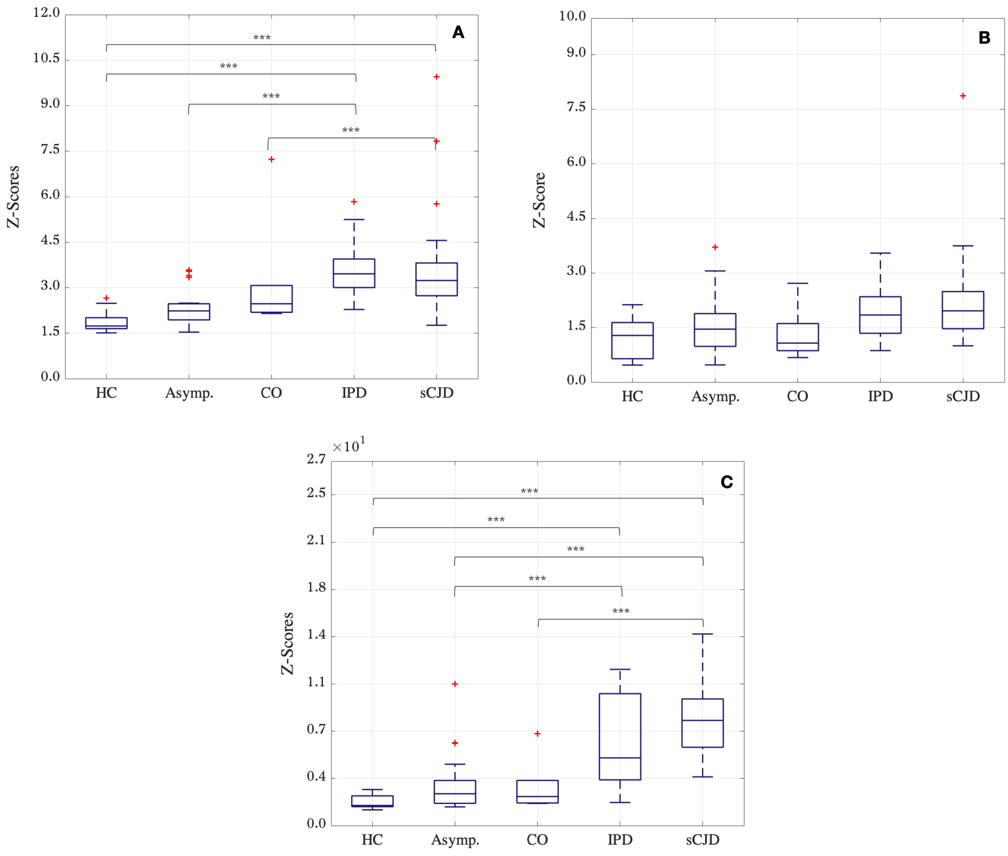


Figure 3.6: Subject-specific feature selection. The initial feature space was reduced to 15 dimensions. A: Structural features obtained from T1w MRI; B: Features extracted from FLAIR scans; C: biomarkers computed using the DWI scans. The red crosses represent outliers, whilst the grey asterisks represent a statistical significance of $p - value < 0.001$, after Bonferroni correction. HC – healthy controls; Asym. – asymptomatic subjects; CO – clinical onset; SI – stage I and SI - stage II of the disease.

groups, whereas the features extracted from FLAIR images have not shown statistical significance across groups. Furthermore, this experiment also indicates that DWI and T1w features enable the diagnosis of sCJD vs HC with high statistical significance, whilst the T1w features identify the disease stage of IPD (*versus* Asymp.) with highest statistical significance.

3.4 Correlation between imaging features and MRC Scale

The MRC Scale captures the clinical features of Prion disease, summarising in a single value the subjects' performance in a set of functional and physiological

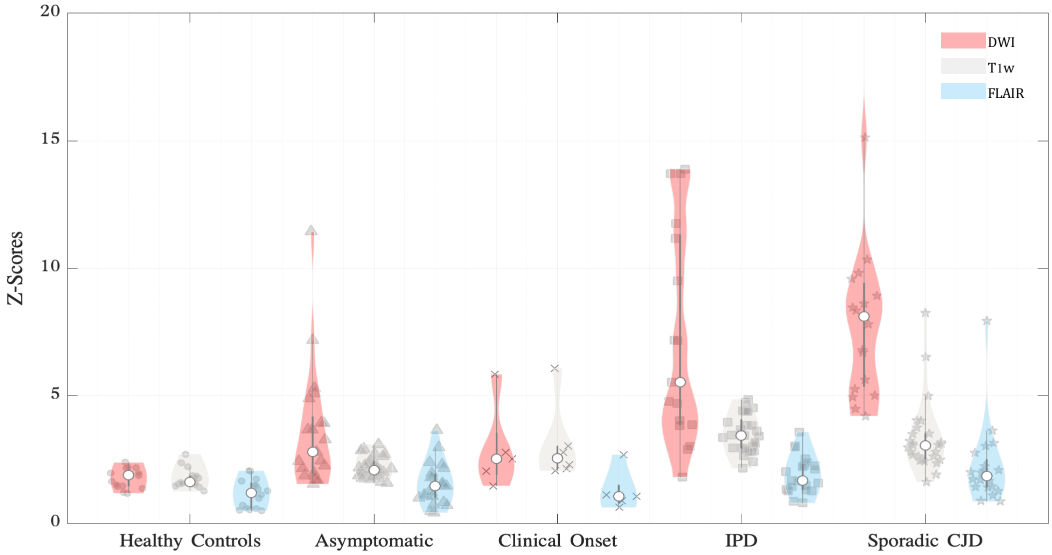


Figure 3.7: The initial feature space was reduce to 15 dimensions. The mean of the 15 features is computed across subjects per group. A: Features extracted from FLAIR scans; B: biomarkers computed using the DWI scans; C: Structural features obtained from T1w MRI. •: Healthy control subjects; ▲: Asymptomatic subjects; ×: Clinical onset subjects; ■: inherited prion disease symptomatic subjects; ★: sporadic CJD; ∘: Median of the distribution.

Table 3.3: Evaluation of the statistical significance of the imaging biomarkers, after feature selection. The Kruskal-Wallis test result is shown with the null hypothesis that the sample data from each group of subjects came from the same distribution. The bold p-values indicate the test rejection of the null hypothesis at 5% significance level, considering the Bonferroni correction, $p-\text{value} < 3.80 \times 10^{-4}$.

	T1w	FLAIR	DWI
HC *vs* Asym.	0.622	0.838	0.986
HC *vs* Conv.	0.082	0.941	0.242
HC *vs* IPD	**9.92E-09**	0.004	**2.53E-06**
HC *vs* sCJD	**5.45E-07**	0.008	**9.96E-09**
Asym. *vs* CO.	0.828	0.999	0.695
Asym. *vs* IPD	**5.63E-08**	0.195	**7.55E-04**
Asym. *vs* sCJD	0.006	0.266	**8.22E-07**
CO. *vs* IPD	**7.32E-05**	0.166	0.065
CO. *vs* sCJD	0.191	0.228	**6.71E-04**
IPD *vs* sCJD	0.115	0.999	0.728
All groups	**2.47E-17**	6.67E-04	**2.21E-13**

HC – healthy controls; Asym. – asymptomatic subjects; CO – clinical onset; SI – stage I and SI - stage II of the disease.

tests [21]. Currently, this appears to be the most valuable tool for assessing disease progression [33]. To determine the correlation between the features selected and the MRC scale, I implemented a non-parametric statistical test between the feature sample and the MRC Scale scores. Table 3.4 shows the results of the Spearman' ρ

for each feature selection method. The features obtained from the subject-specific approach show a higher correlation with the MRC scale, for both FLAIR and DWI features. LASSO method is although more efficient in selecting features extracted from T1w, which are highly correlated with the MRC scale.

As previously detailed, the feature selection is performed based on sample composed by both IPD and sCJD, for all the methods tested. Despite this experimental design, I tested the correlation for each group separately in order to evaluate possible differences between the two groups. The analysis of Table 3.4 suggests that for certain type of features, such as DWI, the correlation for the sCJD patients is negative whereas the IPD subjects show a positive correlation with the MRC Scale.

Table 3.4: Correlation between the selected features and the clinical score, MRC Scale. The correlation between the average of the 15 highest features per group of subjects and the MRC scale is assessed using a Student's t distribution for a transformation of the correlation. Spearman's ρ is presented for both IPD and sCJD forms of CJD. Grey colour highlights the highest correlations, either negative or positive correlation.

	T1w		FLAIR		DWI	
	IPD	*sCJD*	*IPD*	*sCJD*	*IPD*	*sCJD*
z-scores	0.199	-0.358	-0.395	-0.555	0.354	-0.517
Stepwise Regression	0.286	-0.245	0.055	0.549	-0.183	0.245
Isomap	0.146	-0.244	0.181	0.235	0.176	0.215
LASSO	0.564	0.485	0.078	0.368	-0.206	-0.148

3.5 Discussion

Feature Extraction

Previous studies have shown signs of atrophy in temporal, cingulate, frontal, parietal and occipital lobes caused by IPD [15]. By using the structural biomarkers obtained via the proposed framework, I also identified statistical differences in the occipital gyrus, specifically in the cuneus, for both IPD and sCJD. The left and right hippocampus and central opercullum had been identified as meaningful regions to identify CJD, as suggested by De Vita and collaborators [34]. However, I was not able to identify signs of atrophy in the temporal and parietal lobes. My analysis also identified signal abnormalities in DWI scans. Statistical significant differences were observed in the sCJD sample, when compared with healthy controls, in the left and right entorhinal areas, cerebellar vermal lobus I-VII. In turn, DWI signal differences were observed in the right cuneus, anterior cingulate gyrus, angular gyrus and central operculum for IPD subjects. Previous studies [38, 39], have reported signal abnormalities in DWI scans in the caudate, putamen and pulvinar nuclei. However, this study did not reveal statistically significant changes in those regions. This can be justified by the small dataset, or the segmentation of these regions that might have compromise the feature extraction, excluding relevant features.

By comparing the imaging features extracted from healthy controls and symptomatic patients, I observed that CJD disease burden weights equally on each hemisphere. Furthermore, despite the initial assumption of spatial heterogeneity of the brain changes, I identified some regions with higher prevalence among subjects with the same form of CJD (Figure 3.2). These results are explained by the broad spectrum of symptoms stages found in the IPD and sCJD groups, which leads to conflicting MRI signals [166]: paradoxical normal MRI appearances are observed in some brain regions showing pathological changes such as gliosis and spongiosis.

Knowing that CJD mainly causes lesions in the GM tissues, I only computed imaging biomarkers extracted from this brain tissue. As future work, it could be explored the use of WM features, in order to evaluate the effectiveness of these features in contrast with GM features, as suggested by [35]. Furthermore, a future study could also benefit from a larger range of features, including cortical thickness, voxel-based morphometry, fractional anisotropy measurements and clinical features such as blood and CSF biomarkers.

Feature Selection

It is very challenging to select useful biomarkers that may be used to comprehensively characterise all the different subtypes of prion disease and to perform an accurate diagnosis because of the heterogeneity of the clinical manifestations of CJD. Therefore, in this chapter I compared several techniques to establish what is the best framework to select the relevant imaging biomarkers to diagnose CJD.

The supervised feature selection techniques were able to identify relevant features from DWI data. These results are consistent with previous studies, where DWI had shown to be more sensitive to capture brain abnormalities caused by CJD than T1w and FLAIR data [37, 39, 53]. On the other hand, the tested unsupervised method was not able to identify meaningful imaging biomarkers from all the MRI sequences. These results endorse the assumption that supervised methods tend to perform better than unsupervised methods, even when in the presence of few noisy labels and subjects at different stages of the disease equally labelled. Furthermore, the features extracted from T1w and DWI data are insufficient to characterise sCJD patients. These findings support the clinical assumption that prion disease is highly heterogeneous even among subjects with the same mutation; whereby the group-wise methods, previously used in context of neurodegenerative diseases [72, 88, 89, 94, 118], tend to dilute relevant signals that could be used as features in a classification model.

Bearing this in mind, I adopted a subject-specific feature selection method. By selecting subject-specific biomarkers, I ensured that the lack of spatial pattern of biomarkers does not compromise the extraction and selection of features that track subtle brain changes. This feature selection method has also proven to be more efficient in the detection of relevant features to identify symptomatic CJD among healthy controls, when compared with other embedded feature selection methods, as presented in Appendix A. In that supplementary analysis, the subject-specific feature selection showed better results than using an automatic relevance determination (ARD) approach, when implemented in a classification task. The extracted imaging biomarkers (section 3.3.2) have shown significant differences between healthy controls and symptomatic subjects, for both IPD and sCJD. Nonetheless, the intensity based features, computed from FLAIR images, did not show statistical relevance to separate symptomatic subjects from healthy controls, after Bonferroni correction for multiple comparisons.

Note though that by using absolute z-scores, I can potentially hamper the clinical relevance of the features extracted, specifically for the MD measurements. In detail, different regions of the brain can show abnormal MD measurements, at different stages of the disease. However, depending on the microstructural changes caused by the disease, these values can be higher or smaller, when compared with a normal population, evidencing the increasing or decreasing of the diffusivity linked to either spongiosis or astrocytis gliosis changes, respectively. These brain changes, depending on the process that is causing it, can also be correlated with changes in other MRI pulse-sequences, such as T1w. By taking only the absolute z-scores, the information regarding the specific process leading to the abnormal MD values is ignored, as well as its correlation with other features. Therefore, this fact can compromise the specificity of the DWI features, while preserving their sensitivity in the diagnosis of CJD.

The major limitation of this approach is the impossibility to assess what are the brain regions selected as being statistical significantly abnormal since they differ for each subject. In the future, by using a bigger sample, the frequency of the brain regions identified as abnormal and its correlation with the stage of the disease should be investigated.

Correlation between imaging features and MRC Scale

To determine the validity of the features considered as a biological meaningful marker of disease progression and severity, I evaluated the correlation between the features selected to characterise CJD and the clinical scores, such as MRC Scale. Given that the MRC Scale measure of the severity of the clinical manifestations of the disease, a strong correlation between the features selected and this score shows that they are appropriate quantitative measures to be used in clinical context. Furthermore, this experiment also gives insights regarding the best feature selection to be translated in clinical context. From the analysis of Table 3.4, the subject-specific feature selection, z-scores, is the method which extracts the imaging biomarkers better explain the clinical scores, particularly for DWI and FLAIR data.

Previous studies [33], have also shown a strong correlation between the loss of brain tissue (atrophy) and the MRC Scale. However, the group-wise approaches, such as LASSO, seem to be better able to retrieve the features that are higher correlated with the MRC Scale, when compared with subject-specific approaches. Note that De Vita *et al.*, have reported a positive correlation between the tissue volume and the MRC Scale difference in a given period of time; i.e., they found a significant correlation between tissue volume change between the first and last examination, and change in MRC Scale over the same period, across several brain areas, with decreases in MRC Scale accompanied by decreases in local GM and/or, WM tissue volumes [33]. My analysis compares the increasing of abnormality magnitude with the decrease of the MRC Scale score. As a result, my experiments should report a negative correlation when Z-score is used, and a positive correlation for the remaining methods. However, the IPD subjects show a positive correlation between the DWI features and MRC Scale. These results can be explained by the reduced number of IPD subjects with MRC Scale score lower than 16 (only one subject has MRC Scale score below the mentioned threshold), which compromises the statistical power of the Spearman-rank test. In fact, 82% of the IPD sample (14 out of

17 subjects) have MRC Scale score ≥ 18. Consequently, a bigger sample of IPD subjects, and/or a sample with more dispersed MRC Scale scores would be required to properly validate the obtained features.

Lastly, the features extracted from sCJD subjects' data show a higher correlation with the clinical scores when extracted from FLAIR and DWI; conversely, the IPD subjects show a higher correlation for features extracted from T1w images. These results suggest that the T1w can be more informative in the early stages of IPD, whereas DWI and FLAIR are more reliable to characterise the progress of the disease for sCJD subjects. Nonetheless, further analyses are required to validate these hypotheses.

3.6 Summary

Based on the clinical assumption that CJD is highly heterogeneous even among subjects with the same mutation, I chose to extract subject-specific biomarkers. The biomarkers extracted had shown significant differences between healthy controls and symptomatic subjects, for both IPD and sCJD.

Currently, only MRI features are considered. For a better understanding of CJD, quantitative features from other sources could be included, such as blood and CSF biomarkers. However, this is highly dependent of acquisition of more data.

To assess if these features obtained from Chapter 3 are valid and solid biomarkers considered for subjects diagnosis, stratification and differential diagnosis of CJD, I used them as input features in a classification tool. To that end, Chapter 4 introduces a Bayesian framework in which both genetic, demographic and imaging biomarkers are combined within a Gaussian Process classifier, used to calculate the probability of a subject to be diagnosed with CJD. This model assess not only the validity of the imaging biomarkers to be used as features to identify CJD, but also their reliability to be used to characterise the evolution of the clinical manifestations. Therefore, in Chapter 5, I introduce an extension of the model presented in Chapter 4 used to stratify subjects, further validation the imaging biomarkers proposed for CJD characterisation.

Chapter 4

Diagnosis of CJD using Gaussian Process

Aiming to characterise the disease status of each subject based on their multi-source features, I designed a Bayesian framework to find the function that better fits the relationship between imaging features and the subjects' diagnosis. Bayesian frameworks, such as GP, are particularly interesting to study the CJD, since they allow robust modelling even in the circumstances of highly uncertain or incomplete datasets. GP is also able to perform predictions over the long term, being able to be successively better as the number of samples increases [182]. Since GP is a probabilistic model, it also provides an estimation of the likelihood of the predicted class for each subject. Note that the class probability estimations are a measure of the confidence of the predictions, which can be extremely useful in clinical context as a proxy of the diagnosis precision and as an indicator of subjects' prognosis.

The details of the model are described in section 4.1, including the definition of the model parameters and their optimisation, as well as the details regarding the inference method used. The performance of the proposed framework was assessed in section 4.2, where the NPMC was used for validation purposes. The clinical relevance of the framework is discussed in section 4.3, where the performance of other comparable machine learning approaches for the diagnosis of CJD are discussed.

4.1 Model Description

I implement a non-parametric kernel-based model $\mathcal{M}$, as follow:

$$\begin{gathered}\mathcal{M} : y = f(\mathbf{X}) + \varepsilon, \\ f \sim \mathcal{GP}(\mu_f; \mathbf{K} + \mathbf{I}\sigma_f), \;\; \varepsilon \sim \mathcal{N}(\mu_\varepsilon; \sigma_\varepsilon)\end{gathered} \tag{4.1}$$

This model is used to predict the probability of the outcome $y_i \in \mathcal{Y}$, for a subject $i = \{1, \ldots, N\}$, given a set of biomarkers $\mathbf{X} \in \mathcal{X}$ feature space. For the binary discriminative case, such as subjects' diagnosis, the output of the regression model $\mathcal{M}$ is transformed into a class probability using a cumulative density function, *probit* likelihood function, which converts its argument that can lie in the domain $(-\infty, \infty)$ into the range [0, 1]. This procedure guarantees a valid probabilistic interpretation. Therefore, the posterior probability of each class $\mathcal{C}$ for a subject i is then given by Equation 4.2, where $\boldsymbol{\Phi(.)}$ denotes the cumulative density function of the standard normal distribution [112].

$$p(y_i|f(\mathbf{x}_i)) = \boldsymbol{\Phi}(y_i f(\mathbf{x}_i)) = \int_{-\infty}^{y_i f(\mathbf{x}_i)} \mathcal{N}(x\,|0,1)\,dx \tag{4.2}$$

The function f describes the variance of the feature space $\mathcal{X}$ that explains the response variable y. By implementing a GP with a prior mean function $\mu_f = 0$ and covariance kernel matrix $\mathbf{K}$, I determine the pattern of the inductive generalisation of the feature under consideration [112]. For this approach, given the normalisation of the input feature space before the estimation of the model, it is reasonable to assume a GP prior with mean 0.

As detailed in section 2.3.2, the estimation of the model requires the definition of the covariance kernel function (section 4.1.2) and the estimation of its parameters, via marginalisation of the likelihood function (section 4.1.1). By using the optimised parameters, the class probability for a new subject j is given by the approximate predictive mean for the latent function f_j, as demonstrated in section 4.1.3. The full framework is detailed in Figure 4.1, including the encoding of the feature space.

4.1.1 Marginal Likelihood approximation

For the purposes of subjects diagnosis, the likelihood of $p(y_i|f(\mathbf{x}_i))$ is a cumulative density function, hence the posterior (Equation 4.3) is analytically intractable.

$$\begin{aligned} p(\boldsymbol{f}|\mathbf{X}, \boldsymbol{y}) &= \frac{1}{Z} p(\boldsymbol{f}|\mathbf{X}) \prod_{i=1}^{N} p(y_i|f_i) \\ Z = p(\boldsymbol{y}|\mathbf{X}) &= \int p(\boldsymbol{f}|\mathbf{X}) \prod_{i=1}^{N} p(y_i|f_i) d\boldsymbol{f} \end{aligned} \tag{4.3}$$

To address this issue, I use the expectation propagation algorithm (EP) [125] to

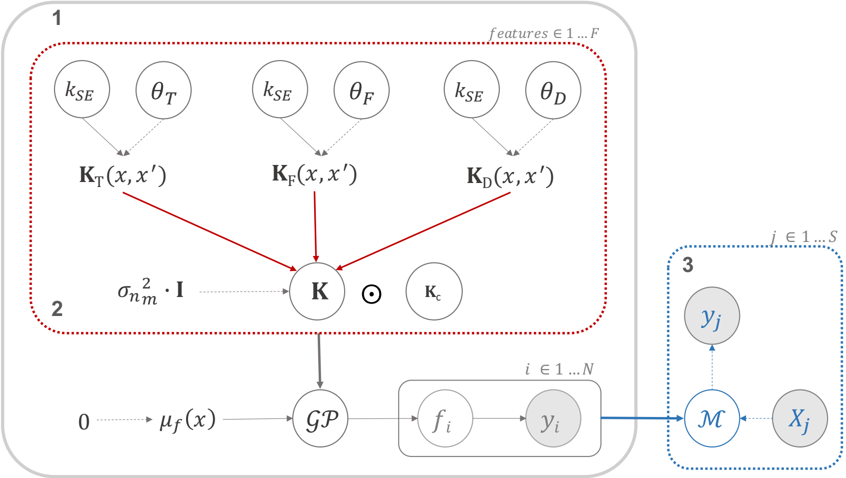


Figure 4.1: Scheme of the generative model used for subjects diagnosis - equation 4.1. The inner section (red outline, component 2) illustrates the definition of the kernel function (section 4.1.2). The component 1 (grey outline) corresponds to the estimation of the hyperparameters of the model detailed in section 4.1.1. The component 3 (blue outline) corresponds to the inference stage of the framework, in which a predictive label for a new subject j is computed using the optimised model $\mathcal{M}$, as described in section 4.1.3. The kernel matrices are estimated via a SE kernel function. The obtained matrix $\mathbf{K}_m$, where m includes T, F and D, encodes the multi-source of features T1w, FLAIR and DWI, respectively.

approximate the likelihood by a local likelihood approximation as:

$$p(y_i|f_i) \simeq t_i(f_i|\widetilde{Z}_i,\widetilde{\mu}_i,\widetilde{\sigma}_i^2) \triangleq \widetilde{Z}_i\mathcal{N}(f_i|\widetilde{\mu}_i,\widetilde{\sigma}_i^2) \tag{4.4}$$

where $\widetilde{Z}_i,\widetilde{\mu}_i,\widetilde{\sigma}_i^2$ are site parameters, as defined by Rasmussen *et al.*, [112]. Note that the tilde-parameters denote the local likelihood approximations, whilst the plain notations is used for the approximate posterior. Considering the Equation 4.4, the posterior is then approximated by $q(\boldsymbol{f}|\mathbf{X},\boldsymbol{y})$ as:

$$\begin{gathered} q(\boldsymbol{f}|\mathbf{X},\boldsymbol{y}) = \frac{1}{Z_{\text{EP}}}p(\boldsymbol{f}|\mathbf{X})\prod_{i=1}^{N} t_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_{p_i}^2) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ \text{with} \quad \boldsymbol{\mu} = \Sigma\widetilde{\Sigma}^{-1}\widetilde{\boldsymbol{\mu}} \quad \text{and} \quad \Sigma = (\mathbf{K}^{-1}+\widetilde{\Sigma}^{-1})^{-1} \end{gathered} \tag{4.5}$$

Finally, the normalisation term $Z_{\text{EP}} = q(\boldsymbol{y}|\mathbf{X})$ can be rewritten as defined by equation 4.6.

$$Z_{\text{EP}} = q(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{f}|\mathbf{X})\prod_{i=1}^{N} t_i(f_i|\widetilde{Z}_i,\widetilde{\mu}_i,\widetilde{\sigma}_i^2)d\boldsymbol{f} \tag{4.6}$$

Considering the formulation detailed by Rasmussen *et al.* [112], the marginal likelihood is:

$$\begin{aligned}\log(Z_{\text{EP}}|\boldsymbol{\Theta}) = -\frac{1}{2}\log|\mathbf{K}+\tilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\tilde{\boldsymbol{\mu}}^T(\mathbf{K}+\tilde{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}} + \sum_{i=1}^{N}\log\Phi\left(\frac{y_i\mu_{-i}}{\sqrt{1+\sigma_{-i}^2}}\right) + \\ +\frac{1}{2}\sum_{i=1}^{N}\frac{(\mu_{-i}-\widetilde{\mu}_i)^2}{2\times(\sigma_{-i}^2-\widetilde{\sigma}_i^2)}\end{aligned} \tag{4.7}$$

where the $\boldsymbol{\Theta}$ denotes the hyperparameters of the covariance function, and $-i$ refers to all the cases except i. Note that first two terms are the marginal likelihood for a regression model for $\tilde{\boldsymbol{\mu}}$, each component of which has independent Gaussian noise of variance $\tilde{\boldsymbol{\Sigma}}$. The remaining three terms come from the normalisation constants $\widetilde{Z}_i$. The first of these penalises the cavity distributions for not agreeing with the classification labels. These approximation is not specific for the model presented in this chapter, hence it is an adaptation of the EP presented by Rasmussen *et al.* [112], fully described at *Gaussian Processes for Machine Learning* book. The estimation of the predictive labels requires to find the best hyperparameters of each kernel covariance function. The hyperparameters $\boldsymbol{\Theta}$ of the kernel functions are estimated via the maximisation of the marginal likelihood of the model, $p(\boldsymbol{\Theta}|Z_{EP})$, as described in Equation 4.8.

$$\begin{aligned}\{\hat{\boldsymbol{\Theta}}\} = \text{argmax}_{\boldsymbol{\Theta}}\, p(\boldsymbol{\Theta}|Z_{EP}) = \\ \text{argmin}_{\boldsymbol{\Theta}}[-\log p(Z_{EP}|\boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta})]\end{aligned} \tag{4.8}$$

4.1.2 Kernel function definition

The covariance kernel function is responsible for encoding the assumptions about the model that is learned. Therefore, it is crucial to define a covariance kernel function that can conveniently explain the feature space, and it is appropriated to describe the evolution of CJD. As demonstrated in section 3.2, the CJD phenotype is better explained by the interaction between several types of features; thus, a basis kernel function is insufficient to describe the variance of the features. Besides, it is reasonable to assume that the features extracted from one MRI sequence do not show a consistent relationship with the features extracted from the subsequent, during all stages of the disease [166]. Assuming that, the inter MRI sequence relationship can be modelled as a multi-task paradigm[1] – a contribution of independent functions that explain the biomarkers progression. A sensible way to model a GP as a multi-task model is using an Additive GP. By implementing an Additive, GP the proposed approach is able (1) to express superposition of different processes contributing for the some output and (2) to improve model interpretability, since it allows to learn the weightings of different functions and their orders of interaction [113, 183]. The latent function f in model $\mathcal{M}$, Equation 4.1, takes then the form of $\boldsymbol{f} = \sum_{m=1}^{M} f_m$, with $f_m \sim \mathcal{GP}(\mu_{f_m}; \mathbf{K}_m + \mathbf{I}\sigma_{f_m})$, where M refers to the number of sources of features – MRI pulse-sequences – taken into consideration in the model. Given the kernel properties, the addition of GP with $\mu_f = 0$ is equivalent to $f \sim \mathcal{GP}(0; \sum_{m=1}^{M} \mathbf{K}_m + \mathbf{I}\sigma_{f_m})$. Therefore, the matrix $\mathbf{K}$, Figure 4.1, which encodes the imaging biomarkers, is obtained by the addition of the kernel $\mathbf{K}_m$ matrices computed individually using the information extracted from the MR pulse-sequences. The imaging biomarkers, encoded in individuals kernel matrices $\mathbf{K}_m, \quad m \in \{\text{S, F, T}\}$ for T1w, FLAIR and DTI respectively, Figure 4.1, using a squared exponential covariance function (SE), Equation 4.9, with hyperparameters $\boldsymbol{\theta} = \left[\sigma^2, l^2\right]$. The SE function is widely-used within binary classification problems, given its main assumptions: smoothness and stationarity.

$$k_{SE}(x, x'|\boldsymbol{\theta}) = \sigma^2 \exp\left(-\frac{1}{2}\frac{(x - x')^2}{l^2}\right) \tag{4.9}$$

The model also accounts for the individualised pattern related to the genetic mutation of the inherited form of CJD, defined as a categorical variable in the kernel matrix $\mathbf{K}_c$. To reduce the bias introduced by the high number of genetic mutations, I grouped the subjects in two clusters according to the expected rate of disease progression associated with each mutation: (1) slow, and (2) fast, defined based in the clinical knowledge[2] about the different mutations, Equation 4.10. For IPD subjects the rate of progression varies as mentioned; whereas the sCJD subtype is always considered as having a fast progression.

$$k_c(x, x') = \begin{cases} 1 & \text{if} \quad x - x' = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

[1] Contrarily to a multi-task feature selection proposed by [120], I extracted the features independently to preserve the complementary information given by the MRI sequences, as proposed and detailed in section 3.6.

[2] According to clinical experience, the slow progression is seen for A117V, P102L, Y163X, 5– and 6–OPRI, whereas E200K, D178N, E196K and sporadic CJD evolve fast.

It is important to note that this is not actually genetic information, but rather a cluster of mutations with similar physiological behaviour. Using this kernel with the aforementioned information, I intent to show the flexibility of my model to deal with both categorical and continuous data, such as genetic and quantitative imaging data respectively. The $\mathbf{K}_m$ is lastly combined with the categorical covariance function by means of the Hadamard product, $\mathbf{K}_c \odot \mathbf{K}_m$ to produce a hierarchical model[3].

The modified latent function $f(\mathbf{X}|\mathbf{\Theta})$ regarding the mutation information is $f \sim \mathcal{GP}(0; \sum_{m=1}^{M}(\mathbf{K}_m + \mathbf{K}_c \odot \mathbf{K}_m) + \mathbf{I}\sigma_{f_m})$, where $\mathbf{\Theta}$ is the vector of parameters of the model, which includes the hyperparmeters of the kernel functions and the sample variance: $\mathbf{\Theta} = \left[\boldsymbol{\theta}_S, \boldsymbol{\theta}_F, \boldsymbol{\theta}_T, \sigma_f^2\right]$.

4.1.3 Predictions

In the final section of the proposed model, illustrated in Figure 4.1 by component 3 (blue outline), I use the optimised model to estimate the predictive label y_{*_j} for a new subject j[4]. The approximate predictive distribution for the binary classification is given by:

$$q(y_* = 1|\mathbf{X}, \boldsymbol{y}, \mathbf{x}_{*j}) = \int \mathbf{\Phi}(f_*)q(f_*|\mathbf{X}, \boldsymbol{y}, \mathbf{x}_{*_j})df_* \tag{4.11}$$

Solving the integral as demonstrated by Rasmussen *et al.* [112], the predictive probability is given by Equation 4.12, which gives a clinical reference to the status of the subject, regarding the highest ranked quantitative biomarkers. Similarly, the mean and variance of function f_* is computed respectively using Equations 4.13 and 4.14. The analysis of the latent models that compose f_* provide the information about the best combination of features to diagnose prion disease.

$$q(y_* = 1|\mathrm{X}, \boldsymbol{y}, \mathbf{x}_{*j}) = \mathbf{\Phi}\left(\frac{\mathbf{k}_{*_j}^T(\mathbf{K} + \widetilde{\mathbf{\Sigma}})^{-1}\widetilde{\boldsymbol{\mu}}}{\sqrt{1 + k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{*_j}^T(\mathbf{K} + \widetilde{\mathbf{\Sigma}})^{-1}\mathbf{k}_{*_j}}}\right) \tag{4.12}$$

$$\mathbb{E}_q\left[f_*|\mathbf{X}, \boldsymbol{y}, \mathbf{x}_{*_j}\right] = \mathrm{k}_{*_j}^T(\mathbf{K} + \widetilde{\mathbf{\Sigma}})^{-1}\widetilde{\boldsymbol{\mu}} \tag{4.13}$$

$$\mathbb{V}_q\left[f_*|\mathbf{X}, \boldsymbol{y}, \mathbf{x}_{*_j}\right] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathrm{k}_{*_j}^T(\mathbf{K} + \widetilde{\mathbf{\Sigma}})^{-1}\mathrm{k}_{*_j} \tag{4.14}$$

[3] As suggested by Luo *et al.* [146], the hierarchical joint models can help to study different behaviours of the same global model regarding a specific variable. This study has shown better results when compared with other non-hierarchical models, once applied in a Bayesian framework design to study Parkinson's disease.

[4] The * notation refers to the inference for an unseen sample, using the optimised model.

4.2 Experiments and Results

The model proposed in this chapter is validated on the NPMC dataset. The features used to train the model are obtained using the subject-specific feature selection, detailed in section 3.3 in Chapter 3. Therefore the 15 hightest ranked features from the three MRI pulse-sequences are considered as input in the following experiments. Given the absence of a reliable state-of-the-art machine learning model used for CJD diagnosis, the performance of the proposed model is compared with a standard multikernel SVM. The experimental design implemented to test both models is detailed below.

4.2.1 Experiments

Proposed model

I evaluate the ability of the proposed model to correctly diagnose subjects for the two subtypes of CJD independently. The diagnosis of these subtypes is performed separately in order to avoid the counfounding effects related to the specific features of each subtype.

For IPD diagnosis, I include the information regarding the rate of progression of each mutation, as described by Equation 4.10; whereas for sCJD the kernel matrix $\mathbf{K}_c$ is filled with 1, since there is not an assumption of different rates of progression for sCJD. Further, to avoid missing information for healthy controls, these are randomly assigned the value 1 or 2 to encode a virtual rate of progression when compared with IPD subjects, assuming the value 1 for sCJD diagnosis. For both IPD and sCJD subtypes, only the subjects clinically labelled as symptomatic at the time of the first scan are considered (sample detailed in Table 3.1, Chapter 3). Exclusively the subjects with the three MRI sequences have been included in this experiment, due to the design of model $\mathcal{M}$, which requires the joint modelling of the three sets of features.

The model is trained using 75% of the overall sample, while keeping the input ratio between the different groups. The testing set corresponds to the remaining 25% of each sample. The proposed approach is coded in MATLAB^{TM}, using the optimisation routines available in GPstuff library[5]. In order to obtain a robust evaluation, a cross-validation scheme with 500 runs is used for all experiments.

[5] GPstuff: Bayesian Modeling with Gaussian Processes, available from http://jmlr.csail.mit.edu/papers/v14/vanhatalo13a.html [184]

Kernel SVM

Following previous studies using kernel SVM to diagnose patients with AD [73, 118, 119], I implement two alternative kernel SVM approaches to diagnose both IPD and sCJD patients independently:

1. **SE-SVM**: The term SE-SVM refers to a kernel SVM model for which a squared exponential function (SE) is used to estimate the kernel matrix. Three SE-SVM models are estimated for the individual source of features: T1w, FLAIR and DWI. The models are estimated using the kernel function defined in Equation 4.9. Both the distance to the class, but also the probability score associated with each prediction are estimated. The class probability for an unseen subject j corresponds to the average of the probability scores obtained from the model estimations using the different sources of features. The model is optimised based on the formulation described by Friedman *et al.* [174], implemented in MATLAB$^{\text{TM}}$ software. This analysis does not accounts for the impact of the rate of progression.

2. **Multi-Kernel Learning algorithm (MKL)**: An MKL model is used to learn: (1) the parameters of the kernel matrices and (2) their relevance for the estimation the of subjects' status [185]. The MKL model is estimated based on a weighted *L2*-norm regularisation with an additional constraint on the weights that encourages sparse kernel combinations. The final model corresponds to a linear combination of multiple kernels. Similarly to the models previously described, three squared-exponential kernels are used to encode the imaging features, as defined by Equation 4.9. The parameters of the model are optimised using an open-source toolbox, denominated SVM-KM [186]. Note that this formulation does not accounts for the estimation of the likelihood of the classes, hence no logaritmic loss has been estimated in this experiment.

These approaches follow the same training scheme described for the validation of the proposed model (section 4.2.1). The evaluation of the aforementioned methods is performed through the estimation of the sensitivity, specificity, accuracy and false rate of discovery (FDR) [187]. The receiver operating curves (ROC) and the area under the curve (AUC) are computed using the formulation for ROC graphs proposed by Fawcett *et al.* [188]. Further details about these metrics can be found in Appendix C.

4.2.2 Results

Figure 4.2 shows the predictive accuracy of the model $\mathcal{M}$ (Equation 4.2) when using imaging biomarkers extracted from the three MRI modalities. The ROC curves show that the model is more effective in the diagnosis of sCJD (AUC = 0.985 ± 0.06), when compared with the IPD classification (AUC = 0.937 ± 0.095), in both cases *versus* a healthy population. Inspection of Figure 4.2 also shows that both the kernel SVM approaches have a better performance in the diagnosis of sCJD subjects (AUC of 0.99 ± 0.04 and 0.93 ± 0.05 for SE-SVM and MKL approaches, respectively), compared with the proposed approach. However, these approaches are outperformed when used for the diagnosis of IPD, showing a AUC equal to 0.92±0.08 and 0.90±0.14 for SE-SVM and MKL approaches, respectively.

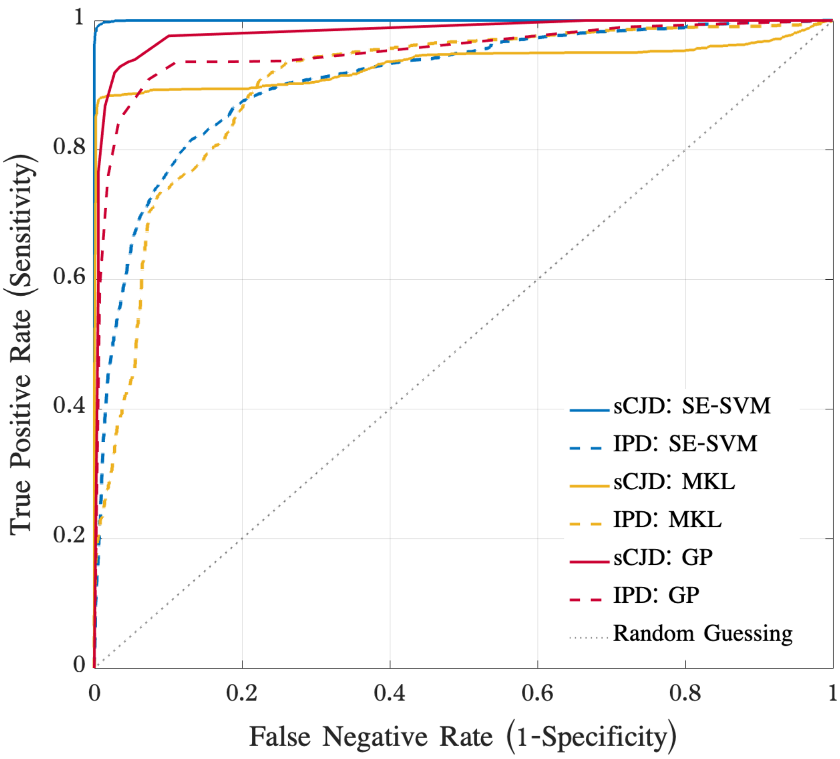


Figure 4.2: Predictive accuracy of the classification models for both IPD and sCJD subjects, when considering a dataset composed by the three MRI sequences (red curves). The predictive accuracy for both IPD and sCJD subjects, using squared exponential SE-SVM (blue curves) and MKL (yellow curves) approaches. The ROC curves are computed considering the predicted labels of 500 iterations, as proposed by Fawcett *et al.*, [188].

To investigate the influence of each feature to the subjects' diagnosis performances, it is also evaluated the accuracy of the predictive classes obtained using the latent models for the GP based model (section 4.2.2) and the individual SE-SVMs models (section 4.2.2). Considering the formulation of the MKL model, I do not investigate the performance of the model when using the different sources, but rather the weight associated with the kernel encoding the multiple sources of features.

Proposed model

Table 4.1 details the precision of the classification of sCJD patients using multi-source of features. The biomarkers extracted from FLAIR images seem to be insufficient to diagnose sCJD subjects during the onset of clinical symptoms. Conversely, the MD measures computed from DWI scans have the strongest influence in the diagnosis of sCJD, followed by the structural features. The results also suggest that the different sources of features show contradicting information, since including multiple sources has worsening the classification performance. In particular, the inclusion of features such as FLAIR in the model estimation when already including both DWI and T1w.

The predictive accuracy of the model for the diagnosis of IPD subjects is also evaluated, Table 4.2. Similarly to sCJD diagnosis, the intensity based features extracted from FLAIR promote a lower accuracy when used as single feature in the proposed approach. It can be observed that including the rate of progression associated with specific mutations yields an improvement of the predictive accuracy. Note also that the inclusion of the three source of features does not necessarily lead to the best performance for all metrics, which can be justified by the introduction of noise due to the features' interactions.

Finally, the distribution of the logarithmic loss[6], $\mathcal{L}$, across bootstrapping iterations, Figure 4.3, for classification of IPD symptomatic subjects shows a lower predictive power than the classification of sCJD patients. The lower $\mathcal{L}$ translates the higher certainty of the model in the classification of sCJD subjects; whereas, due to the less evident symptoms of IPD during the initial stages of the disease, the probability of the individual predictive class is lower, translating the uncertainty of the model in the diagnosis of this form of CJD.

[6] The logarithmic loss describes the uncertainty of the estimation associated to the predictive labels, as detailed in Appendix C, Equation C.5.

Table 4.1: Evaluation of the full model used for subjects diagnosis, for Sporadic CJD (sCJD). The mean value and standard deviation of 500 iterations is computed for all the metrics used for performance evaluation. The AUC is computed considering the results of all iterations. The false discovery rate (FDR) is also evaluated. All the evaluation measures are presented in percentage, excepting the AUC and the $\mathcal{L}$.

	Accuracy	Sensitivity	Specificity	FDR	AUC	$\mathcal{L}$
T1w	90.01 (10.34)	89.55 (16.60)	90.48 (16.76)	9.52 (0.17)	0.95 (0.10)	0.54 (0.18)
FLAIR	60.98 (15.48)	85.19 (22.29)	36.77 (29.25)	63.23 (0.29)	0.60 (0.25)	0.69 (0.01)
DWI	98.61 (4.62)	97.22 (9.23)	99.90 (< 0.10)	0.1 (< 0.01)	0.99 (0.01)	0.51 (0.18)
T1w + FLAIR	88.29 (13.33)	87.83 (19.33)	88.76 (20.00)	11.24 (0.03)	0.94 (0.13)	0.54 (0.19)
T1w + DWI	94.84 (8.79)	93.12 (14.47)	96.56 (10.59)	3.44 (0.05)	0.99 (0.03)	0.40 (0.22)
FLAIR + DWI	96.16 (10.13)	93.65 (16.41)	98.68 (10.22)	1.32 (0.02)	0.99 (0.06)	0.51 (0.18)
T1 + FLAIR + DWI	94.51 (9.96)	92.86 (15.52)	96.16 (12.21)	12.21 (0.06)	0.99 (0.06)	0.34 (0.15)

Table 4.2: Performance of the model for IPD diagnosis. The mean value and standard deviation of 500 iterations is computed for all the metrics used for performance evaluation. I included the impact of the rate of progression (RP) of the several mutations as a categorical variable in the model. In the full model, I modelled the join contribution of the DWI, FLAIR, T1w and the RP. Accuracy, sensitivity, specificity and false rate of discovery (FDR) are shown in percentage.

	Accuracy	Sensitivity	Specificity	FDR	AUC	$\mathcal{L}$
T1	93.70 (8.77)	93.00 (11.89)	94.40 (13.46)	5.40 (12.70)	0.95 (0.09)	0.35 (0.29)
FLAIR	56.36 (17.64)	80.73 (24.75)	31.98 (23.79)	67.82 (23.94)	0.53 (0.19)	0.70 (0.07)
DWI	77.63 (13.58)	80.40 (18.15)	74.84 (25.65)	24.93 (22.31)	0.69 (0.20)	0.69 (0.07)
T1 + FLAIR	93.48 (8.87)	93.00 (11.89)	93.95 (13.96)	5.85 (13.31)	0.95 (0.09)	0.35 (0.28)
T1 + DWI	92.53 (9.21)	93.00 (11.89)	92.05 (15.68)	7.75 (15.13)	0.94 (0.09)	0.36 (0.28)
FLAIR + DWI	70.85 (16.25)	74.22 (21.86)	67.48 (25.73)	32.32 (25.60)	0.67 (0.21)	0.69 (0.08)
T1 + FLAIR + RP	93.50 (8.87)	93.12 (11.74)	94.00 (13.93)	5.80 (13.28)	0.94 (0.09)	0.36 (0.28)
T1 + DWI + RP	93.05 (9.16)	92.85 (12.02)	93.10 (15.09)	6.70 (14.50)	0.94 (0.10)	0.37 (0.28)
FLAIR + DWI + RP	70.78 (15.74)	74.23 (20.92)	67.32 (26.16)	32.48 (26.02)	0.69 (0.21)	0.69 (0.09)
T1 + FLAIR + DWI	91.93 (9.07)	93.00 (11.89)	90.85 (15.90)	8.95 (15.37)	0.94 (0.09)	0.37 (0.28)
T1w + FLAIR + DWI + RP	92.45 (9.09)	93.14 (11.78)	91.90 (15.48)	7.90 (14.93)	0.94 (0.10)	0.37 (0.27)

[a] RP - Rate of Progression.

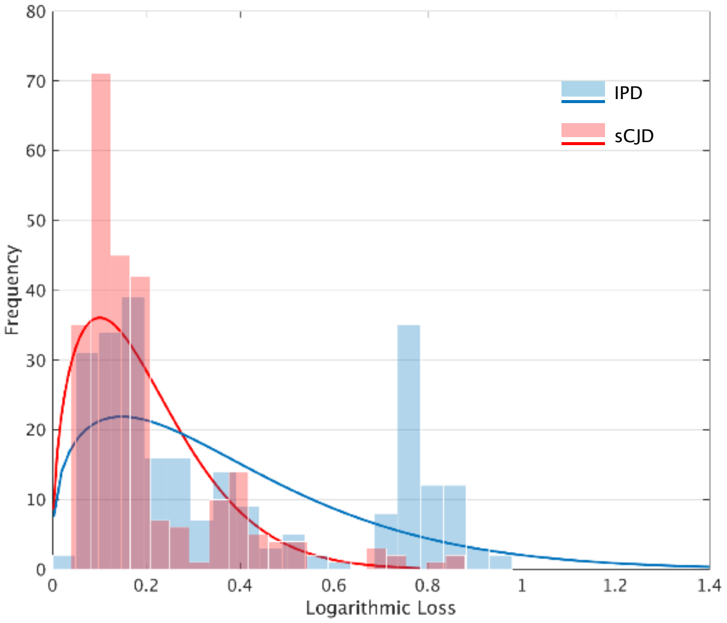


Figure 4.3: Distribution of the logarithmic loss $\mathcal{L}$ of the binary classification task, for both IPD and sCJD subjects. The histogram is computed across 500 iterations of the model. The fit is computed using a Weibull distribution.

Kernel SVM

The predictive accuracy of the SE-SVM model is evaluated for each set of biomarkers: T1w, FLAIR and DWI. Tables 4.3 and 4.4 show that the accuracy obtained from the binary classification using SE-SVM is comparable with the predictive accuracy of the proposed model, namely on the sCJD diagnosis. Furthermore, this result is not outperformed by the MKL approach, which achieved an accuracy of 93.00±0.05 for sCJD diagnosis. SE-SVM identifies the DWI features as the most relevant biomarkers to diagnose sCJD (Table 4.3); whilst, the features extracted from T1w had shown higher performance in the diagnosis of IPD subjects (Table 4.4). These results are in agreement with the results obtained with the GP-based model, where the latent models trained using the aforementioned input features show higher predictive performance.

Finally, Table 4.5 lists the mean value (and standard deviation) of the weights associated to the kernel matrices in the final model, over 500 iterations. The weight of the kernel matrices is intimately related to the relevance of these features to determine the subjects' status. The results show that the diagnosis of sCJD subjects is achieved mainly using DWI features, which is in agreement with the results obtained using both SE-SVM and the GP model. On the other hand, the diagnosis of IPD is based mainly on the T1w biomarkers, but both DWI and FLAIR equally contribute to the model final prediction, showing a weight of 0.21 and 0.23 in the final model.

4.3 Discussion

Proposed Model

The imaging biomarkers extracted, as detailed in Chapter 3, were used in a non-parametric Bayesian approach to predict the subjects status. The predictive labels are based on a probabilistic labelling of the subjects based on the joint modelling of the biomarkers pattern by a Gaussian Process. Both sCJD and IPD were independently diagnosed by evaluating the predictive accuracy of the labels for both subtypes. The reported results, reported in section 4.2.2, are indicative of the effectiveness of the model to detect prion disease patients, among healthy controls. Furthermore, the model was also able to diagnose subjects in the early stages of CJD, particularly for IPD symptomatic subjects with MRC scale of 20, a time at which the diagnosis can be otherwise very challenging. The results also suggest that the diagnosis of CJD can be achieved without all three MRI sequences, Table 4.1. From the results obtained using the proposed model, I concluded that the DWI scans are more informative to diagnose sCJD. Thus, the diagnosis of sCJD benefits from the use of a single feature, in specific the MD measurements. Nevertheless, the logarithmic loss $\mathcal{L}$ shows that the full model was more robust in the diagnosis of sCJD, for which the uncertainty regarding the predictive label was lower.

Differently, the diagnosis of IPD subjects show a higher accuracy when using only biomarkers extracted from T1w images, Table 4.2. Despite the fact that the diagnose of IPD benefited mostly of T1w MRI images, by including other modalities such as DWI the sensitivity of the predictive labels increases. Note that the jointly modelling of T1w with either with FLAIR or DWI and the rate of progression is equally sensitive in the IPD diagnosis, when compared with the joint modelling of the three MRI images. In both scenarios, including the three MRI sequences, did not show significantly better results when compared with the aforementioned latent models. In addition, the inclusion of the kernel $\mathbf{K}_c$ had proved the flexibility of the proposed model to deal with both categorical and continuous data, such as genetic and quantitative imaging data respectively. Future work should make use of this kernel matrix to encode any other relevant genetic data, such as SNP information.

This model could be an improvement in clinical environment, since it provides

Table 4.3: Evaluation of the full model used for subjects diagnosis, for sporadic CJD (sCJD) using a Squared Exponential SVM. The mean value and standard deviation of 500 iterations is computed for all the metrics used for performance evaluation. The AUC is computed considering the results of all iterations. All the evaluation measures are presented in percentage, excepting the AUC and the $\mathcal{L}$.

		Accuracy	**Sensitivity**	**Specificity**	FDR	AUC	$\mathcal{L}$
SE-SVM	T1w	93.63 (8.01)	93.20 (11.25)	94.05 (11.56)	5.95 (11.56)	0.94 (0.08)	1.96 (0.87)
	FLAIR	66.47 (15.21)	67.90 (25.61)	65.05 (24.36)	34.95 (24.36)	0.67 (15.21)	0.80 (0.25)
	DWI	99.10 (3.33)	98.20 (6.66)	99.90 (> 0.10)	0.1 (< 0.01)	0.99 (0.03)	0.51 (0.18)
	T1w + FLAIR + DWI	98.82 (3.90)	97.75 (7.50)	99.90 (1.58)	1.0E-3 (1.58)	0.99 (0.04)	1.13 (< 0.01)
MKL	T1w + FLAIR + DWI	94.54 (3.14)	93.80 (2.43)	94.90 (2.64)	1.0E-3 (2.24)	0.93 (0.05)	—

Table 4.4: Evaluation of the full model used for subjects diagnosis, for inherited prion disease (IPD) using a Squared Exponential SVM. The mean value and standard deviation of 500 iterations is computed for all the metrics used for performance evaluation. The AUC is computed considering the results of all iterations. All the evaluation measures are presented in percentage, excepting the AUC and the $\mathcal{L}$.

		Accuracy	**Sensitivity**	**Specificity**	FDR	AUC	$\mathcal{L}$
SE - SVM	T1w	91.82 (8.25)	87.97 (14.77)	96.79 (7.74)	3.21 (7.74)	0.92 (0.08)	2.94 (1.45)
	FLAIR	58.82 (13.42)	69.48 (24.98)	48.51 (24.75)	51.49 (24.75)	0.59 (0.14)	0.75 (0.14)
	DWI	86.36 (9.54)	73.77 (17.80)	90.90 (> 0.10)	0.1 (< 0.01)	0.87 (0.09)	0.51 (0.18)
	T1w + FLAIR + DWI	91.73 (8.89)	84.68 (16.37)	90.92 (1.26)	8.0E-4 (1.24)	0.92 (0.08)	0.78 (< 0.01)
MKL	T1w + FLAIR + DWI	90.80 (9.65)	89.34 (11.04)	91.15 (10.23)	13.60 (16.22)	0.90 (0.14)	—

Table 4.5: Kernel matrices contributions of the MKL approach for subjects' diagnosis. The relevance of the different source of features is ranged from 0 (not relevant) to 1 (only relevant feature). The grey colour highlights the most relevant set of features.

	T1w	FLAIR	DWI
IPD	0.560 (0.134)	0.206 (0.112)	0.234 (0.114)
sCJD	0.202 (0.103)	0.016 (0.040)	0.782 (0.110)

relevant information regarding the biomarkers that are more useful for the earlier diagnosis of CJD, avoiding unnecessary exams or a prioritisation of the clinical assessment in more severe stages of the disease.

Notwithstanding the good results, the current formulation of the model does not account the progression of the disease. Further, the proposed model only takes in consideration cross-sectional data. Due to the reduced number of subjects and high heterogeneity of the symptoms across patients, there is not an established function to describe the progression of symptoms over time, hampering the use of longitudinal data. This consists of a limitation of the formulation of the proposed approach. Hence longitudinal information could improve the diagnosis of the subjects. Currently, this does not hamper the performance of the model on the classification of CJD patients. Nonetheless, to identify signs of the onset of symptoms, promoted by brain alterations occurring between two clinical assessments, the model should take into consideration longitudinal information.

Kernel SVM

To assess the validity and accuracy of the proposed model, I compared the GP-based model with two kernel based approaches: MKL and SE-SVM. The results, Tables 4.3 and 4.4, show that the kernels based models were accurately predicting CJD. In fact, both SE-SVM and MKL approaches outperformed the GP model when diagnosing sCJD subjects. Nevertheless, the GP showed a significantly lower logarithmic loss in both tasks, which is translated in a lower uncertainty of the predictions given by this approach. Therefore, even with a higher accuracy, the SE-SVM is not suitable to be used in clinical context given the uncertainty of the predicted classes.

For the diagnosis of IPD subjects, the GP model outperformed the kernel-based approaches, suggesting that this approach is more sensitive to identify earlier stages of the diseases, where the clinical manifestations are less evident and noisy.

Note also that the MKL approach identifies the features extracted from FLAIR as relevant as the features extracted from DWI. These findings support the results obtained using the GP model, where the combination of T1w with either DWI or FLAIR features leads to an increase of the model sensitivity; whereas the SE-SVM completely discards the relevance of the features extracted from FLAIR. Despite the results obtained using SE-SVM and GP model that indicated a very low relevance of the FLAIR features, the MKL showed that the biomarkers extracted from FLAIR conditioned the optimisation of the model parameters with equal relevance than the DWI features. Therefore, the FLAIR should not be disregard as a source of information to characterise CJD.

The good results obtained by the kernel methods sustained the hypothesis that subject-specific features are suitable to diagnose CJD, since these features even when used in a different classifier lead to a good identification of prion disease.

4.4 Summary

This chapter introduced a non-parametric Bayesian approach to predict the subjects status. I evaluated the effectiveness of the proposed method in a cohort of patients with inherited and sporadic forms of prion disease. The model had shown to be effective in the prediction of both inherited CJD (93.7% of accuracy) and sporadic CJD (98.5% of accuracy). Compared with state-of-the-art approaches, the framework achieved comparable results, outperforming the state-of-the-art approaches when diagnosing IPD subjects.

This model is particularly useful if implemented in clinical context as a computer-aided-diagnosis tool, potentially reducing the current misdiagnosis rate of prion diseases. In fact, this is one of the objectives of applying machine learning models to study prion diseases. By using these type of diagnostic models, the clinicians could be alerted to the potential presence of CJD, even when the clinical manifestations resemble more common forms of dementia. Therefore, this chapter introduced a clinically relevant algorithm to address the current clinical challenges of CJD.

However, the model did not provide yet an effective characterisation of the different stages of the disease, neither a prediction of clinical onset for IPD. Further, the proposed method only took in consideration cross-sectional data. Due to the reduced number of subjects and high heterogeneity of the symptoms across patients, there is not an established function to describe the progression of the disease over time.

To improve the knowledge about the evolution of the disease over time, in the Chapter 5, I extend the current framework to perform subject's staging according to the MRC Scale and by consequence the severity of brain changes. The new model can be seen as a disease progression model, defined as an additive multi-class GP.

Chapter 5

Multi-class Gaussian Process for CJD characterisation

To date, there is no accurate measure that can be used to quantify the evolution of symptoms over time, as a proxy of subjects' prognosis, or to anticipate the clinical onset of asymptomatic subjects. Being able to diagnose CJD at the early stages of the disease would enable the patients to be involved in clinical trials, which is currently challenging as patients can die in less that 12 months from diagnosis [18]. Therefore, the prediction of the time to clinical onset of IPD patients is one of the aims of this study. To predict the subjects prognosis, I extended the model introduced in Chapter 4 to perform a multi-class classification aiming the subjects staging according to the clinical symptoms. Moreover, to address the current misclassification rate of CJD, I also applied the multi-class GP framework to identify CJD among other neurodegenerative diseases.

This chapter describes a multi-class Gaussian Process Classification (GPC) used aiming both the CJD stratification and its identification among other neurodegenerative diseases. The multi-class GP framework is detailed in section 5.2, followed by its evaluation when used on clinical data in section 5.3. By using a common model, which includes the same kernel function and optimisation scheme, I demonstrate the potential of model $\mathcal{M}$ to be extended to work both as a prognosis and differential diagnosis tool, as well as its current limitations if translated to clinical context (sections 5.4 and 5.4, respectively).

5.1 Context

Subjects' Stratification

The models used to predict subject's prognosis in context of neurodegenerative diseases require the age/time normalisation among the subjects. This step is crucial for these approaches, since only with a proper normalisation it will be possible to analyse the subjects at different stages and time-points jointly [149, 150, 159]. The subjects' age normalisation is a challenging step for CJD patients, due to the wide range of ages at the clinical onset, as described in Chapter 3. Attending to the data limitations, namely the reduced number of subjects with data before and after clinical onset, the estimation of the exact time to clinical onset became an ill-posed problem when formulated as a regression task. To tackle these issues, I define the subjects' prognosis as a multi-class classification task, where the subjects' status is a class in an ordinal scale based on the severity of the symptoms. This formulation is not a continuous measure of time to onset, in years; hence, it does not answer to the problem introduced in this chapter. Nevertheless, the stratification of the subjects according to the severity of symptoms, or the proximity to clinical onset stage, can be interpreted as a surrogate measure of the subject's outcome. Note also that the probabilistic outcome also give information regarding the transition between stages; i.e., admitting that the disease stages follow an ordinal distribution, it is sensible to assume that a subject i at the time-point t will progress to the closest upper stage as $y_{i,t+1} = C + 1$.

Differential Diagnosis

Due to its rarity, CJD is, in fact, commonly mistaken for other neurodegenerative disease, which results in a higher rate of undiagnosed subjects. As a consequence, these patients are not conveniently treated according to their symptoms. Besides, by reducing the misdiagnosed cases, the sample size of CJD studies will increase, leading to a more robust analysis and to a better understanding of the disease. Taking advantage of the flexibility of the GP formulations, I here adapt the model $\mathcal{M}$ to be used as a differential diagnosis tool, in particular to identify CJD among other symptoms of dementia, caused by diseases such as YOAD.

5.2 Model definition

The generative model, Equation 4.1, can be adapted to predict the stage of the disease for a subject i given the set of features $\mathbf{X} \in \mathcal{X}$. The estimated probabilistic class provides a clinical input regarding the severity of symptoms of CJD. I implement a multi-class classification GP based on individualised likelihood factors computed for the target classes defined by $y_i = \{C_1, \dots, C_{\mathcal{C}}\}, \mathcal{C} > 2$ for the subject i. The estimation of the class probability is given by a multinomial *probit* likelihood[1], which can be generalised to account for non-constant error variances. For this purpose the Equation 4.2 has been modified as following:

$$p(y_i|\mathbf{f}_i) = \mathrm{E}_{p(u_i)}\left\{\prod_{j=1,j\neq y_i}^{\mathcal{C}} \mathbf{\Phi}(u_i + f_i^{y_i} - f_i^j)\right\} \tag{5.1}$$

where f_i is a vector $\mathbf{f}_i = [f_i^1, ..., f_i^{\mathcal{C}}]^T$ to account for the number of classes under consideration for a subject i. In the Equation 5.1, the auxiliary variable u_i is distributed as $p(u_i) = \mathcal{N}(u_i|0,1)$.

5.2.1 Marginal Likelihood approximation

Similarly to binary classification, the posterior of a multi-class GPC is analytical intractable; thus, it requires the approximation of the likelihood. To keep the consistency across classification tasks, this approximation is achieved by means of EP algorithm. The EP approximation is in fact particularly efficient for the multi-task likelihood problems. Note however that in case of binary GPC, the estimation of the tilde distributions (defined in Equation 4.4) requires solving one-dimensional integrals. Assuming the *probit* likelihood function, these univariate integrals can be computed efficiently without numerical quadratures [112, 125]. For the multi-class paradigm the solution is more complex, since it is required to evaluate the multi-dimensional integrals [127]. For this, the approximation of the tilde variables can be done by Laplace approximation [112]. The problem with the Laplace approximation approach is that the mean is replaced with the mode of the distribution and the covariance with the inverse Hessian of the logarithmic density at the mode. Because of the skewness of the tilde distribution caused by the likelihood function, the Laplace approximation method can lead to inaccurate mean and covariance estimates in

[1] The likelihood used here could be potentially be replaced by an ordinal likelihood, as proposed by Chu *et al.*, [189]. However, for simplicity, the multinomial likelihood is used here, which still ensures the multi-class nature of the problem even if it disregards the ordinal behaviour of the categories where lie the disease's stages.

which case the resulting posterior approximation does not correspond to the full EP solution. To overcome this issue, following the method proposed by Riihimäki *et al.*, the marginalisation of the likelihood is achieved via nested EP approximation, which does not require numerical quadratures or sampling to estimate the predictive probabilities [127, 190].

5.2.2 Kernel function definition

In this particular case of the model $\mathcal{M}$, the imaging biomarkers are encoded in the model using a linear combination of logistic kernel functions (Equation 5.2). Here, the logistic function $h(x)$ is fully defined by the w scalar that is the weight of the mean function, by the intercept of the linear part b and the regression coefficient of the linear part a. Given a zero mean function, as assumed in this model, on the Gaussian prior for weight $w \sim N(0, \sigma^2)$ the prior for $h(x)$ is $h(x) \sim N(0, \mathbf{H}(x)\mathbf{H}(x)^T \boldsymbol{\sigma}^2)$ where $\mathbf{H}(x) = [h(x_1), ..., h(x_n)]^{\mathrm{T}}$. Finally, for N input samples, the diagonal matrix $\mathbf{\Sigma}$ contains the prior variances of the b_N terms. This function is used to encode the variance of the biomarkers over the different stages of the disease.

$$\begin{aligned} h(x) &= w(\text{logit}^{-1}(ax + b) - 0.5) \\ \mathbf{H}(x) = [h(x_1), ..., h(x_n)]^{\mathrm{T}}, \mathbf{\Sigma} &= \text{diag}(\sigma_1^2, ..., \sigma_N^2) \\ k_{logit}(x, x' | \mathbf{\Theta}) &= \mathbf{H}(x)\mathbf{\Sigma}\mathbf{H}(x')^{\mathrm{T}} \end{aligned} \tag{5.2}$$

The particular use of Equation 5.2 is motivated by the results achieved by the automatic kernel selection scheme, proposed by Duvenaud *et al.* [113]. This algorithm was modified to be used on the specific scenario of the study of neurodegenerative diseases, as detailed in Appendix B. Note that the automatic selection of the kernel function is not performed specifically for CJD, due to the sample size. In fact, using the same sample for both kernel selection and the optimisation of the model parameters for subjects' staging would lead to double-dipping, hampering the robustness and generalisation of the results. Hence, an independent sample, composed of 320 patients diagnosed with inherited Alzheimer's disease, is used to select the appropriate kernel to encode both clinical and imaging data. The data is part of the Dominantly Inherited Alzheimer Network (DIAN) study [191].

5.3 Experiments and Results

5.3.1 Subjects Stratification

Experiments

Using cross-sectional data, the model is trained to perform subjects' stratification on the HC, asymptomatic subjects, IPD and sCJD. The target classes defined by $y_i = \{\mathcal{C}_1, \ldots, \mathcal{C}_C\}, \mathcal{C} \in \{1, \ldots 5\}$ for the subject i correspond to the five predefined stages of the disease: (1) healthy control (HC), (2) asymptomatic subjects (Asym.), (3) subjects with an MRC Scale score of 20 i.e., asymptomatic[2] or early symptomatic but no accrued neurodisability, within one year inside of the clinical onset window (CO), (4) to (5) symptomatic subjects divided in 2 severity quantiles according to their MRC Scale scores [59]. In detail, the (4) stage includes the subjects with MRC Scale score between 19 and 15, and (5) comprises the subjects with MRC Scale score below 14. Only 2 subjects had an MRC Scale score below 10, and those were included in the stage 5. In this specific experiment the sCJD and IPD subjects are jointly classified, according to their MRC Scale score. Similarly to experiments described in Chapter 4 (section 4.2.1), to model the different rates of progression of the subtypes of CJD and IPD mutations, the $\boldsymbol{K}_c$ is considered in to the model. It is assumed that sCJD patients show a disease progression rate analogous to the IPD subjects with the fastest progression rate. Both the SE-SVM model, defined in Chapter 4 (section 4.2.2), and the proposed approach for subjects staging are evaluated using the same sample. The models are trained using 75% of the overall sample, whereas the testing set corresponds to the remaining 25% of each sample, while keeping the input ratio between the different groups. The hyperparameters of the model are optimised using the previously mentioned open-source library[3]. In order to obtain a robust evaluation, I apply a cross-validation scheme with 500 runs for all experiments. Lastly, I evaluate the performance of the model when used as a prognostic tool. For that purpose, I consider the available longitudinal data for the CO subjects, where I tracked the biomarkers evolution over time and the conversion to IPD by predicting the class label independently for each time-point. Even though the current formulation does not provide information about the time to clinical onset of each subject, the probability distribution for each subject with

[2] The asymptomatic subjects considered as part of clinical onset (CO) had have the diagnosis confirmed in later MRI images.

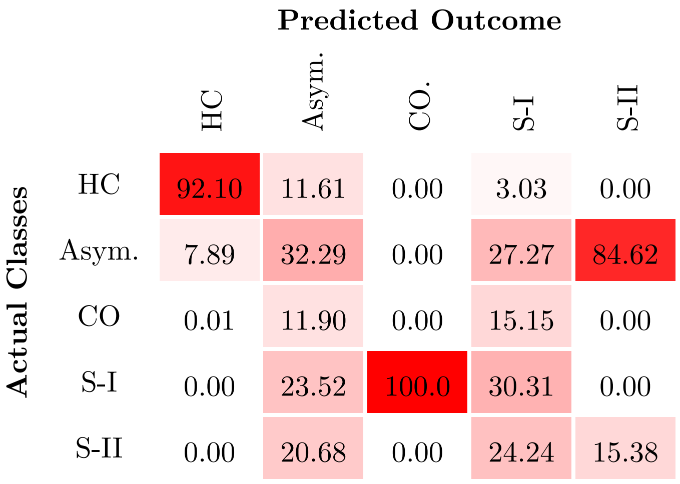
[3] GPstuff: Bayesian Modelling with Gaussian Processes, available from http://jmlr.csail.mit.edu/papers/v14/vanhatalo13a.html [184]

respect to the six classes provides information about the severity of the symptoms and consequently the stage of the disease. All the aforementioned experiments are assessed based on the metrics defined in Appendix C.

Results

Figure 5.1 shows the normalised confusion matrix for the testing set, when using the SE-SVM to perform the subjects stratification. The results suggest that the model is able to effectively differentiate the healthy controls from patients' showing signs of CJD. However, this approach is inadequate to characterise the several stages of the symptoms severity over the course of the disease. In fact, the SE-SVM model is strongly impacted by the presence of asymptomatic patients during the training stage, resulting in a high percentage of false positive diagnosis for this group. Lastly, the SE-SVM approach failed in the identification of any of the subjects at the clinical onset.

Figure 5.1: Subjects Stratification via SE-SVM. The discrete confusion matrix was computed based on the sum of 500 iterations of the model. The dark red highlights the higher percentage of subjects classified with a given label across iterations. HC – healthy controls; Asym. – asymptomatic subjects; CO – clinical onset; SI – stage I and SI - stage II of the disease.



Predicted Outcome

Actual Classes	HC	Asym.	CO.	S-I	S-II
HC	92.10	11.61	0.00	3.03	0.00
Asym.	7.89	32.29	0.00	27.27	84.62
CO	0.01	11.90	0.00	15.15	0.00
S-I	0.00	23.52	100.0	30.31	0.00
S-II	0.00	20.68	0.00	24.24	15.38

Figure 5.2 shows the normalised confusion matrix for the testing set. The qualitative analysis of the confusion matrix suggests that the model is able to correctly identify the extreme stages of the disease, while being less accurate in the differentiation of the intermediate stages of the disease. Note that the results reported in Figure 5.2 are deterministic and they do not account for the fuzziness of the classes estimated, particularly for the asymptomatic stage.

Figure 5.2: Subjects stratification via multi-class GP. The discrete confusion matrix was computed based on the mean of 500 iterations of the model. The values correspond to the mean percentage of subjects labelled as belonging to a given class. The intensity of the color increases with percentages. HC – healthy controls; Asym. – asymptomatic subjects; CO – clinical onset; SI – stage I and SI - stage II of the disease.

		Predicted Outcome				
		HC	Asym.	CO.	S-I	S-II
Actual Classes	HC	90.68	7.43	0.09	6.27	0.86
	Asym.	4.33	76.38	17.54	14.46	1.05
	CO	0.03	6.60	22.39	24.86	0.01
	S-I	3.76	8.71	50.83	36.93	29.35
	S-II	1.20	0.88	9.15	17.48	68.74

Figure 5.3 shows the correlation between the categorical (discrete) labels and the average probability given to each class, computed through bootstrapping. The probability distribution across classes gives a more intuitive interpretation of subjects' clinical status, as well as the confidence of the predictions for each group.

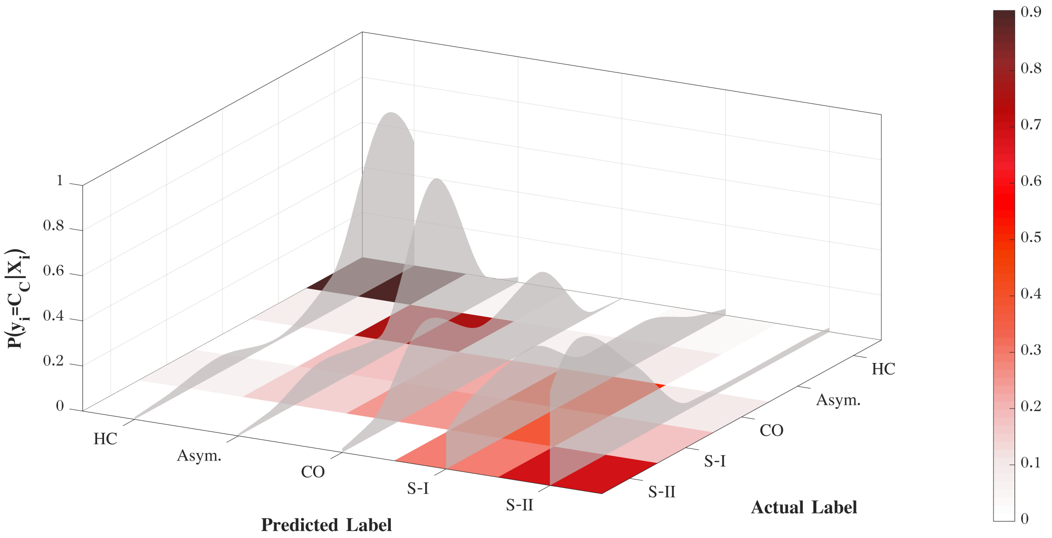


Figure 5.3: Prion disease subjects stratification using the proposed framework.. The discrete confusion matrix is normalised by the number of subjects included in the classification task. The shadow area is the average distribution of probabilities per class.

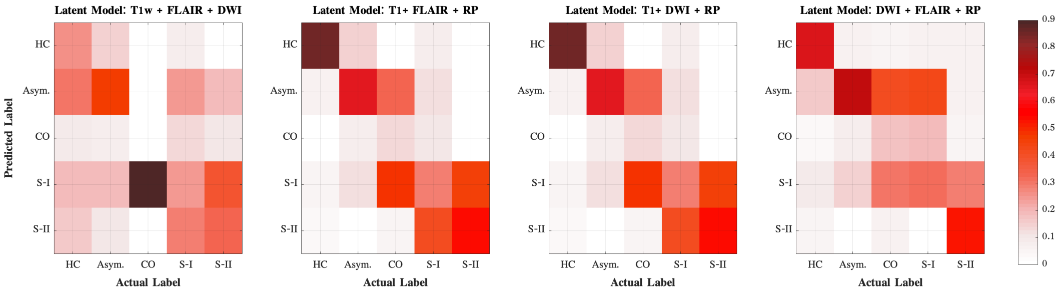


Figure 5.4: Prion disease subjects stratification, latent models. The confusion matrices show the results of the latent models for the stratification task.

I further investigate what is the best combination of features to achieve a good stratification of subjects, by computing the predictive classes for the latent models. Table 5.1 presents these findings, where the average accuracy across stages is higher for the jointly modelling of the three set of features (Average Acc = 83.2%) combined with the progression rate of the individual mutation. Nonetheless, both macro-recall and macro-precision reported are low ($Recall_M$ = 40% and $Precision_M$ = 39%), suggesting that the model is not sensitive to detect the different stages of the disease.

Figure 5.4 also show the inefficiency of certain latent models in the stratification of patients showing signs of CJD. It is highlighted the importance of different MRI sequence as a source of information to characterise different stages of the disease progression. As an example, T1w seems to have a predominant role in the identification of the early stages of the disease (second and third panels), where its influence reduce the false positives – i.e., asymptomatic subjects being classified as CO; whilst, the biomarkers extracted from DWI images are more relevant to identify severe symptoms (fourth panel), characteristic of the latest stages of the disease progression. The logarithmic loss, computed across classes, is also lower for the full model ($\mathcal{L} = 1.74 \pm 0.44$), supporting the assumption that by using the three MRI sequences, the CJD symptoms are better explained resulting in a more accurate subject's prognosis[4].

I apply the trained model to study the evolution of imaging biomarkers over time, for each subject considered at clinical onset. This experiment aims to demonstrate the possible use of the model as prognostic tool, based on the class probabilities over time. The reported results, Figure 5.5, are evaluated for each of subject

[4] The low logarithmic loss translates the higher certainty of the classes correctly label, whilst the classes wrongly predicted have often higher uncertainty related to them.

Table 5.1: Performance of the model when used for disease staging. The mean value and standard deviation of 500 runs is computed for all the metrics used for multi-class evaluation. The values are presented in percentage, excepting the logarithmic Loss.

	Average Acc	$Precision_M$	$Recall_M$	$\mathcal{L}$
T1w	69.43 (2.16)	4.80 (1.50)	19.96 (0.80)	3.14 (1.11)
FLAIR	69.17 (1.68)	5.24 (3.71)	20.08 (2.14)	2.60 (0.13)
DWI	69.17 (1.68)	5.24 (3.71)	20.08 (2.14)	2.60 (0.13)
T1w + FLAIR	70.26 (2.03)	6.83 (5.28)	20.67 (4.07)	2.66 (0.23)
T1w + DWI	70.26 (2.03)	6.83 (5.28)	20.67 (4.07)	2.66 (0.23)
FLAIR + DWI	69.09 (1.78)	5.09 (3.36)	19.94 (2.35)	3.84 (1.66)
T1w + FLAIR +RP	78.74 (5.11)	35.18 (13.15)	39.40 (10.45)	2.05 (0.16)
T1w + DWI + RP	78.74 (5.11)	35.18 (13.15)	39.40 (10.45)	2.05 (0.16)
FLAIR + DWI + RP	77.62 (5.60)	35.31 (17.37)	42.85 (13.26)	2.47 (0.68)
T1w + FLAIR + DWI	69.88 (1.96)	6.48 (5.81)	20.97 (4.21)	6.03 (2.01)
T1w + FLAIR + DWI + RP	85.39 (4.21)	57.34 (11.91)	58.83 (11.93)	1.74 (0.44)

RP – Rate of Progression.

at clinical onset independently. Note that in some cases the evaluation before clinical onset is omitted due to missing data (subjects B and C). The results show the inability of the model to predict the actual stages of the disease for these subjects. However, the changes of the likelihood of the predicted classes can be interpreted as a possible change of the confidence of the model to predict a class given the set of features, which might show signs of disease progression to more severe stages. As an example, Subject A initially is classify as asymptomatic with a high confidence ($p(y_A|\boldsymbol{f}_A) \approx 0.89$); whereas for the second time-point this probability decreases ($p(y_A|\boldsymbol{f}_A) \approx 0.80$), suggesting that the prediction is less certain for the second time-point. This pattern is also visible for Subject C. However, the results for Subjects B and D show a less evident correlation between the decrease of the likelihood of the predicted class in the two different time-points and the conversion from asymptomatic to CO, or to more severe stages of the disease.

5.3.2 Differential Diagnosis

Experiments

Thanks to the flexibility of the proposed approach, I also implement a multi-class classification GP to perform a differential diagnosis of CJD. This formulation aims to demonstrate the possibility of using this approach to either diagnosis the subtypes of CJD or to identify CJD among other neurodegenerative diseases.

Therefore, I compare the CJD subtypes against a clinically related form of

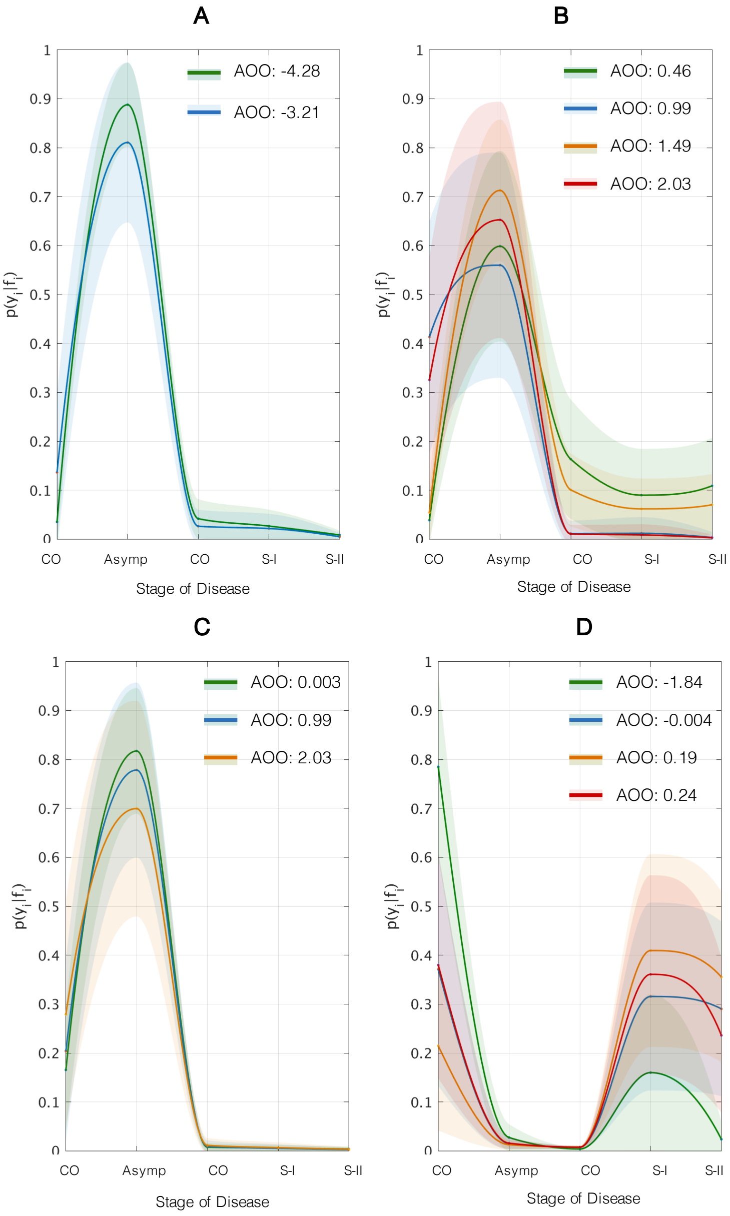


Figure 5.5: Classification of subjects at clinical onset. Longitudinal stratification of subjects with MRI scans acquired before and after the clinical onset. The likelihood of the predicted class are represented by the mean function and variance for each time-point. The age of onset (AOO) is used as reference of the scanning time-points, before (negative values) and after clinical onset (positive values). Subject A: MRC scale 21 (blue), 21 (green); Subject B: MRC scale 19 (blue), 12 (green), 17 (gold), 15 (red); Subject C: MRC scale 20 (blue), 20 (green), 18 (gold); Subject D: MRC scale 21 (blue), 20 (green), 19 (gold), 10 (red).

dementia. For this experiment, I consider the HC, IPD, sCJD and YOAD groups. The asymptomatic subjects have been excluded to avoid the presence of confounding effects during the training of the model[5]. The YOAD dataset comprises 32 subjects (10 males) with a mean age of 61 years old, as detailed in Chapter 3 (Table 3.1). The YOAD subjects are part of a larger study of young onset Alzheimer's disease, for which ethical approval was obtained from the National Hospital for Neurology and Neurosurgery Research Ethics Committee. T1w and DWI images were acquired using the same scanner than the NPMC subjects, which ensures the compatibility of the images acquired. In detail, the T1w pulse sequence is identical to the one used for NPMC. The DWI acquisition however differ for this dataset. Multiple shells were acquired for the YOAD dataset. For better harmonisation, this work only used the shell that had the most similar b-value (b=700) to the one used for the prion data acquisition (b=1000). I acknowledge this limitation as a potential bias in the results of the classification. The FLAIR images were not acquired for this group of patients. Hence, the features used to characterise YOAD subjects are obtained only from DWI and T1w MRI scans, which have been processed using the framework detailed in Chapter 3 (sections 3.2 and 3.3). To keep the agreement across datasets, only DWI and T1w imaging features are considered to characterise CJD patients. Note that the feature selection section of this framework is tailored to maximise the information related to CJD symptoms; thus, the features do not encode the spatial pattern that characterises YOAD.

By following the model formulation described in section 5.1, I compute individualised likelihood factors for the target classes defined by $y_i = \{\mathcal{C}_1, \dots, \mathcal{C}_C\}, \mathcal{C} \in \{1, \dots 4\}$ for the subject i, where (1) corresponds to healthy controls, (2) IPD, (3) sCJD and (4) YOAD. Lastly, the Equation 5.2 is used to encode the imaging biomarkers. Once again, the effectiveness of the proposed approach is evaluated based on the metrics described at Appendix C.

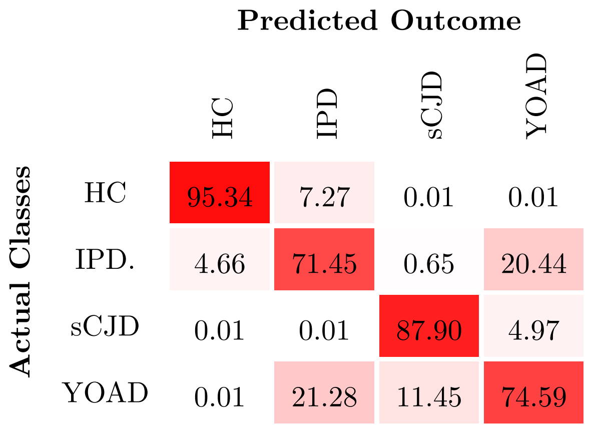
Results

By modelling the joint contribution of the three sets of features, it is achieved a good differentiation between the symptomatic CJD subjects and the YOAD patients, as reported in Figure 5.6. The confusion matrix shows that only approximately 8% of CJD subjects are labelled as HC. This result is in agreement with the results reported in Chapter 4 (section 4.2.2), which suggests that including more signs does not perturb the sensitivity of the model in identifying CJD. Most of the CJD patients

[5] Asymptomatic subjects form indeed a heterogeneous group as individuals can be days or decades from clinical onset.

misclassified as HC are IPD patients. I presume that the slower rate of progression of IPD patients and the higher number of subjects with MRC scale of 20 lead to less evident symptoms and consequently proximity to the HC biomarkers pattern.

Figure 5.6: Differential diagnosis of CJD subtypes. The confusion matrix shows the mean percentage of predictive labels across the 500 runs of the model. The higher percentages of subjects classified with a given label across iterations are shown with an intense colour. HC – healthy controls; IPD – inherited prion disease; sCJD – sporadic CJD; YOAD – young onset Alzheimer's disease.


Predicted Outcome
Actual Classes

	HC	IPD	sCJD	YOAD
HC	95.34	7.27	0.01	0.01
IPD.	4.66	71.45	0.65	20.44
sCJD	0.01	0.01	87.90	4.97
YOAD	0.01	21.28	11.45	74.59

The results also indicate that 75% of the YOAD subjects have been correctly labelled, showing a likelihood of 0.61 of being YOAD (Figure 5.7). The CJD subjects, both IPD and sCJD, have shown a probability of 0.35 of being wrongly labelled as YOAD subjects, with the majority of the uncertainty associated with IPD misclassification (likelihood of being YOAD, for IPD patients, ≈ 0.40). The overlap between these two classes is due to the similarities of the phenotype of the diseases, which differ from sCJD. These results can be improved by a specific kernel matrix to explain the spatial differences between the two diseases. Nonetheless, this section is an illustrative example of the flexibility of the proposed model and the possibility of being used as a differential diagnosis tool, particularly to identify CJD among other types of dementia.

The analyses of the latent models performance show that the categorical kernel used to encode the progression of IPD mutations, highly improves the accuracy and precision of the differential diagnosis tool. Figure 5.8 highlights the DWI features as the most sensitive to differentiate IPD from YOAD, whereas T1w images combined with DWI show better results in the identification of sCJD among YOAD patients. Finally, Table 5.2 summarises the performance of the latent models in the prediction of the subjects' status. The approach that includes the features extracted from the

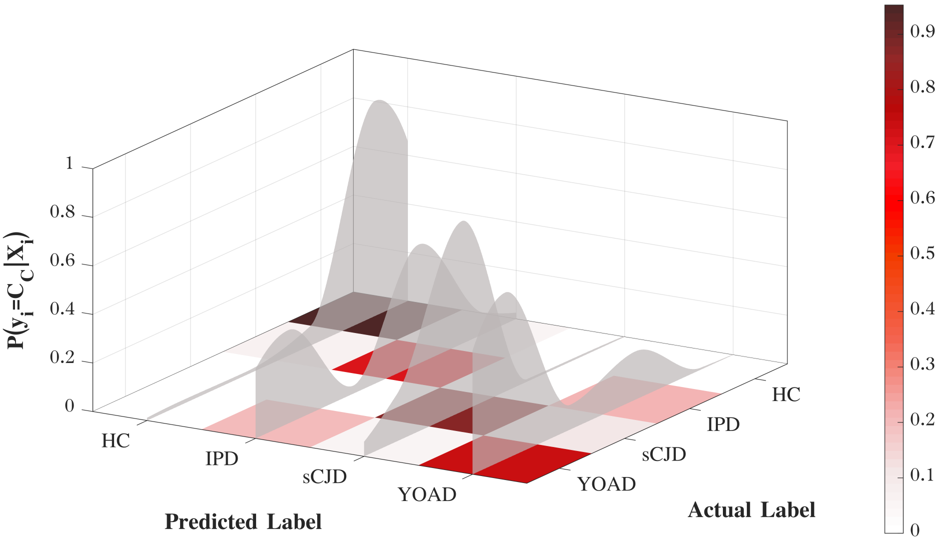


Figure 5.7: Likelihood of the predictive classes of differential diagnosis. The shadow area is the average distribution of probabilities per class.

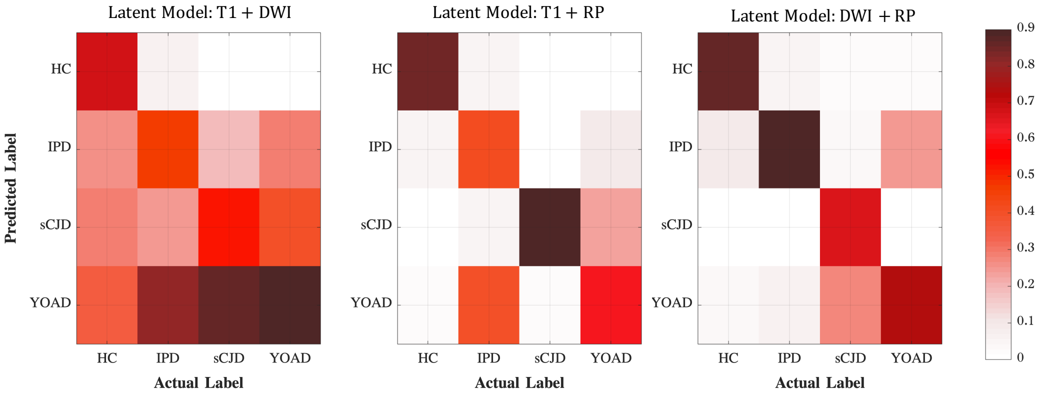


Figure 5.8: The confusion matrices show the results of the latent models for the differential diagnosis task.

two MRI pulse-sequences, as well as the rate of progression, outperforms the latent models for all the metric analysed.

5.4 Discussion

Subjects Stratification

Notwithstanding the promising results obtained for subjects' diagnosis, there is not an effective characterisation of the different stages of the disease, neither the prediction of clinical onset for IPD. To improve the knowledge regarding the evolution of symptoms over time, I extended the initial model (Chapter 4) to perform the subjects staging according to the MRC Scale. The extended model $\mathcal{M}$ can be seen as a stratification tool working as disease progression model, defined by an

Table 5.2: Performance of the model for the differential diagnosis. The mean value and standard deviation over 500 runs is computed for all the metrics used for performance evaluation. The average accuracy, macro precision and macro recall are shown in percentage. RP refers to the rate of progression.

	Average Acc	$Precision_M$	$Recall_M$	$\mathcal{L}$
T1w	60.99 (4.97)	7.55 (6.40)	25.40 (2.90)	3.40 (1.68)
FLAIR	60.69 (4.89)	5.91 (3.66)	25.09 (1.96)	2.79 (0.26)
DWI	60.69 (4.89)	5.91 (3.66)	25.09 (1.96)	2.79 (0.26)
T1w + DWI	63.75 (6.08)	17.49 (13.63)	34.44 (10.81)	2.94 (0.61)
T1w + DWI + RP	80.95 (8.33)	67.06 (18.90)	63.80 (13.15)	1.51 (0.20)
T1w + DWI + RP	88.90 (6.89)	80.86 (11.23)	77.09 (9.28)	0.97 (0.31)

additive multi-class GP. Contrary to the current disease progression models used to study neurodegenerative diseases [138, 192], this approach does not assume a known order of events to stage the subjects in specific clinical status, neither an expected time-to-onset based on the familial clinical onset. Alternately, it finds the correlation between subjects at a similar stage of the disease, by means of the covariance kernel function. The predicted stages of the disease were then computed based on the highest probability across classes. The overall accuracy (Average Acc = 83.2%), suggests that the model had been successful in stratifying the subjects according to the MRC Scale. However, the analysis of the confusion matrix (Figure 5.2) suggests that the model was not sensitive to classes with closer intervals of the MRC Scale. The creation of well defined clinical stages of CJD goes beyond the scope of this work, but a future study should investigate alternatives to MRC Scale to define the labels used to train the model aiming the subjects staging. For a clinical application, as the aim of this work, the probability of the prediction associated with the predicted label gives relevant insights regarding the model performance.

The proposed approach is hampered by the noise introduced by the asymptomatic subjects, which show confounding features given their similarity either with the group of healthy controls (when far from clinical onset) or the subjects at the initial stages of the disease (when closer to the onset of the clinical symptoms). Despite the fuzziness of the results for these three classes, the model seems to differentiate with high accuracy the healthy controls group *versus* asymptomatic subjects. This can be explained by the existence of features that the model captures for the subjects closer to the clinical onset, included in the asymptomatic group. However, these results could be further investigated as future work, where two experiments should be performed:

1. Classification of asymptomatic subjects among a HC population - This experiment would aim to validate the results obtained in this chapter, as well as to evaluate the predictive probability obtained for each asymptomatic subject. This probability would verify the precision of the labels recognised as ground truth for the asymptomatic subjects. Consequently, the subjects with lower probability to be labelled as asymptomatic would present higher similarity to the HC group, hence these subjects would not be close to the clinical onset. Otherwise, a subject showing a higher probability of being label as asymptomatic would more likely to be close to the clinical onset. Note that these assumptions rely on the fact that the model can differentiate these two groups with accurately.

2. Stratification of the patients, excluding the HC group - The bias towards the HC population would be eliminated since only existing stages of the disease would be considered. Therefore, the model would better fit the several stages, ignoring the separation of the extreme stages, which has already been addressed by the diagnosis tool (Chapter 4).

Even though these two experiments address the limitations inherited by the asymptomatic group, further improvements in the model could potentially reduce the effect of noisy labels used at the training stage. Recent studies have applied an interval censoring approach to deal with the disease progression problem for neurodegenerative diseases [193–195]. Interval censoring, in statistics, defines a sampling scheme or an incomplete data structure. Specifically, a random variable of interest is known to only lie within an interval instead of being directly observed, hence its study is performed according to these pre-defined intervals instead of a continuous model [196]. Typically, in survival analysis and disease progression studies, the random variable is defined as the time taken until the occurrence of an event such as the clinical onset, death, a disease recurrence or a distant metastasis. Similarly, many clinical trials and longitudinal studies also generate interval-censored data for more sensitive modelling of the pattern of the considered features [196]. Example of that are the longitudinal studies that entail periodic follow-up. In this situation, an individual due for the pre-scheduled observations for a clinically observable change in disease or health status may miss some observations and return with a changed status. Accordingly, it is only known that the true event time is greater than the last observation time at which the change has not occurred, as well as less than or equal to the first observation time at the point the change has been observed to occur. As a result, for this particular sample, the time of occurrence of the change is an interval which contains the real (but unobserved) change. Therefore, this concept could be easily translated to the scenario of Prion disease, where for the particular case of the asymptomatic subjects it could be assumed, for each time-point, the probability of the true event (becoming symptomatic) is higher than for the previous observation. The interval censoring approach was already successfully applied to machine learning models, including both shallow [193, 194] and deep learning approaches [195], to conveniently model Alzheimer's disease patients. Therefore, the future work could benefit of a similar approach in the stratification of IPD patients.

Despite the fact that the formulation of the present approach as a multi-class task simplifies the problem greatly, it may reduce its clinical validity, given that the stages of the disease should follow an ordinal structure. The ordinal structure can be imposed by an ordinal likelihood function, as proposed by Chu *et al.* [189]. The proposed likelihood function is a generalisation of the *probit* function for Gaussian process [189]. Similarly, Doyle *et al.* [197], also used the concept of an ordinal likelihood function to estimate the progression of AD in a multivariate ordinal regression framework. Since the results of this study illustrate the efficiency of this framework to characterise the evolution of the disease over time, the proposed model could be adapted in a similar fashion.

Lastly, given that the proposed approach only takes in consideration cross-sectional data, the number of patients is reduced and the symptoms are highly heterogeneous, it is impossible to establish a function that describe the progression of symptoms over time. Thanks to the probabilistic nature of the predictions, the model gives not only information regarding the predicted class for a given time-point, but also about an estimation of the closest class for that time-point. This information can be used as a prognosis tool, since the transition between classes can infer the severity of symptoms and consequently the stage of the disease. Bearing this in mind, I used the trained model to predict the several stages of the disease for the subjects with scans before and after the clinical onset. Only five subjects had all three modalities available before and after onset, therefore the results are necessarily inconclusive. The results suggested that the proposed framework is currently inefficient to predict the time to the clinical onset, even if within a one year window to the first symptoms. Moreover, the different time-points for these subjects were modelled independently, which consists in to a source of bias in the model, since there is no dependency between results for a same subject. Nevertheless, these are interesting results that prove the need of a spatio-temporal covariance kernel function, used to link the several time-points for a same subject and to model the longitudinal information of the biomarkers.

Differential Diagnosis

I also investigated the possibility of using the model as a differential diagnosis tool. The current framework had proven to be able to recognise the individual features of prion disease among another form of dementia, in particular YOAD. Note however that the results reported in section 5.3.2 were achieved without a particular modelling for the two types of dementia. The current formulation of the model relies only on the proximity of features pattern for the subjects with the illness. This approach only correlates the magnitude of symptoms and the correlation between the features selected to a specific form of dementia.

By analysing the predictive accuracy of the latent models, this approach provides information regarding the combination of input features that better describes the signs of dementia for the two illnesses. Specifically, the latent models had shown that in clinical environment, DWI and T1w are the more relevant to identify sCJD. Therefore, the micro-structural changes happening in the brain of IPD symptomatic patients visible in DWI and FLAIR can be used as the main feature to distinguish this type of CJD from other neurodegenerative syndromes. From the results, I can conclude that the proposed framework can be used as personalised diagnostic tool, optimised to learn the best model for a specific aim, aside of learning the best kernel function used to explain the variance of the features that characterise the subjects' symptoms.

Following the good results obtained in this illustrative example, a new diagnostic tool based on quantitative measures could be created, which should account for the uncertainty of the diagnosis, given the similarity of prion diseases to other syndromes. This new diagnostic algorithm, developed to identify prion disease among other neurodegenerative diseases, would improve the detection accuracy of this illness, and thus address the current high rate of misdiagnosis patients[6]. In the future, I intend to adapt the model to learn what is the best covariance kernel function for

[6] Note that the initial symptoms of the disease are usually mistaken as depression symptoms, and even in later stages of the disease, the symptoms may be seen as dementia symptoms, particularly in the inherited form of CJD in which the progression is slower.

different neurodegenerative diseases.

5.5 Summary

In this chapter, I extended to work presented in Chapter 4 to: (1) stratify the CJD patients according to the severity of the symptoms as an indirect prognosis prediction, and (2) to identify CJD among other neurodegenerative diseases precursors of dementia. The results suggested that the model is effective in the identification of CJD among other neurodegenerative diseases. However, the proposed approach is less effective in the subjects stratification, failing in the prognosis of the subjects at clinical onset.

In the future, thanks to the flexibility given by the GP, this work can be extend to account for the longitudinal information available. This will allow not only a more accurate stratification of subjects based on the extracted biomarkers, but also the subjects prognosis in a given time frame. This can be addressed by integrating a spatio-temporal covariance model, such as the Kronecker form proposed by [198], to provide a unified framework to model jointly the time-series of biomarkers measurements with different natures, for a given subject.

More generally, due to their statistical nature, the performances approaches are negatively impacted by small sample size in the presence of normal or pathological variability. To address the limitation raised by the reduced sample size, I developed a new framework to deal with the missing data. This framework, detailed in Chapter 6, imputes the missing biomarkers, prior to implement the framework used for subject stratification, described in this chapter.

Chapter 6

Uncertainty embedding for partial data in Gaussian Processes

Due to the severity of the symptoms in the latest stages of CJD, it is often impossible to collect all the expected data, yielding an incomplete dataset. Because of the very limited amount of data available, it is key to use all available information, hence the need for data imputation. Current approaches to address the challenge of missing data either require a high number of samples or assume a simple relationship between data. However, such approaches are not compatible with the reduced sample size and heterogeneity present in Prion disease data cohorts. To tackle this issue, I employ a Bayesian framework to predict the missing values as well as the uncertainty associated with those predictions. Therefore, instead of excluding the subjects without data from the three MRI modalities, these subjects increase the sample in approximately 21% from the original sample of patients: from the 67 patients considered for training and testing, the new sample includes the full set of CJD patients – 85 patients. I then extend the classical Gaussian Process approach to take as input the uncertainty associated with each of the estimated features. The proposed framework is combined with the subject specific multi-modal feature extraction described in Chapters 3 and 4. Finally, the resulting model was finally used to stratify subjects with inherited forms of Prion disease into 5 ordinal stages from asymptomatic to extremely affected, in a regression fashion.

In this chapter, I present the robust GP used to impute the missing data from the available biomarkers (section 6.2.1), followed by another GP model used for subjects staging considering the uncertainty of the missing values (section 6.2.2). The proposed approach is extensively validated and compared with simplest approaches to deal with missing data using Prion disease dataset, as detailed in section 6.3.

Lastly, the effectiveness and usefulness of the proposed approaches is discussed in section 6.5, as well as the next steps and further improvements of the presented model.

6.1 Context

In research studies, it is common practice to acquire complementary information in the form of multi-modal imaging as well as non-imaging data. The acquired information can then be used to characterise a disease process or to train diagnosis or prognosis classifier [120, 199]. However, it is challenging to collect all anticipated data, yielding incomplete datasets. Common causes of missing data include patient drop-out, imaging artefacts and algorithmic failures, amongst others. A common strategy to use these incomplete datasets is to discard the partial samples, which is practical when a dataset contains a small number of missing data, and when the complete samples do not induce a sampling bias. Discarding incomplete samples becomes problematic when dealing with small samples, as the missing observations compromise the performance of classification algorithms [200, 201].

The limitation of missing data in neuroimaging studies has been mainly addressed by multi-task approaches [120]. By using a multi-task scheme for subjects' classification, it is possible to include different sources of data and sample sizes in a common model. Instead of removing samples with missing data, the subjects are grouped according to the sources of data available and a unique classifier is trained for the different learning tasks. Once the inference of the models is complete, the predictions resulting from the different tasks are combined for a more robust and accurate classification [120]. The performance of these models can be further improved by learning common features of interest used among different tasks [201]. Despite being able to deal with missing data, the multi-task approaches are limited and often not appropriate for the analysis of small samples. Due to the scheme used to separate the data in different classification tasks, the models do not use the maximum number of samples in each task, excluding observations during the training of the model. Note that the partitioning of the dataset according to the different tasks results in small and unbalanced sub-samples of the full dataset. Thus, most of the multi-task approaches are not appropriate to study rare or acute diseases given the need for a minimum number of samples to train each task. Furthermore, multi-source feature learning methods, such as the method proposed by Yuan *et* *al.* [201], assumes the linearity between the data and label. By jointly learning the same features from different modalities, this method also ignores the heterogeneity of features among the data sources, which hampers the information complementary given by the multi-modality datasets. The restrictive assumption of linear data-to-label relationship is tackled by Thung *et al.* [202]. They proposed a multi-input multi-output deep learning framework to deal with the incomplete multimodal data via multi-task learning, without involving data synthesis. The model also assumes a non-linear relationship between the data and the labels, and it allows multi-class classification. However, the model is limited by the need of a big sample to be conveniently trained.

Alternatively, imputation techniques are commonly used to estimate the missing values and preserve the original sample size, leading to a better characterisation of the dataset. Imputation methods are commonly developed based on statistical algorithms, such as quasi-randomisation inference (data-driven models) [200]. Among the different approaches used to deal with missing data, the expectation maximisation algorithm [203] and the k-nearest neighbour principles [204] have shown good results in estimating the missing values based on the observed samples. However, similarly to the aforementioned methods, the performance of imputation methods also rely either in high number of samples or they assume simple relationships between data.

In case of very rare diseases, such as Prion disease [7], is even more important to increase the sample size, in order to study in detail the biomarkers patterns that characterise the disease. But, because of the very limited amount of data available, it is key to use all available information and to avoid the partitioning of the sample, hence the need for data imputation in place of the multi-task approaches. However, some of the missing observations can be poorly imputed, when they do show a weak correlation with the remaining data considered. The symptoms associated with prion disease are highly heterogeneous, even among subjects at the same stage of the disease's course [181]. As a result, the imputed values can be considered as outliers, showing a high error in their estimation. Due to the error inherent to the missing observations estimation, the performance of the classification model is significantly hampered with biased and noisy observations. To tackle this problem, I propose a novel framework to subject stratification via additive Gaussian Process conditioned by the uncertainty of the imputed values. The algorithm is composed of two steps (1) imputation of missing data and their uncertainty using a GP, followed by (2) subjects stratification considering the complete dataset. The formulation of the imputation step accounts for the data heterogeneity by considering a robust observation model, such as the Student-t distribution, as detailed in section 6.2.1. Given the new sample composed by both the observed and imputed data, I implement an additive Gaussian Process to stratify the Prion patients according to the severity of the disease. To reduce the impact of poorly imputed values in the model estimation, I consider the uncertainty of the predictions in the hyperparameters estimation. Therefore, the proposed framework is able to handle missing data, without compromising the signal-to-noise ratio of the dataset. This method is an added-value in the study of rare or acute diseases, in which the imputation of the missing values is highly challenging and can compromise the performance of current classification models.

6.2 Model definition

Figure 6.1 shows the graphical representation of the full model that includes the (A) the imputation of the missing values, followed by (B) the predictive model definition, in which I include the impact of the uncertainty of the imputed values during the estimation of the hyperparameters.

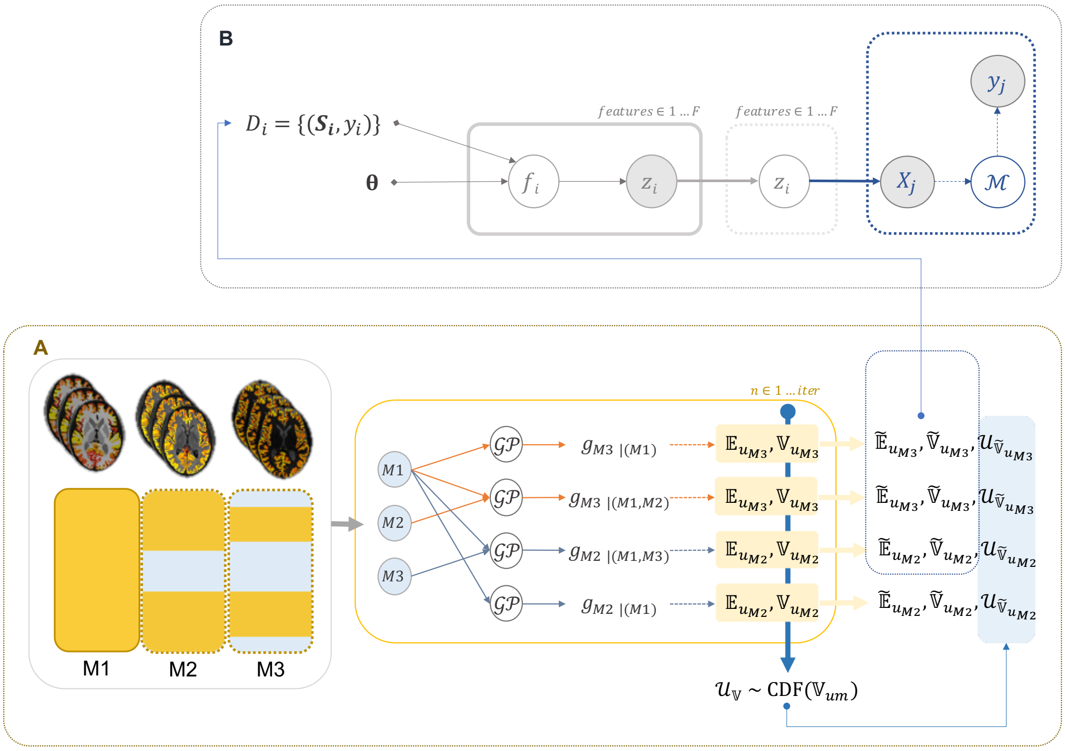
The subjects' stratification is obtained by a non-parametric kernel-based model to predict the status of Prion disease patients, according to multi-modality data available, as defined in Chapter 5. The model $\mathcal{M}$ is defined as follow:

$$\begin{gathered}\mathcal{M} : y = f(\mathbf{S}) + \varepsilon, \\ f \sim \mathcal{GP}(\mu_f; \mathbf{K} + \mathbf{I}\sigma_f), \ \varepsilon \sim \mathcal{N}(\mu_\varepsilon; \sigma)\end{gathered} \tag{6.1}$$

where for the subject i, $i = \{1, \dots, N\}$, the outcome $y \in \mathcal{Y}$ is inferred regarding a set of biomarkers $\mathbf{S} \in \mathcal{S}$ feature space. The function f describes the variance of the feature, which explains the response variable y using a GP model with μ_f and covariance kernel matrix $\mathbf{K}$, obtained by a stationary kernel function k, such as SE covariance function as detailed in Appendix B [112].

As described in the previous chapter, the relationship between features is modelled by an Additive Gaussian process [166, 181]. Considering the formulation of the model in Equation 6.1, I can write f as $f = \sum_{m=1}^{M} f_m$, with $f_m \sim \mathcal{GP}(\mu_{f_m}; \mathbf{K}_m + \mathbf{I}\sigma_{f_m})$, where M refers to the number of modalities taken into consideration in the model. Given the kernel properties, the addition of GP with $\mu_f = 0$ is equivalent to

Figure 6.1: Scheme of the full framework to deal with missing samples. Section A represents the imputation model $\mathcal{I}$. The samples from three different data sources M_1, M_2 and $M3$ correspond to features extracted from T1w, FLAIR and DWI, normalised as z-scored values. In the yellow outline section is depicted the estimation of the robust GP for imputation of missing values (section 6.2.1). Orange arrows represent the models used to estimate M3, whereas the blue arrows correspond to the estimation of M2. The blue shadow area represents the estimation of the uncertainty, $\mathcal{U}_{\mathbb{V}}$, based on the cumulative distribution function of the likelihood of the predictions (details in section 6.2.1). Section B represents the subjects stratification considering the complete dataset (section 6.2.2). The data for a subject i, D_i, is composed by both the observed biomarkers and the imputed values with associated variance, $\mathbf{S}_i$, and the subject's labels, y_i. The hyperparameters of the model, $\boldsymbol{\theta}$, are considered as input of the model. The function f_i generates samples of latent variables z_i by evaluating random non-linear mappings of latent inputs and then drawing mean-field samples parameterised by the mapping. These latent variables aim to follow the posterior distribution for a discriminative model $\mathcal{M}$, conditioned on data D_i. For a new sample X_j, the model $\mathcal{M}$ estimates the new label y_j. The observed variables are represented by shadow circles, whereas the estimated variables are represented by unfilled circles.


B
$D_i = \{(\boldsymbol{S}_i, y_i)\}$
$\boldsymbol{\theta}$
features ∈ 1 ... F
features ∈ 1 ... F
f_i
z_i
z_i
X_j
$\mathcal{M}$
y_j
A
M1
M2
M3
$n \in 1 \ldots iter$
$M1$
$M2$
$M3$
$\mathcal{GP}$
$\mathcal{GP}$
$\mathcal{GP}$
$\mathcal{GP}$
$g_{M3}\ |(M1)$
$g_{M3}\ |(M1,M2)$
$g_{M2}\ |(M1,M3)$
$g_{M2}\ |(M1)$
$\mathbb{E}_{u_{M3}}, \mathbb{V}_{u_{M3}}$
$\mathbb{E}_{u_{M3}}, \mathbb{V}_{u_{M3}}$
$\mathbb{E}_{u_{M2}}, \mathbb{V}_{u_{M2}}$
$\mathbb{E}_{u_{M2}}, \mathbb{V}_{u_{M2}}$
$\widetilde{\mathbb{E}}_{u_{M3}}, \widetilde{\mathbb{V}}_{u_{M3}}, \mathcal{U}_{\widetilde{\mathbb{V}}_{u_{M3}}}$
$\widetilde{\mathbb{E}}_{u_{M3}}, \widetilde{\mathbb{V}}_{u_{M3}}, \mathcal{U}_{\widetilde{\mathbb{V}}_{u_{M3}}}$
$\widetilde{\mathbb{E}}_{u_{M2}}, \widetilde{\mathbb{V}}_{u_{M2}}, \mathcal{U}_{\widetilde{\mathbb{V}}_{u_{M2}}}$
$\widetilde{\mathbb{E}}_{u_{M2}}, \widetilde{\mathbb{V}}_{u_{M2}}, \mathcal{U}_{\widetilde{\mathbb{V}}_{u_{M2}}}$
$\mathcal{U}_{\mathbb{V}} \sim \mathrm{CDF}(\mathbb{V}_{um})$

$f \sim \mathcal{GP}(0; \sum_{m=1}^{M} \mathbf{K}_m + \mathbf{I}\sigma_{f_m})$. Therefore, the matrix $\mathbf{K}$, which encodes the imaging biomarkers, is obtained by the addition of the kernel matrices computed individually using the information extracted from different sources of information. For simplification purposes, here I ignore the contribution of the rate of progression, defined as $\mathbf{K}_{\text{RP}}$ in Chapter 4. The missing observations in each kernel $\mathbf{K}_m$ are imputed by means of a robust Gaussian process regression, detailed in section 6.2.1.

Lastly, the estimation of $\boldsymbol{y}$ requires to find the best hyperparameters associated to each kernel covariance function. The hyperparameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \sigma_{f_m}\}$ related to the kernel functions are estimated via the maximisation of the marginal likelihood of the model, $p(\boldsymbol{\Theta}|\mathcal{M})$. Note that $\boldsymbol{\theta}$ is the vector of parameters estimated for each kernel as: $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]$. The marginalisation over the hyperparameters is performed by variational inference (Equation 6.8), for which the input features are not point samples, but distributions obtained from the imputation step around the expected value for an observation $\mathbf{S}$, as detailed in section 6.2.2.

6.2.1 Robust Gaussian Process for missing data estimation

I impute the missing values by means of a GP regression. Specifically, given set of modalities available $\mathbb{M} : \{1, ..., M\}$ and a corresponding task t, the missing values for the missing modality $m \in \mathbb{M}$ are obtained by the estimation of their correlation with the observed imaging modalities $\mathbb{M}^t = \mathbb{M} \setminus \{m\}$. The imputation model is defined as $\mathcal{I} : u_m = g_{\mathbb{M}^t}(\mathbf{x}_{\mathbb{M}^t}) + \epsilon$, where $\mathbf{x}_{\mathbb{M}^t}$ is the matrix of observed features extracted from the available modalities and u_m is the imputed value for the modality m, conditioned by the observed values of the other MRI pulse sequences, such as $g_{\mathbb{M}^t} \sim \mathcal{GP}(0, \mathbf{K}_{\mathbb{M}^t} + \mathbf{I}\sigma_u)$ with $\epsilon \sim (\mu_\epsilon, \sigma_\epsilon)$. For the sake of simplicity, the notation $g_{\mathbb{M}^t}$ is equivalent to g_m, as well as $\mathbf{K}_{\mathbb{M}^t}$ reads $\mathbf{K}_m$. The $\mathbf{K}_m$ is computed using a squared exponential kernel function as described by Equation 4.9 in Appendix B, with hyperparameters $\boldsymbol{\Omega} = \{l_{u_m}, \sigma^2_{u_m}\}$. Note that despite the equivalent notation used to define the kernel matrix, $\mathbf{K}_m$, here the features are different from the ones considered in the full model $\mathcal{M}$. As a results, the noise of the population, σ_u, differs from the noise σ_{f_m}.

To reduce the effect of the heterogeneity of symptoms across subjects, I use a robust observation model, using the Student-t likelihood function [205, 206]. This observational model is specially adequate to predict the features related to heterogeneous diseases, since it reduces the influence of the outlying observations and improves the predictions. Formally, the observational model is defined as an outlier-prone of order n, if $p(u|\boldsymbol{g}_{1,m},...,\boldsymbol{g}_{i+1,m}) \rightarrow p(u|\boldsymbol{g}_{1,m},...,\boldsymbol{g}_{i,m})$ as $\boldsymbol{g}_{i+1,m} \rightarrow \infty$. Note that this formulation contrasts with the Gaussian observational model for which each observation influences the posterior independently of its distance to the other observations. The Student-t distribution, equation 6.2, where v is the degree of freedom that take the value 2, and σ_p is the scale parameter, will reject up to ω outliers if there are at least 2ω observations, to be optimised during the training of the model.

$$p(u_m|g_m,\sigma_p^2,v) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{v\pi\sigma_p}}\left(1+\frac{(u_m-g_m)^2}{v\sigma_p^2}\right)^{-\frac{v+1}{2}}. \tag{6.2}$$

The marginalisation over the kernel parameters is achieved via EP, as described in Chapters 4 and 5. The EP algorithm was previously used by Jylänki *et al.*, as approximation method when using the Student-t likelihood in a GP framework [205, 206]. Adapting the Equations 4.13 and 4.14, for the individual tasks defined as M1, M2 and M3 (Figure 6.1), I estimate the expected value, Equation 6.3, and the variance of the missing biomarker, Equation 6.4. The $*$ notation refers to the testing sample that can comprise only an unseen subject j, or several samples.

$$\mathbb{E}_{u_m|g_m}\left[u_{*,m}|\mathbf{X},g_m,\mathbf{x}*\right] = \mathbf{K}_{*,u_m}\nabla \mathrm{log}p(g_m|\boldsymbol{u}_m) \tag{6.3}$$

$$\mathbb{V}_{u_m|g_m}\left[u_{*,m}|\mathbf{X},g_m,\mathbf{x}_*\right] = k_{*,*} - \mathbf{K}_{*,u_m}\left(\mathbf{K}_{u_m,u_m}+\mathbf{I}\sigma_u\right)^{-1}\mathbf{K}_{u_m,*} \tag{6.4}$$

The imputation is made per brain region, before the feature selection step, described in Chapter 4. This ensures the spatial coherence between the features extracted from different MRI pulse-sequences, whilst maximising the predictive power of the imputation model.

The variance distribution of the imputed values over the population is used in order to estimate the associated imputation uncertainty. Using bootstrapping, I compute the cumulative density function (CDF) of the variances across the subjects for each run. Even without an explicit correlation between the variance of the predicted labels and the uncertainty of the estimations, I use this measure as a proxy of the uncertainty of the imputed values, as proposed by Quiñonera-Candela *et al.* [207]. Thus, given a new subject j with variance $\mathbb{V}_j$, I can estimate its uncertainty $\mathcal{U}_{j,m}$ for modality m given the predictions for the population. The obtained values of uncertainty are in the range [0,1] that denotes low and high uncertainty respectively.

In the following section, I introduce a scheme to include these information in a GP regression, in order to condition the relevance given to the subjects with poorly imputed modalities in the estimation of the final model.

6.2.2 Embedding the uncertainty of imputed data

The existence of uncertain data in samples used to infer data-driven models is common. Most of the current techniques address the impact of noisy labels in the training of the model, reducing random classification noise and improving the classification accuracy [208, 209]. However, the impact of uncertain samples can be also investigated from the inputs perspective. To reduce the impact of noisy inputs, Henao *et al.*, suggest to combine inline active set selection with hyperparameter optimisation [210]. The method uses an active set selection method to select the most relevant inputs, in which the selection criterion is based on the weight of the sample during the training phase. In GP classification this is equivalent to using the predictive distribution to select the best set of inputs in each iteration, until the performance of the model converges. However this approach requires a large dataset, hence is not appropriate to study rare diseases. Differently, Dallaire *et al.*, included both certain and noisy inputs to infer the model, by means of a modified covariance function to account for the uncertainty effects [211]. Therefore, they suggested that by assuming a Gaussian distribution with known variances over the inputs and a Gaussian covariance function, it is feasible to marginalise out the inputs uncertainty, whilst keeping an analytical posterior distribution over the function. The new model outperformed a classical GP regression proving the advantage of considering the uncertainty effect on the hyperparameters estimation. Despite showing good results, this approach does not model explicitly the effect of the uncertainty in the estimation of the hyperparameters, but only models the distribution of the input values.

To overcome the aforementioned issues, I propose a strategy to include the effect of uncertainty directly in the estimation of the GP. Conversely to the other techniques, the method does not depends on a large sample, neither requires the exclusion of observations. In fact, the model $\mathcal{M}$, defined in equation 6.1, is optimised in order to account for the inputs uncertainty by modelling the input data through a Gaussian process latent variable model (GP-LVM), as proposed by Lawrence *et al.* and Michalis *et al.* [212, 213]. Therefore, the marginal likelihood of the data is expressed as:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{S})p(\mathbf{S})d\mathbf{S}$$
$$p(\mathbf{S}) = \prod_{m'=1}^{M}\prod_{i=1}^{N}\mathcal{N}(\mathbf{s}_{i,c}|0, \mathbf{I}\sigma_{m'}) \tag{6.5}$$

where **S** is the latent variable and *m'* indexes the M modalities (not to be confused with missing modality m defined in the previous section). Note that for each subject i, the feature $\mathbf{s}_{i,m'}$ is defined as:

$$\begin{cases} \mathbf{s}_{i,m'} \sim \mathcal{N}(\mathbb{E}_{u_m|g_m}, \mathbb{V}_{u_m|g_m}), & \text{if} \quad m' \notin \mathbb{M}_i^t, \\ \mathbf{s}_{i,m'} \quad \text{otherwise,} \end{cases} \tag{6.6}$$

where $\mathbb{M}_i^t$ is the set of existing imaging modalities for subject i. This way, the prior $p(\mathbf{S})$ encodes the variance of the estimations of the imputed values, since each prediction becomes a set of possible points distributed in the space according to results of the model $\mathcal{I}$. Note that poorly imputed values will show a wider distribution around the mean expected value obtained by $\mathcal{I}$. The marginal likelihood of the data is intractable. Following Michalis *et al.* [213], I apply an approximate variational distribution $q(\mathbf{S})$ to approximate the true posterior distribution $p(\mathbf{S}|\mathbf{y})$ over the latent variables, described by:

$$q(\mathbf{S}) = \prod_{m'=1}^{M}\prod_{i=1}^{N}\mathcal{N}(\mathbf{s}_{i,m'}|\mu_{i,m'}, \sigma_{i,m'}) \tag{6.7}$$

where $\mu_{i,m'}$ and $sigma_{i,m'}$ take the form of the distributions of $\mathbf{s}_{i,m'}$ defined in Equation 6.6. Using this variational distribution, I can express the Jensen's lower bound on the $\log p(\mathbf{y})$ that takes the form:

$$F(q) = \int q(\mathbf{S}) \log \frac{p(\mathbf{y}|\mathbf{S})p(\mathbf{S})}{q(\mathbf{S})} d\mathbf{S}$$
$$= \widetilde{F}(q) - \mathrm{KL}(q||p) \tag{6.8}$$

The computation of the $\widetilde{F}(q)$ is performed using the lower bound estimation via variational sparse GP detailed in Michalis *et al.* [213], where auxiliary inducing variables $\mathbf{Z}$ are used in sparse GP models. As a result, the hyperparameters $\mathbf{\Theta} = \{\boldsymbol{\theta}, \sigma\}$ are optimised in order to maximise the logarithmic density of the predictions.

For simplification purposes, I define the likelihood of the predictions as Gaussian likelihood, rather than a *probit* function used to define a multi-class classification, as described in Chapter 5.

6.3 Experiments and Results

6.3.1 Robust GP for data imputation

Experiments

The proposed approach to deal with missing data is tested using the data from the NPMC, including both IPD and sCJD patients. These subjects are grouped based on the severity of symptoms y_i for subject i, as detailed in Chapter 5. These labels correspond to the five predefined stages of the disease: (1) healthy control (HC), (2) asymptomatic subjects (Asym.), (3) subjects within a year of clinical onset (CO), (4) early symptomatic subjects and (5) late symptomatic subjects as defined by their clinical assessment scores [21].

The section A of the proposed framework, Figure 6.1, is trained using the incomplete dataset, which includes the subjects with all three MRI pulse-sequences available. The training set includes all the groups aforementioned to avoid overfitting on the second task of this approach; i.e., to avoid the inclusion of any information regarding the subjects status in the estimation of the features, the training set includes subjects from all the possible stages of the progression of the disease, relying on the capacity of the model to identify which subjects should be considered in the estimation of the missing modality m, for a new subject j through the kernel matrix estimation.

For comparison purposes, I use also used a GPR with Gaussian likelihood to impute the missing biomarkers.

Figure 6.2 illustrates the CDF of the variance of the imputed biomarkers. The variance is estimated for both the datasets with 50% and 20% of missing data. The robust GP regression shows higher variance when used in a dataset with only 20% of missing data (dark red line), for DWI features, when compared with normal GP regression (blue line). Differently, the estimations through robust GP regression shows lower variance (green line) when used in a dataset with 50% of missing data. These results suggest that the robust GP is only advantageous when the percentage of missing data is higher and the estimation of the missing biomarkers is more

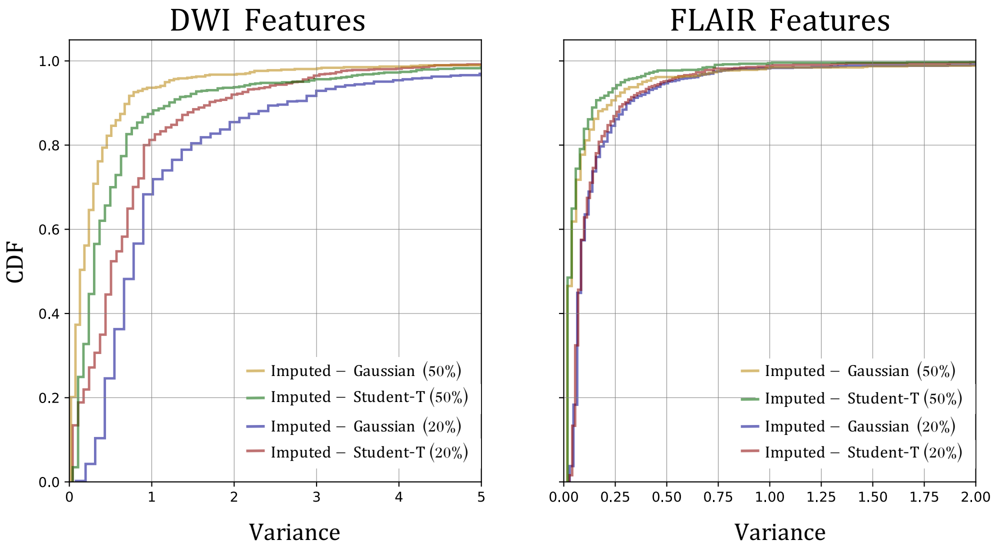


Figure 6.2: Example of the CDF of the variances across the subjects for the biomarkers imputed. The variance is estimated for both the datasets with 50% and 20% of missing data.

difficult due to the lack of information. On the other hand, there is no strong evidence of the advantage of using the robust GP for the estimation of biomarkers extracted from FLAIR (Figure 6.2, right panel.) Note that these results are just an example of the variance across one of the runs of the bootstrapping, requiring further validation.

6.3.2 Embedding the uncertainty of the imputed data

In order to assess the performance of the proposed model, I considered three experiments:

1. **Incomplete data:** Subject staging using the original dataset that has missing data and is therefore defined as the incomplete dataset;

2. **Completed data:** Subject staging using the original dataset and imputed data obtained with the model presented in section 6.2.1;

3. **Embedding uncertainty:** Subject staging using the original dataset and imputed data with uncertainty embedding as described in section 6.2.2. This experiment includes:

   (a) The model trained using inducing points exclusively picked from the real dataset;

(b) The model trained using a randomly selected sample from the full training set, which includes the augmented and observed data.

To evaluate the effect of size of the sample containing the missing data in the performance of our model, the incomplete dataset, which presents around 20% of missing data, was further perturbed to have 50% and 80% of missing data. I split the samples containing all three sequences into training and validation data, using the following two split proportions – 50%, and 80%. The samples missing one or more sequences were only used as training data. The remaining complete dataset is hold-out for evaluation purposes. A 5-fold stratified cross-validation was computed to assess the robustness of the results. For all analyses both the mean squared error (MSE) and the explained variance (η) are evaluated in order to quantitatively compare the models. In all experiments the GPy framework in Python was used to code the algorithms [214].

Results

The results from the three experiments are summarised in Table 1. When in presence of a higher percentage of missing data (50% of the sample missing), the model achieves a lower error in the regression of the stage of the disease (1.36 ± 0.22) for a training set containing 50% of the sample) and a higher percentage of explained variance (0.51 ± 0.08). However, when the training set is increased to 80% of the sample, the error of the regression task is lower when the information regarding the uncertainty is neglected. This indicates that the imputation using the robust Gaussian Process is actually being efficient in the prediction of the missing values when an adequate training set is used. Furthermore, in the case of small datasets the imputation using a classical GP is outperformed by the robust GP, showing a MSE superior in both training schemes. Conversely, the model seems to be less efficient when the percentage of missing data is lower (Table 1, 20% of missing data). In this case, by using the incomplete dataset, the regression task achieves better performance.

6.4 Discussion

The presented results suggest that the imputation model $\mathcal{I}$ is increasing the noise in the sample and therefore lowering the performance. The results can be

Table 6.1: Incomplete data $\mathcal{O}$ refers to the initial dataset, with missing data; $\mathcal{I}$ Gaussian refers to the refers to the dataset with the imputed missing samples through a GP regression with Gaussian likelihood; $\mathcal{I}$ (Student-T) is the augmented dataset via Robust GP; Ours (Z = $\mathcal{O}$) is the model with augmented data, which embeds the uncertainty of the estimations using as inducing points Z the original dataset $\mathcal{O}$; lastly, Ours (Z = Ω) represents the model with augmented data and its uncertainty, for which the inducing points are picked randomly from the sample. The mean squared error (MSE) and explained variance score(η) are evaluated for different training set sizes: 50%, 80% of the data. The grey shadow represents the best result in each test. Mean (standard deviation).

	50% missing data			
	MSE		η	
	50%	80%	50%	80%
Incomplete $\mathcal{O}$	1.428 (0.366)	0.970 (0.354)	0.44 (0.154)	0.458 (0.195)
$\mathcal{I}$ (Gaussian)	1.490 (0.129)	0.924 (0.304)	0.462 (0.039)	0.458 (0.199)
$\mathcal{I}$ (Student -T)	1.480 (0.189)	**0.800 (0.355)**	0.456 (0.059)	0.466 (0.160)
Ours (Z = $\mathcal{O}$)	1.408 (0.146)	0.880 (0.286)	0.491 (0.047)	**0.491 (0.172)**
Ours (Z = Ω)	**1.356 (0.215)**	0.882 (0.261)	**0.508 (0.077)**	0.475 (0.168)
	20% missing data			
	MSE		η	
	50%	*80%*	*50%*	*80%*
Incomplete $\mathcal{O}$	**1.297 (0.185)**	0.963 (0.119)	**0.657 (0.091)**	0.720 (0.055)
$\mathcal{I}$ (Gaussian)	1.459 (0.127)	**0.820 (0.184)**	0.461 (0.026)	**0.786 (0.032)**
$\mathcal{I}$ (Student -T)	1.458 (0.078)	0.928 (0.194)	0.549 (0.015)	0.736 (0.061)
Ours (Z = $\mathcal{O}$)	1.454 (0.178)	0.846 (0.224)	0.547 (0.051)	0.762 (0.064)
Ours (Z = Ω)	1.377 (0.114)	0.853 (0.249)	0.591 (0.016)	0.757 (0.073)

explained by the distribution of missing data across all the stages of the disease. Specifically, the percentages of missing data are not uniform for all stages but mainly predominant in the most advanced ones. This may lead to a poor characterisation of such stages and consequently an inaccurate imputation. However, further analysis with different datasets is required to assess the generability of the results. The results also support the assumption that including the uncertainty of the imputed values in the subsequent machine learning tasks improves performance, mainly when the percentage of missing data is high.

One of the challenges in validating this method is that there are no methods that can provide a direct comparison with this approach, since current deep learning techniques are not directly translatable to the study of prion disease. However, further validation is required, namely using different datasets to evaluate the generability and robustness of the method. Once this validation is performed, I can assess which approach is more robust in small datasets, such as those including medical data and used for the diagnosis and prognosis of rare and acute diseases.

6.5 Summary

I have proposed a novel framework for regression and stratification tasks in the presence of both reduced sample size and missing data, which is common in the diagnosis and prognosis of rare or acute diseases. Given preliminary results, I believe that this framework could improve the clinical assessment of rare and acute diseases by including valuable incomplete data that is usually discarded. In the future, I aim to extend the model to perform both regression and multi-class classification model.

Chapter 7

Conclusions and Future Perspectives

7.1 Conclusions

The human form of prion disease is a rare and fatal neurodegenerative disorder. Currently, there is no disease progression model that is able to describe the evolution of symptoms and the changes occurring in the brain over time. In fact, due to the high heterogeneity of the clinical manifestations of this illness, it has been very challenging to select useful biomarkers that may be used to characterise the different subtypes of CJD.

This thesis presents a first attempt to identify quantitative imaging biomarkers to diagnose CJD and characterise the evolution of the clinical manifestations during the course of the disease. I used a subject-specific feature selection framework, followed by a tailored Gaussian Process approach to correlated symptoms with disease types and stages. I applied the model to three different tasks: diagnosis, differential diagnosis and stratification. I obtained promising results on all three tasks. This work addresses an unmet need as it would enable to automatically identify patient with or at risk developing Prion disease.

In detail, Chapter 3 introduced the framework used to extract relevant biomarkers to detect CJD and to develop a disease progression model. The obtained quantitative features have shown promising results when used to identify symptomatic subjects among healthy controls. The key aspect of this framework is the subject-specific feature selection. This step enables the recognition of the relevant features, despite the heterogeneity of the clinical manifestations across subjects and the lack of spatial pattern of the brain changes. The biomarkers selected using this framework are broadly in agreement with previous study, where identical brain regions identified as abnormal due to CJD [33, 34, 55]. The proposed framework could benefit from the inclusion of more biomarkers, namely CSF biomarkers and blood samples. Besides, a bigger sample size would increase the power of the statistical analysis conducted on the selected features, demonstrating their validity in the clinical environment. Nevertheless, the objective of this chapter was reached, since I was able to handcraft meaningful features to be used in the diagnosis algorithm.

Chapter 4 presented the new diagnostic model, developed to identify CJD. This model aimed to improve the accuracy of detection of CJD while addressing the current high rate of clinical miss-diagnosis. Given that the initial symptoms of the disease are usually mistaken by depression symptoms, the new model could potentially alert the clinicians to the presence of prion disease. Even in later stages of the disease, the symptoms may be seen as dementia symptoms, particularly for the inherited form of CJD in which the progression is slower, compromising the diagnosis of CJD. The results showed the potential to use this framework to diagnose both forms of CJD, validating also the relevance of the features selected in Chapter 3.

In Chapter 5, I extended the model developed in Chapter 4 to perform the differential diagnosis of CJD. As a result, the new model identified CJD among healthy controls and other forms of dementia. This experiment consisted of a proof-of-concept, and it was validated on YOAD data, given the early onset of this illness. The promising results suggest that the proposed framework can recognise the individual features of prion disease among another form of dementia. In the future, the model will need to be improved to recognise the individual features of the other forms of dementia to reduce the ambiguity of the predictions. From that improvement, it would be possible the development of diagnostic criteria based on accurate and quantitative measures, and accounting for the uncertainty caused by the similarities of the clinical manifestations to other syndromes. Once more, this would consist of an advance to the clinical practice when treating prion disease patients.

In Chapter 5, I also presented the framework used for subjects stratification. The framework was developed intending to work as prognosis tool. Alternately, to the current disease progression model, my model found the correlation between subjects at the similar stage of the disease, by means of the covariance kernel function. The predicted stages of the disease were then computed based on the highest probability across classes. The inspection of the results suggested that the model was not sensitive to classes with close intervals of the MRC Scale values. The creation of well defined clinical stages of CJD goes beyond the scope of this thesis, but a future study should investigate alternatives to MRC Scale to define the labels used to train the model for subject's staging. Besides, another factor that compromised the effectiveness of the applied models, as referred in Chapter 5, is the non-correlation between the age of clinical onset of the asymptomatic subjects and their relatives, who already had shown symptoms. This issue constrained the translation of the models used to study Alzheimer's disease as prognosis models.

Lastly, Chapter 6 introduced a way to deal with missing data. This is particularly relevant in case of rare and heterogeneous diseases, where the current state-of-the-art approaches are faulty, as they lead to noisy and uncertain predictions of the missing values. To tackle the limitations of the current approaches, I proposed a novel two-step framework for classification, regression or stratification, where Gaussian Processes were conditioned by the uncertainty of the imputed values. Firstly, an imputation scheme is used to account for data heterogeneity through a robust observation model. Second, when training using the real and imputed values, I considered the uncertainty of each individual imputation in the optimisation of the overall model. I applied the proposed technique for patient stratification, as a regression task, to the NPMC dataset. The results were insufficient to attest the advantageous of the proposed model. Nevertheless, these are preliminary experiments to assess the feasibility of this framework to be used to deal with missing samples in neurodegenerative studies, and it requires further validation with a bigger dataset.

Overall, my findings are broadly consistent with the theoretical assumptions drawn at the beginning of this project. I was able to accomplish the main objectives of this work.

7.2 Future Perspectives

Considering the work developed to date and its limitations, the future work related to the topic of prion disease could follow two main directions:

1. The development of a diagnostic criterion for the sporadic form of CJD;
2. The development of an accurate progression model for prion disease, which will enable to predict the age of clinical onset of asymptomatic subjects.

The two research paths are beyond the scope of this thesis. Consequently, I have not developed any work attempting to address them. Nevertheless, some preliminary analyses developed in this thesis can be used as starting points to approach these research paths. The first path requires a bigger sample size, composed by sCJD patients, to perform an adequate statistical analysis to select and validate specific features to characterise sCJD. Furthermore, this would also require an intimate collaboration between engineers and clinicians to conveniently model the sCJD features and test their clinical validity. Lastly, other approaches to extracted meaningful features, such as deep learning models, should be considered.

The second research path is intimately linked to the work developed in Chapter 5. The development of an accurate disease progression model would enable to: (1) achieve a more sensitive subjects' stratification based on the clinical symptoms and (2) to predict the age of clinical onset of asymptomatic subjects. This path of research requires to model the longitudinal data of IPD subjects to properly identify the biomarkers evolution over time (section 7.2). However, this work is still hampered by the difficulty of normalising the time of the clinical assessment across subjects.

The work developed till the date can be also further improved by addressing its main limitations. Therefore, in the following sections, I detail possible improvements to the work presented in this thesis.

Longitudinal analysis of prion patients

Thanks to the flexibility given by GP, the model proposed in this thesis for subjects' stratification can be modified to account for the available longitudinal information. This will allow not only a more accurate stratification of subjects based on the extracted biomarkers, but also the subjects prognosis in a given time frame. That modification would consist in the integration of a spatio-temporal covariance model, such as the Kronecker form proposed by Lorenzi *et al.* [198], to provide a unified framework to model jointly the time-series of biomarkers measurements with different natures, for a given subject.

Lorenzi *et al.*, also proposed a disease progression model within a probabilistic setting to quantify the diagnostic uncertainty of individual disease severity in an hypothetical clinical scenario, with respect to missing measurements, biomarkers, and follow-up information [138]. This model is especially designed to account for longitudinal information, particularly by assuming that each individual measurement is made with respect to an absolute time-frame through a time-warping function. Theoretically, this model is directly translated to any time-frame function, which can include a non-linear function to conveniently represent the reality of CJD progression. Both of these approaches can be extended to integrate the longitudinal information in the models proposed in this thesis.

Currently, I am working on a hierarchical GP, as proposed by Hensman *et al.*, [215] to model the longitudinal information. This model also includes a coregionalised kernel to link multiple outputs, in order to improve the subjects stratification considering the mutual influence of multiple labels[1] available [216]. This model will in one hand improve the characterisation of the CJD symptoms during the course of the disease by reducing the effect of noisy labels, and in the other hand take advantage of the longitudinal information available to increase the sample size.

Differential diagnosis tool

The differential diagnosis tool would benefit from a more extended sample for training and validation purposes. The sample should comprise IPD, sCJD and other forms of dementia that simulate both the clinical manifestations of prion disease and the MRI signs, particularly YOAD, multiple sclerosis and AD. Therefore, future work should start by the collection and organisation of a complete dataset to perform the aforementioned analysis.

On the other hand, considering the main limitation of the differential diagnosis tool proposed in Chapter 5, the model could be improved by including a feature set that includes handcrafted features that are not only prion specific, but also spatial information required to accurately identify other forms of dementia. By training the model in a more extensive feature set, I would decrease the bias created by the subject-specific features, mainly used to characterise CJD.

Also, the identification of meaningful features used in the differential diagnosis tool would benefit of recently emerged deep learning techniques, which have proven to be accurate in the characterisation of CJD patients, when in the presence of other forms of rapidly progression dimension [217].

Lastly, since CJD symptoms can be interpreted as depression symptoms at the earlier stages of the disease, the differential diagnosis tool would also benefit from a training sample including subjects with depression. Many studies have reported relevant neuroimaging features that can be used to diagnose depression [63, 117, 218], namely features extracted from T1w MRI and fMRI. These features can be easily included in the current formulation of the model, if guaranteed the spatial

[1] The labels available include: MRC Scale score, subjects' status and time of conversion of symptomatic subjects.

correlation of the features across the different diseases. However, given how rare CJD is compared to depression the model training would need to be adapted in order to deal with an imbalanced dataset. This imbalanced dataset problem consists of learning a concept from the class that has a small number of samples, as defined by Fernandez *et al.*, [219]. In this particular case, the CJD sample size is significantly smaller when compared with the depression sample, as well as its prevalence in real clinical environment. Techniques to overcome this problem include algorithms such as the synthetic minority over-sampling technique (SMOTE) [220] and adaptive synthetic sampling (ADASYN) [221]. SMOTE is an oversampling approach that creates synthetic minority class samples using the information available in the data. For each sample from the minority class a set of k nearest neighbours is defined. To create a new synthetic data point, it is considered the vector between one of those k neighbours, and the actual sample. By multiplying this vector by a random number x which lies between 0, and 1 and adding this to the current data point, the new synthetic data point is created [222]. It potentially performs better than simple oversampling and it is widely used [222]. On the other hand, ADASYN builds on top of SMOTE, by shifting the importance of the classification boundary to those minority classes which are difficult to estimate. I.e., ADASYN is based on the idea of adaptively generating minority data samples according to their distributions: more synthetic data is generated for minority class samples that are harder to learn compared to those minority samples that are easier to learn. Consequently, the ADASYN method can not only reduce the learning bias introduced by the original imbalance data distribution, but can also adaptively shift the decision boundary to focus on those difficult to learn samples [221]. These algorithms provide effective solutions to over-sample the minority sample, without loss of information by under-sampling the majority sample, avoiding the under-fitting of the model and poor generalisation to the test set. Note that all these techniques could potentially be used not only for the training of the model using depression data, but also for any disease that has a sample size significantly bigger than CJD, improving the predictive accuracy of the differential diagnosis tool.

Dealing with missing data

Given that CJD is a very rare disease and highly heterogeneous, thus very challenging to model, further improvements to the proposed imputation method, detailed in Chapter 6, could lead to a more accurate characterisation of CJD. In particular, the model presented in Chapter 6 should be modified to perform both imputation and the optimisation of the parameters of the classifier (or regression model) at the same time. The new model would exclude the 2-step approach, as well as the errors associated to the imputed values. To achieve that, the new model would include a variational inference approach as proposed by Dalca *et al.* [223], where the biomarkers estimated would directly constrain the global cost function used to estimated both the model parameters and the missing values.

In more general notes, the dataset used in this study is one of the largest worldwide, comprising the data from all United Kingdom [11]. Given the rarity of the of prion disease, the research of this illness would benefit from the creation of an international dataset, covering both IPD and sCJD patients. The new dataset would increase the sample size, as well as including a broader spectrum of phenotypes that could validate the assumptions presented in this thesis.

Appendix A

GP optimisation on presence of limited samples

Liane S Canas, Benjamin Yvernault, Carole H Sudre, Jorge Cardoso, John Thornton, Frederik Barkhof, Sebastien Ourselin, Simon Mead, and Marc Modat. Multikernel Gaussian Processes for patient stratification from imaging biomarkers with heterogeneous patterns. In Learning from Limited Labeled Data: Weak Supervision and Beyond, NIPS*, Long Beach, 2017*

A.1 Context

The approaches used to study other types of neurodegenerative diseases, such as Alzheimer's disease, are not appropriate to capture the progression of the human form of Prion disease. This is largely due to the heterogeneity of the phenotypes associated with Prion disease. The biomarkers heterogeneity, combined with the rarity of the disease and thus the limited amount of available data, hampers the ability of state-of-the-art models to stratify patients in disease stages accurately. The proposed model for subjects' stratification, based on GP models, allows to stratify the patients even if in presence of small sample size. The proposed design also tackles the insufficient number of training data and the presence of heterogeneity among subjects' biomarkers. This is achieved by using a subject-specific multi-modal feature extraction scheme in order to normalise the data across subjects. The model is compared with other schemes to feature selection in GP modelling, such as the ARD scheme [112]. Through a simulated dataset, it is highlight the rationale and added value of the proposed technique before applying it to real data. The model considered here is the model defined in Chapter 5, for subjects' stratification. The target classes, defined by $y_i \in \{1, ..., 7\}$ for the subject i, comprise seven stages of the disease: HC – healthy control, Asymp. – asymptomatic subjects, CO – subjects at clinical onset , SI to SVI – symptomatic subjects divided in 4 groups according to their clinical scores (MRC Scale) [59].

A.2 Experiments and Results

The experiments and corresponding results detailed in the sections below are part of the *Multikernel Gaussian Processes for patient stratification from imaging biomarkers with heterogeneous patterns* study, presented at Learning with Limited data Workshop, NIPS 2017. Those include both experiments using synthetic data, section A.2.1, and real clinical data, section A.2.2.

A.2.1 Synthetic Dataset

Experiments

To demonstrate the efficacy of the proposed approach as well as its rationale, a synthetic dataset is created. The dataset is derived from three functions designed to simulate the three MRI modalities used in this study. For each one of them, referred to as modality A, B and C in Figure A.1, it is generated 10 different biomarkers.

Biomarkers from the same modality are set to follow a common evolution over time but all deviate differently from the main function. The three main functions are altered to simulate the three different rate of disease progression previously mentioned, defined as clusters 1, 2 and 3 and represented by $\star$, $\triangle$ and $\circ$, respectively (Figure A.1, bottom row). The biomarkers $\tau \in [1, 10]$ corresponding to class y_i as a function $\mathbf{f}_i(\tau) = (f^i_{M_A}(\tau) + \epsilon_i, f^i_{M_B}(\tau) + \epsilon_i, f^i_{M_C}(\tau) + \epsilon_i), \epsilon_i \sim \mathcal{N}(0; 0.25)$. f_{M_A}, f_{M_B} and f_{M_C} are respectively to a monotonically increasing sigmoid function, a second order polynomial function and a monotonically decreasing sigmoid function, for a subject i. Note that the number of subjects per class is uniformly distributed. In order to simulate the spatial heterogeneity of features among subjects, one to three biomarkers τ are selected randomly to deviate from the controls samples. The estimation of $\hat{y}_{i,j}$ requires to find the best hyperparameters for the regression task and for each kernel. The hyperparameters Θ are estimated following the marginalisation of the likelihood detailed in Chapter 5 (section 5.2.1).

The model is used to estimate the probability of a subject to belong to class

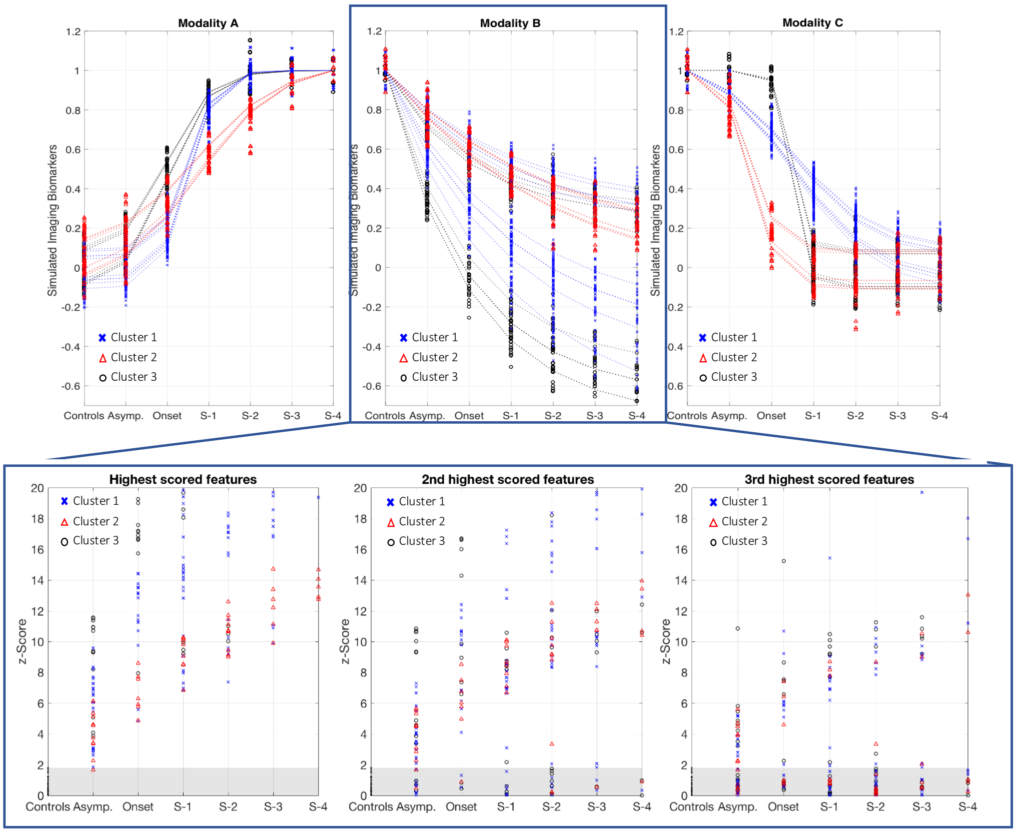


Figure A.1: Upper row: Synthetic data generated to simulate the three different modalities. The three different colour define the virtual clustering of subjects according to the rate of progression of IPD: fast, medium and slow, represented by $*$, △ and $\circ$. Lower row: Example of the data extracted from modality B after being normalised and ranked by the most significant features for each subject.

C. Given that, for this experiment, the synthetic data aims to replicate the pattern of the features of the actual clinical dataset, the response variable includes seven stages, labelled according to the clinical stages of the disease. Thus, the synthetic dataset shows the following labels, ordered from lower to higher z-scores: healthy controls (HC), asymptomatic (Asymp.), clinical onset (Onset), SI to SIV for the symptomatic stages of the disease.

For comparison purposes, it is also considered an additive GP with a squared exponential kernel (GP SE) in which all available features are considered, an additive GP with ARD scheme (GP - ARD) used for feature selection. All the models are trained with several sample sizes ($N \in \{100, 500, 1000, 2500\}$) to assess the validity of the assumptions of the proposed framework. For validation purposes, the samples are split into two sub-samples: a training set with 75% of subjects, and a testing set with the remaining 25%. The robustness of the estimations and the stability of the

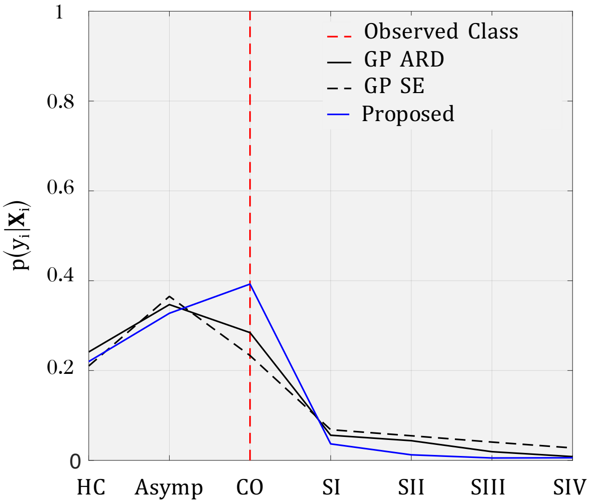


Figure A.2: Probability of the subjects at the clinical onset being correctly classified. Average of the predicted class over bootstrapping runs for an unseen subject j at CO. The predicted class is obtained from the argmax $p(\mathbf{y}_j|\mathbf{X}_j)$.

results are assessed through bootstrapping (500 runs).

Results

Table A.1 reports the results of the proposed approach on synthetic data, as well as the the models used for comparison: GP simple and GP-ARD. The mean percentages of misclassified subjects per class indicate that the proposed method is able to better capture the granularity of the disease stages than the standard approaches. The design is only outperformed by the GP-ARD when the sample size is very large, as GP-ARD is able to find correlations between subjects even using heterogeneous biomarkers. Note that in this case, the number of samples is considerably higher than the number of features: $N = 2500 >> F = 30$. Figure A.2 highlights the likelihood of an unseen subject being classified correctly as CO.

A.2.2 IPD dataset

Experiments

Using real patient data, I perform the stratification of IPD subjects based on their clinical diagnosis, using both imaging and genetic data. The dataset includes the baseline scans of control, symptomatic and asymptomatic subjects, as well as scans of subjects who have shown clinical symptoms within a year after their scan, as detailed in Chapter 3 (Table 3.1). The whole sample consists of 25 controls,

Table A.1: Percentage of misclassified subjects per class. N defines the sample size (number of subjects). The first models corresponds to an additive GP; second model is an additive GP, as implemented before, with ARD scheme for feature selection; finally, third corresponds to the approach proposed in this paper. The bold values correspond to the best approach in each test.

N	Model	Healthy Controls	Asymp. Subjects	Clinical Onset	MRC S. (20-16)	MRC S. (15-11)	MRC S. (10-6)	MRC S. (5-0)	Mean Percentage
100	*GP SE*	1.38	4.63	5.00	9.63	10.13	13.75	**4.75**	7.04
	GP-ARD	2.25	6.50	6.13	9.75	10.63	12.88	5.38	7.65
	Proposed	**1.00**	**1.13**	**1.38**	**5.13**	**8.50**	**7.88**	5.50	**4.36**
500	*GP SE*	**0.40**	1.36	2.00	8.00	11.44	10.96	7.12	5.90
	GP-ARD	0.72	2.16	3.28	7.04	10.08	10.08	7.44	5.83
	Proposed	0.56	**0.80**	**1.20**	**3**.28	**6.72**	**9.52**	**5.60**	**3.95**
1000	*GP SE*	**0.40**	1.60	2.67	6.93	10.27	13.33	**6.67**	5.98
	GP-ARD	0.93	4.40	6.27	8.80	10.67	10.27	6.93	6.90
	Proposed	**0.40**	**1.20**	**1.60**	**3.33**	**6.80**	**9.73**	6.80	**4.27**
2500	*GP SE*	**0.08**	**0.08**	0.48	3.76	11.04	10.48	4.96	4.41
	GP-ARD	0.24	0.24	0.48	1.92	**5.20**	**6.96**	4.48	**2.79**
	Proposed	**0.08**	0.16	**0.08**	**1.84**	6.24	8.16	**4.40**	2.99

Table A.2: Percentage of misclassified subjects per class using clinical data. N defines the sample size (number of subjects).

N	Model	HC	Asymp.	CO	SI. (20-16)	SII. (15-11)	SIII. (10-6)	Overall
89	*GP SE*	14.58	21.53	**4.66**	11.41	**2.18**	**0.20**	7.79
	GP-ARD	14.48	16.17	5.36	11.31	2.68	**0.20**	7.17
	Proposed	**12.30**	**14.88**	4.96	**9.23**	2.28	**0.20**	**6.26**

29 asymptomatic, 5 close to clinical onset and 30 symptomatic subjects, yielding unbalanced classes. The features considered in this experiment are obtained as described in Chapter 3, section 3.3.2.

The trained of the model is performed on a sample composed of 75% of the groups mentioned above. The testing set includes the remaining 25% of each group. Consequently, even with unbalanced classes, the model is consistently trained and tested on the same proportion of subjects per class. This training scheme does not overcome the bias towards the classes with higher sample size. Nevertheless, due to the small sample size of the dataset, it is not possible to overcome this issue without data augmentation. The robustness of the estimations and the generability of the results are assessed through bootstrapping (500 runs). Similarly to the experiment performed using synthetic data, three models are considered: an additive GP with a SE function in which all available features are considered, an additive GP with ARD scheme and the proposed model. The number of subjects is kept fixed for this experiment, comprising a total of 89 subjects.

Results

Table A.2 reports the results of the proposed approach to deal with small samples, when used to stratify IPD patients. The classification rate for SIV class is not evaluated due to the absence of individual in IPD dataset with MRC Scale score lower than 5. The proposed method reduces the misclassification rate per class, when compared with the other methods tested. Note that the results reported in Table A.2 are deterministic and they do not account for the fuzziness of the classes estimation. The results support the hypothesis that for rare diseases, like IPD, due to the reduced number of subjects the problem is ill-posed. In this specific case, the results achieved by the GP-ARD are not as accurate as the results obtained by an approach in which the number of features are pre-selected to avoid confounding effects raised by the inconsistency of features across subjects.

A.3 Summary

The results obtained with the proposed framework suggested that the feature selection, using z-scores normalisation, is more sensitive in the prediction of clinical onset than the ARD scheme. This approach was specifically design for the study of prion disease, accounting for its rarity and the lack of biomarker geometrical consistency across subjects. In fact, the model showed better results when compared with the current frameworks used in context of neurodegenerative diseases, both on synthetic and clinical data.

Appendix B

Optimal kernel construction for neuro-degenerative clinical onset prediction

Liane S. Canas and, Benjamin C. Yvernault, David M. Cash, Erika Molteni, Tom Veale, Tammie L. Benzinger, Sebastien Ourselin, Simon Mead, and Marc Modat. Gaussian Processes with optimal kernel construction for neuro-degenerative clinical onset prediction. In Medical Imaging 2018: Computer-Aided Diagnosis of SPIE*, volume 10575, Houston, 2018*

B.1 Context

The performance of the GP is highly dependent of the kernel function used in a specific context [112, 225]. In order to improve the performance of kernel-based prediction models, Duvenaud *et al* [113] have proposed a structure discovery algorithm through a compositional kernel search. Their study has successfully shown that, in supervised prediction tasks, the automatically learning of kernels outperforms both variety of kernel classes commonly used and kernel combination methods. As a brute force scheme, Duvenaud's approach is impractical due to the highly dimensional problem, their algorithm searches over a space of based kernels and operations using a greedy search approach: at each stage it chooses the highest scoring kernel and expands it by applying operations, such as addition or multiplication, with other basis kernel functions. However, this method does not account for the replacement of the kernel selected in a previous level. Furthermore, the method is design to preferentially select the best model based on the bayesian information criterion

(BIC) [226], which it only takes into account the balance between the marginal likelihood of the predictions estimated based on the training set, and the complexity of the model.

An extension of this approach is here proposed to overcome the potential pitfalls of the approach proposed by Duvenaud *et al.*, [113]. The optimal kernel search is then based on depth-first search in a pre-pruned tree, applying a greedy search to select the highest scored kernel in each layer of the tree. The approach also considers possible that a given branch of the search tree could be better than the one selected in the previous layer. It is further introduced a new energy function to evaluate the kernels function performance and a cost-function to select the optimal kernel. The energy function introduced takes into account the model predictions for the validation set, but it also considers the balance between the model performance and its complexity by including a term in the energy function in which the BIC is introduced.

B.2 Algorithm

The algorithm, Figure B.2, takes as input a set of basis covariance kernel functions $b \in \mathcal{B}$, defined in this paper as basis kernel functions, the design matrix of the features $\mathbf{X}$, a *c-by-N* matrix of c brain regions, extracted from N subjects, and the response variable vector $\mathbf{y}$, a *1-by-N* vector with the time to onset corresponding to each of the observations in the matrix $\mathbf{X}$. The basis covariance functions, illustrated in Figure B.1, are used to capture complex relationships in data which do not have a simple parametric form, through their product and/or addition. The basis covariance functions k are parameterised by the hyperparameters $\boldsymbol{\theta}$, as follow: the linear kernel $k_{Lin}, \boldsymbol{\theta}_{Lin} = \{\sigma_b, \sigma_v, w\}$ (Equation B.1), periodic kernel $k_{Per}, \boldsymbol{\theta}_{Per} = \{\sigma, p, w\}$ (Equation B.2), squared exponential $k_{SE}, \boldsymbol{\theta}_{SE} = \{\sigma, w\}$ (Equation B.3), matérn functions $k_{MA}, \boldsymbol{\theta}_{MA} = \{\sigma, r, w\}$ (Equation B.4) and linear logistic $k_{Logit}, \boldsymbol{\theta}_{Logit} = \{W, a, b\}$ (Equation B.5). Note that the kernel functions present the distance between two scalar samples $(x - x')$, which is replaced by $||x - x'||^2$ when the samples x are in fact vectors.

$$k_{Lin}(x, x') = \sigma_b^2 + \sigma_v^2 (x - w)(x' - w) \tag{B.1}$$

$$k_{Per}(x, x') = \sigma^2 \exp\left(- \frac{2 \sin^2 (\frac{\pi (x - x')}{p})}{w^2} \right) \tag{B.2}$$

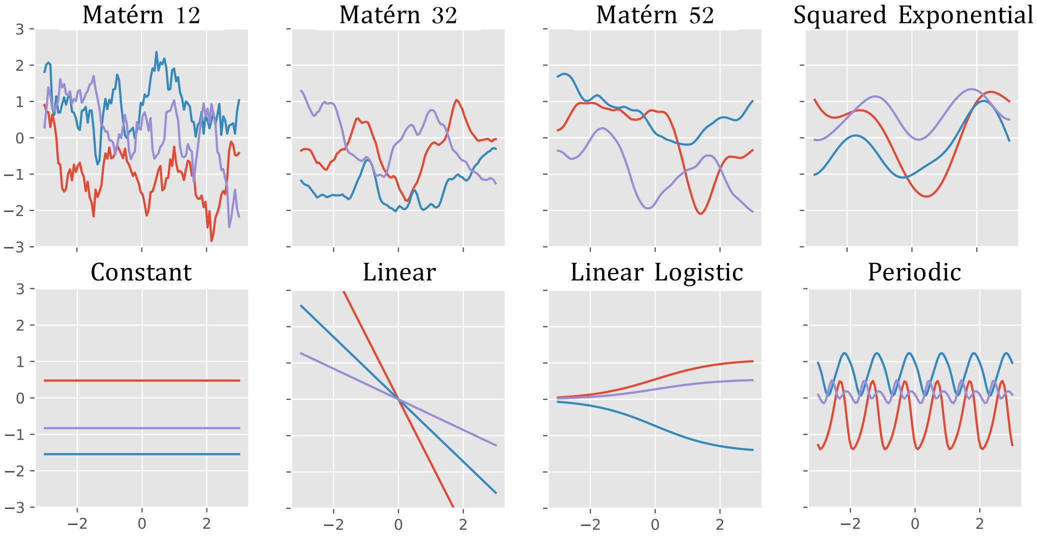


Figure B.1: Basis covariance functions included in the optimal kernel search. Example of the basis kernel functions described by Equations B.1 to B.5, when evaluated in 1D feature vectors. The x-axis has the same range on all plots.

$$k_{SE}(x,x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2w^2}\right) \tag{B.3}$$

$$k_{MA}(x,x') = \sigma^2 \left(1 + \frac{\sqrt{r}(x-x')}{w}\right) \exp\left(-\frac{\sqrt{r}(x-x')}{w}\right), r \in \{1,3,5\} \tag{B.4}$$

$$k_{Logit}(x,x') = W \text{logit}^{-1}(ax+b) - 0.5 \tag{B.5}$$

The model initialisation in the first layer $\mathcal{S}$ is $\varnothing$ followed by the evaluation of the algorithm performance when each one of the basis kernel functions is used. $\mathcal{S}'$ is the result of the expansion of $\mathcal{S}$ by operations with $\mathcal{B}$. This procedure (Figure B.2 – component 1, dark blue outline) is performed until all basis kernel functions are tested. The best kernel in each layer is selected by maximisation of the objective function, Equation B.6, where α and β are constants, $\hat{y}_{i,b}$ is the model estimations and $y_{i,b}$ is the vector of observed values correspondent to the response variable. Note that this objective function is one of the contributions of this approach, aiming the selection of the optimal kernel based on the accuracy of model, while balancing the complexity of it. Therefore, the first term of the equation B.6 evaluates the accuracy

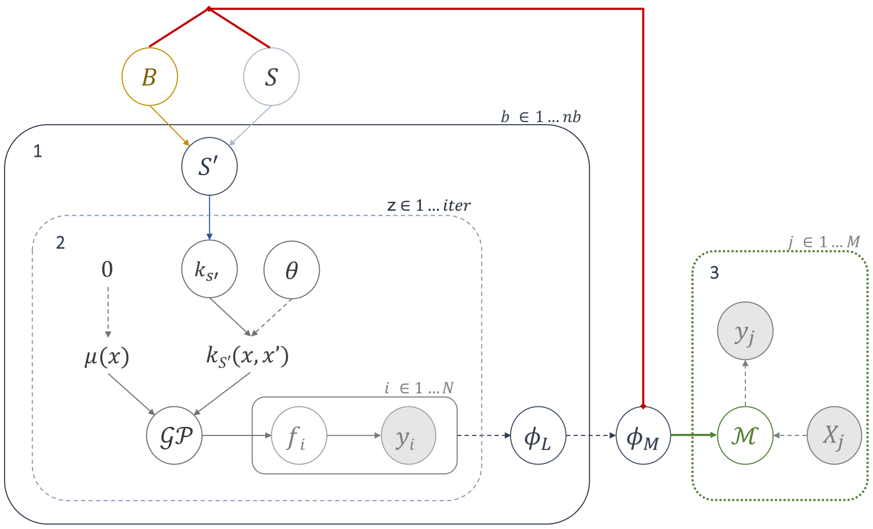


Figure B.2: $\mathcal{B}$ is the set of basis kernels. $\mathcal{S}$ is the kernel function initialised at each layer, whilst $\mathcal{S}'$ represents the kernel function after the expansion. f denotes the latent function, whereas y is the vector of values predicted using the estimated latent function. The subscripts i denote sampled latent values for each point to a maximum of N data points. $\mathcal{M}$ is the model selected and used to estimate y_j predictions of j observation given a set of features $\mathbf{X}_j$. The process 1 is repeated until the cost function ϕ_M converges - red line. The white circles represent functions or set of functions and the dark circles represent values of variables.

of the prediction of the model in the testing set by computing the explained variance of the model. The second term constrains the complexity of the model based on the marginal likelihood of the predictions of the training set, by computing the BIC associated with each kernel function tested, where N is the number of samples, z the number of parameters and $\hat{\mathcal{L}}$ refers to the maximised value of the likelihood function. Regarding the parameters of ϕ_L, it is required to compute the predictions of the model based on the kernel in analysis. Further, the estimation of $\hat{y}_{i,b}$ requires to find the best hyperparameters of each kernel. The hyperparameters θ of the kernel functions are estimated via the maximisation of the marginal likelihood of the model, $p(y|X, \boldsymbol{\theta})$, as described in equation 4.8; i.e., the marginalisation over the kernel parameters is performed by maximum *a posteriori* algorithm (MAP), and that the hyperparameters $\boldsymbol{\theta}$ are estimated by bootstrapping.

$$\operatorname{argmax}_{b \in B}(\phi_L(b)) = \alpha\left[1 - \sum_{i=1}^{N}\left(\frac{(\hat{y}_{i,b} - y_{i,b})}{(y_{i,b})}\right)\right] + \beta\left[1 - \exp\left(\frac{\mathrm{BIC}_b}{\max(\mathrm{BIC})}\right)\right], \qquad \text{where} \quad \mathrm{BIC} = \log(N)z - 2\log(\hat{\mathcal{L}}) \tag{B.6}$$

For simplification purposes, for each predictor p, an individual length scale is computed $\mathbf{w} \in \{w_1, \ldots, w_p\}$ using the ARD method. This method allows to establish different weights (relevance) for the features, as well as it allows to include in the same model biomarkers with different scales. By including the ARD approach, the relevant features are automatically selected among redundant predictors, avoiding the transformation of the feature space before the training of the model. The following layers of the tree correspond to the composition of a kernel function based on a set of operations between the basis kernel functions $b \in \mathcal{B}$ and the covariance function $\mathcal{S}$ obtained in the previous layer. Addition, product, replacement of basis kernel and replacement of the branch previously chosen are the valid operations in this algorithm. Attending to the pre-pruning behaviour of this algorithm, the maximum search depth is not defined. Instead, a cost-function is used as stopping criterion, Equation B.7, where l is the number of layers.

$$\phi_M : (|\phi_L(l-1) - \phi_L(l)| \leq 0.01 \cup \phi_L(l) \geq 0.9) \tag{B.7}$$

B.3 Experiments and Results

Experiments

The effectiveness of the approach is evaluated on Dominantly Inherited Alzheimer Network (DIAN) dataset [191]. The model is optimised to estimate the years to clinical onset of Inherited Alzheimer's disease patients, using clinical and imaging data. The sample is composed of 320 controls, 240 symptomatic subjects and 70 subjects that converted to symptomatic. The subjects' T1-weighted images are processed using the Geodesical Information Flows (GIF) algorithm [165] algorithm, and the volumes of brain region are then used as features. The impact of confounding effects in the features, such as age and the total intracranial volume, is regressed. The relevant features are selected using an elastic net regression. The sample is split into three sub-samples: training set that corresponds to 70% of the asymptomatic and converted subjects, the validation set comprises 20% of the subjects belonging to the aforementioned groups and testing set, which correspond to the remaining 10% of the sample. The optimal kernel is optimised using the training and testing sets. The validation of this approach aims to justify that the optimal kernel selection is an effective way to detect the pattern of the features considered, without a selection of the function that better explains their evolution over time.

Results

When compared with kernel functions defined *a priori* to explain time-series, the proposed approach selected a kernel function with equivalent coefficient of determination R^2, but lower BIC (table B.1). This fact suggests that this approach is able to find a function that explains conveniently the variance of the features, with lower level of complexity considering the likelihood of the prediction attending to the BIC achieved. The algorithm selected a linear combination of linear logistic functions, which supports the biological assumptions regarding the features used. The results also support the assumption that this approach may be extended to other diseases, without modelling explicitly the pattern of features used to characterised them. The performance of the model for the prediction of the time to clinical onset of the subjects is also evaluated. Note, however, that the prediction of the time to onset is highly dependent of the data used to train the model. In this study, a large number of subjects considered have not converted to symptomatic to the date; whereby, the time to onset used as response variable is not the real age of clinical onset, but an estimation based on the age of clinical onset of the their family. In the future, I aim to validate the proposed framework in a dataset that accounts for the uncertainty of the labels used as response variable.

Table B.1: Evaluation of the kernels used to predict the time to clinical onset, given a set of structural features. The results obtained with different functions commonly used to explain time-series datasets, a variation of the Compositional Kernel Learning (CKL) [113], and the proposed approach.

Approach	R^2	*RMSE*	*BIC*
GP - LIN	0.401	5.61	-12.5
GP - (LIN+SE)	0.392	5.64	53.2
GP - SE (ARD)	0.423	5.48	48.1
GP - CKL	0.407	5.60	-2.67
Proposed approach	0.432	5.52	12.3

B.4 Summary

The results have shown that the learned structure of the covariance kernel function is capable of an accurate extrapolation in a complex time-series, such as the evolution of brain volumes over time, and it is competitive with the other methods tested given this prediction task. Even though some kernel functions such as the periodic kernel function are not expected to be relevant to encode the features pattern, I included them in order to make a fair comparison with the approach of Duvenaud *et al.* [113]. Regardless, the optimal kernel function obtained by the algorithm does not include those kernel functions, confirming the theoretical assumptions about the features pattern and the effectiveness of the approach.

The main limitation of the proposed method is the complexity of the kernel structure obtained when in the presence of a high level of noise in the data. Furthermore, the evaluation of the structures found requires the estimation of the full model; thus, the metrics taken into account in table B.1 to evaluate the performance of the model include the errors associated to the hyperparameters estimation and inference. In the future, this approach should be further validated to find additional support to the assumption that the model is as general as possible and it can characterise any type of features despite their nature.

Appendix C

Evaluation Metrics

I evaluated the performance of the models present in this thesis with respect to the robustness and accuracy of their predictions. The robustness of the estimations and the stability of the results are accessed by bootstrapping, as described in Chapters 4 to 6. The metrics used to evaluate both binary and multi-class classification are comparable, with small adaptations to account for the multiple classes scheme. The following sections describe the metrics used to evaluate the model performance in this study.

C.1 Notation

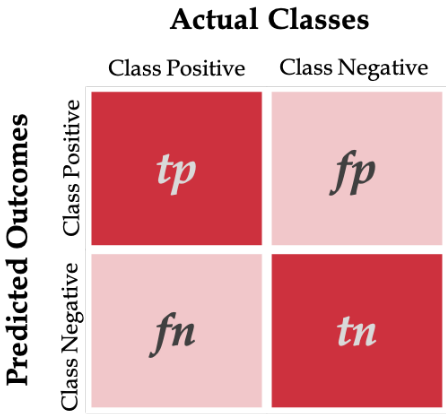
Consider any type of labels when data entries of subject i, $\mathbf{X}_i \in \mathcal{X}$ have to be assigned into predefined classes $C = \{C1, \ldots, C_{\mathcal{C}}\}$.

Given a classification task, it is generally defined as true negative (tn) cases the samples correctly identified as not belonging to a class $\mathcal{C}$; conversely, the cases correctly identified as class $\mathcal{C}$ are defined as true positive (tp). The false positive and false negative are to the cases incorrectly identified as being $\mathcal{C}$ or $\neg\mathcal{C}$, respectively. Figure C.1, illustrates a confusion matrix for binary classification.

C.2 Binary Classification

The probabilistic predictions are threshold at a particular value (th=0.5) to get binary labels used to computed the evaluation measures. The subjects diagnosis is accessed by quantitative measures, such as sensitivity (Sens), specificity (Spec), accuracy (Acc) and false rate of discovery (FDR) as proposed by Sokolova *et al.*, [187]:

- **Accuracy** is the overall effectiveness of a classifier (Equation C.1);



		Actual Classes	
		Class Positive	Class Negative
Predicted Outcomes	Class Positive	*tp*	*fp*
	Class Negative	*fn*	*tn*

Figure C.1: Confusion matrix from binary classification.

- **Specificity** measures how effectively a classifier identifies a negative labels (Equation C.2);
- **Sensitivity** evaluates the effectiveness of a classifier to identify positive labels (Equation C.3);
- **False discovery rate** is the proportion of false positives, as shown in Equation C.4.

$$Acc = \frac{\sum tp + \sum tn}{\sum(tp + fn + fp + tn)} \tag{C.1}$$

$$Spec = \frac{\sum tn}{\sum(fp + tn)} \tag{C.2}$$

$$Sens = \frac{\sum tp}{\sum(tp + fn)} \tag{C.3}$$

$$\mathrm{FDR} = \frac{\sum fp}{\sum(tn + fp)} \tag{C.4}$$

The receiver operating curves (ROC) and the area under the curve (AUC) are computed according the formulation for ROC graphs proposed by Fawcett *et al.*, [188]. To evaluate the model performance regarding the probabilistic predictions, I also compute the **logarithmic loss** (equation C.5), which takes into account the uncertainty of the predictions based on how much it varies from the actual label:

$$\log(\mathcal{L}) = -\big(y \log(p_{\hat{y}}) + (1 - y)\log(1 - p_{\hat{y}})\big) \tag{C.5}$$

where y is the observed label and $p_{\hat{y}}$ is the probability of the $\mathcal{C} = +1$.

C.3 Multi-class Classification

Due to the multi-class paradigm and the unbalanced nature of the data, I consider both macro-averaging measures, generalised from the measures used to assess the performance of binary classification evaluation to evaluate the accuracy of the predictive labels obtained from the multi-class model [187]. The averaging accuracy (Equation C.6), macro-recall (Equation C.7), macro-precision (Equation C.8) are computed for model evaluation. The predicted label for each subject is obtained based on the class with the highest probability among all the possible classes.

$$Acc = \frac{\sum_{c=1}^{\mathcal{C}} \left(\frac{tp_c + tn_c}{tp_c + tn_c + fp_c + fn_c} \right)}{\mathcal{C}} \tag{C.6}$$

$$R_M = \frac{\sum_{c=1}^{\mathcal{C}} \left(\frac{tp_c}{tp_c + fn_c} \right)}{\mathcal{C}} \tag{C.7}$$

$$P_M = \frac{\sum_{c=1}^{\mathcal{C}} \left(\frac{tp_c}{tp_c + fp_c} \right)}{\mathcal{C}} \tag{C.8}$$

I also compute the **multiclass logarithmic Loss** (equation C.9):

$$\log(\mathcal{L}) = -\sum_{c=1}^{\mathcal{C}} \log(p_{\hat{y},c}) \tag{C.9}$$

where $p_{\hat{y},c}$ is the probability of observation of class $\mathcal{C}$.

Appendix D

List of brain regions

Brain Regions	**Tissue type**	
Cerebellar Vermal Lobules I-V	**GM**	
Cerebellar Vermal Lobules VI-VII	**GM**	
Cerebellar Vermal Lobules VIII-X	**GM**	
Left ACgG anterior cingulate gyrus	**GM**	
Left AIns anterior insula	**GM**	
Left AOrG anterior orbital gyrus	**GM**	
Left Accumbens Area	**WM**	**Deep GM**
Left Amygdala	**GM**	**Deep GM**
Left AnG angular gyrus	**GM**	
Left Basal Forebrain	**Deep GM**	
Left Caudate	**WM**	**Deep GM**
Left Cerebellum Exterior	**GM**	
Left Claustrum	**Deep GM**	
Left Cun cuneus	**GM**	
Left Ent entorhinal area	**GM**	
Left FO frontal operculum	**GM**	
Left FRP frontal pole	**GM**	
Left FuG fusiform gyrus	**GM**	
Left GRe gyrus rectus	**GM**	
Left Hippocampus	**GM**	**Deep GM**
Left IOG inferior occipital gyrus	**GM**	
Left ITG inferior temporal gyrus	**GM**	
Left LOrG lateral orbital gyrus	**GM**	
Left Lesion	**Deep GM**	
Left LiG lingual gyrus	**GM**	
Left MCgG middle cingulate gyrus	**GM**	
Left MFC medial frontal cortex	**GM**	
Left MFG middle frontal gyrus	**GM**	
Left MOG middle occipital gyrus	**GM**	

Table D.1 continued from previous page

Left MOrG medial orbital gyrus	**GM**	
Left MPoG postcentral gyrus medial segment	**GM**	
Left MPrG precentral gyrus medial segment	**GM**	
Left MSFG superior frontal gyrus medial segment	**GM**	
Left MTG middle temporal gyrus	**GM**	
Left OCP occipital pole	**GM**	
Left OFuG occipital fusiform gyrus	**GM**	
Left OpIFG opercular part of the inferior frontal gyrus	**GM**	
Left OrIFG orbital part of the inferior frontal gyrus	**GM**	
Left PCgG posterior cingulate gyrus	**GM**	
Left PCu precuneus	**GM**	
Left PHG parahippocampal gyrus	**GM**	
Left PIns posterior insula	**GM**	
Left PO parietal operculum	**GM**	
Left POrG posterior orbital gyrus	**GM**	
Left PP planum polare	**GM**	
Left PT planum temporale	**GM**	
Left Pallidum	**WM**	**Deep GM**
Left PoG postcentral gyrus	**GM**	
Left PrG precentral gyrus	**GM**	
Left Putamen	**WM**	**Deep GM**
Left SCA subcallosal area	**GM**	
Left SFG superior frontal gyrus	**GM**	
Left SMC supplementary motor cortex	**GM**	
Left SMG supramarginal gyrus	**GM**	
Left SOG superior occipital gyrus	**GM**	
Left SPL superior parietal lobule	**GM**	
Left STG superior temporal gyrus	**GM**	
Left TMP temporal pole	**GM**	
Left TTG transverse temporal gyrus	**GM**	
Left Thalamus Proper	**WM**	**Deep GM**
Left TrIFG triangular part of the inferior frontal gyrus	**GM**	
Left Ventral DC	**WM**	**Deep GM**
Left Ventricular Lining	**Deep GM**	
Left vessel	**Deep GM**	
Optic Chiasm	**Deep GM**	
Right ACgG anterior cingulate gyrus	**GM**	
Right AIns anterior insula	**GM**	
Right AOrG anterior orbital gyrus	**GM**	
Right Accumbens Area	**WM**	**Deep GM**
Right Amygdala	**GM**	**Deep GM**

Table D.1 continued from previous page

Right AnG angular gyrus	**GM**	
Right Basal Forebrain	**Deep GM**	
Right CO central operculum	**GM**	
Right Calc calcarine cortex	**GM**	
Right Caudate	**WM**	**Deep GM**
Right Cerebellum Exterior	**GM**	
Right Claustrum	**Deep GM**	
Right Cun cuneus	**GM**	
Right Ent entorhinal area	**GM**	
Right FO frontal operculum	**GM**	
Right FRP frontal pole	**GM**	
Right FuG fusiform gyrus	**GM**	
Right GRe gyrus rectus	**GM**	
Right Hippocampus	**GM**	**Deep GM**
Right IOG inferior occipital gyrus	**GM**	
Right ITG inferior temporal gyrus	**GM**	
Right LOrG lateral orbital gyrus	**GM**	
Right Lesion	**Deep GM**	
Right LiG lingual gyrus	**GM**	
Right MCgG middle cingulate gyrus	**GM**	
Right MFC medial frontal cortex	**GM**	
Right MFG middle frontal gyrus	**GM**	
Right MOG middle occipital gyrus	**GM**	
Right MOrG medial orbital gyrus	**GM**	
Right MPoG postcentral gyrus medial segment	**GM**	
Right MPrG precentral gyrus medial segment	**GM**	
Right MSFG superior frontal gyrus medial segment	**GM**	
Right MTG middle temporal gyrus	**GM**	
Right OCP occipital pole	**GM**	
Right OFuG occipital fusiform gyrus	**GM**	
Right OpIFG opercular part of the inferior frontal gyrus	**GM**	
Right OrIFG orbital part of the inferior frontal gyrus	**GM**	
Right PCgG posterior cingulate gyrus	**GM**	
Right PCu precuneus	**GM**	
Right PHG parahippocampal gyrus	**GM**	
Right PIns posterior insula	**GM**	
Right PO parietal operculum	**GM**	
Right POrG posterior orbital gyrus	**GM**	
Right PP planum polare	**GM**	
Right PT planum temporale	**GM**	
Right Pallidum	**WM**	**Deep GM**

Table D.1 continued from previous page

Right PoG postcentral gyrus	**GM**	
Right PrG precentral gyrus	**GM**	
Right Putamen	**WM**	**Deep GM**
Right SCA subcallosal area	**GM**	
Right SFG superior frontal gyrus	**GM**	
Right SMC supplementary motor cortex	**GM**	
Right SMG supramarginal gyrus	**GM**	
Right SOG superior occipital gyrus	**GM**	
Right SPL superior parietal lobule	**GM**	
Right STG superior temporal gyrus	**GM**	
Right TMP temporal pole	**GM**	
Right TTG transverse temporal gyrus	**GM**	
Right Thalamus Proper	**WM**	**Deep GM**
Right TrIFG triangular part of the inferior frontal gyrus	**GM**	
Right Ventral DC	**WM**	**Deep GM**
Right Ventricular Lining	**Deep GM**	
Right vessel	**Deep GM**	

Bibliography

[1] Robert G Will. Acquired prion disease : iatrogenic CJD , variant CJD , kuru. *British Medical Bulletin*, 66:255–265, 2003.

[2] Piero Parchi, Maura Cescatti, Silvio Notari, Walter J Schulz-schaeffer, Sabina Capellari, Armin Giese, Wen-quan Zou, Hans Kretzschmar, Bernardino Ghetti, and Paul Brown. Agent strain variation in human prion disease: insights from a molecular and pathological review of the National Institutes of Health series of experimentally transmitted disease. *Brain : a journal of neurology*, 133:3030–3042, 2010.

[3] B Caughey, G J Raymond, D Ernst, and R E Race. N-terminal truncation of the scrapie-associated form of PrP by lysosomal protease(s): implications regarding the site of conversion of PrP to the protease-resistant state. *Journal of virology*, 65(12):6597–6603, 1991.

[4] Keh-ming Pan, Michael Baldwin, Jack Nguyen, Maria Gasset, Ana Serban, Darlene Groth, Ingrid Mehlhorn, Ziwei Huang, Robert J Fletterick, Fred E Cohenu, and Stanley B Prusiner. Conversion of a-helices into ,f-sheets features in the formation of the scrapie prion proteins (ceilular prion protein/protein conformation/secondary structure/amyloid/post-translational modification). In *Proc. Natl. Acad. Sci. USA*, volume 90, pages 10962–10966, 1993.

[5] Olivia May. Prion Proteins: Unfolding Infectious Potential, 2008.

[6] Richard T Johnson. Prion diseases. *Lancet Neurology*, 4:635–642, 2005.

[7] Simon Mead. Prion disease genetics. *European Journal of Human Genetics*, 14:273–281, 2006.

[8] Durrenajaf Siddique, Harpreet Hyare, Stephen Wroe, Thomas Webb, Rebecca MacFarlane, Peter Rudge, John Collinge, Caroline Powell, Sebastian Brand-

ner, Po Wah So, Sarah Walker, Simon Mead, Tarek Yousry, and John S Thornton. Magnetization transfer ratio may be a surrogate of spongiform change in human prion diseases. *Brain*, 133(10):3058–3068, 2010.

[9] M. Pocchiari, M. Puopolo, E. A. Croes, H. Budka, E. Gelpi, S. Collins, V. Lewis, T. Sutcliffe, A. Guilivi, N. Delasnerie-Laupretre, J. P. Brandel, A. Alperovitch, I. Zerr, S. Poser, H. A. Kretzschmar, A. Ladogana, I. Rietvald, E. Mitrova, P. Martinez-Martin, J. De Pedro-Cuesta, M. Glatzel, A. Aguzzi, S. Cooper, J. Mackenzie, C. M. Van Duijn, and R. G. Will. Predictors of survival in sporadic Creutzfeldt-Jakob disease and other human transmissible spongiform encephalopathies. *Brain*, 2004.

[10] Stephen DeArmond. Degenerative Diseases and other Dementias, 2017.

[11] Medical Research. MRC Prion Unit, 2017.

[12] Maren Breithaupt, Carlos Romero, Kai Kallenberg, Christian Begue, Pascual Sanchez-Juan, Sabina Eigenbrod, Hans Kretzschmar, Gabi Schelzke, Eduardo Meichtry, Analia Taratuto, and Inga Zerr. Magnetic resonance imaging in E200K and V210I mutations of the prion protein gene. *Alzheimer Disease and Associated Disorders*, 27(1):87–90, 1 2013.

[13] A LeBlanc R Medori, HJ Tritschler. Fatal Familial Insomnia, a Prion Disease with a mutation at codon 178 of the prion gene. *The New England Journal of Medicine*, 326(7):444–449, 1992.

[14] L G Goldfarb, P Brown, E Mitrovà, L Cervenáková, L Goldin, A D Korczyn, J Chapman, S Galvez, L Cartier, R Rubenstein, and D C Gajdusek. Creutzfeldt-Jacob disease associated with the PRNP codon 200LYS mutation: An analysis of 45 families. *European Journal of Epidemiology*, 7(5):477–486, 1991.

[15] K Alner, H Hyare, S Mead, P Rudge, S Wroe, J D Rohrer, G R Ridgway, S Ourselin, M Clarkson, H Hunt, N C Fox, T Webb, J Collinge, and L Cipolotti. Distinct neuropsychological profiles correspond to distribution of cortical thinning in inherited prion disease caused by insertional mutation. *Journal of Neurology Neurosurgury and Psychiatry*, 83(1):109–114, 2012.

[16] Laura Pirisinu, Michele A. Di Bari, Claudia D'Agostino, Stefano Marcon, Geraldina Riccardi, Anna Poleggi, Mark L. Cohen, Brian S. Appleby, Pierluigi Gambetti, Bernardino Ghetti, Umberto Agrimi, and Romolo Nonno. Gerstmann-Sträussler-Scheinker disease subtypes efficiently transmit in bank voles as genuine prion diseases. *Scientific Reports*, 6, 2 2016.

[17] R G Will, J W Ironside, M Zeidler, K Estibeiro, S N Cousens, P G Smith, A Alperovitch, S Poser, M Pocchiari, and A Hofman. A new variant of Creutzfeldt-Jakob disease in the UK. *The Lancet*, 347(9006):921–925, 1996.

[18] Barry M Bradford, Pedro Piccardo, James W Ironside, and Neil A Mabbott. Human prion diseases and the risk of their transmission during anatomical dissection. *Clinical Anatomy*, 27(6):821–832, 2014.

[19] Diego Cardoso Fragoso, Augusto Lio da Mota Gonçalves Filho, Felipe Torres Pacheco, Bernardo Rodi Barros, Ingrid Aguiar Littig, Renato Hoffmann Nunes, Antnio Carlos Martins Maia Júnior, and Antonio J da Rocha. Imaging of Creutzfeldt-Jakob Disease: Imaging Patterns and Their Differential Diagnosis. *RadioGraphics*, 37(1):234–257, 2017.

[20] S Mead, M Ranopa, G S Gopalakrishnan, A G B Thompson, P Rudge, S Wroe, A Kennedy, F Hudson, A MacKay, J H Darbyshire, J Collinge, and A S Walker. PRION-1 scales analysis supports use of functional outcome measures in prion disease. *Neurology*, 77(18):1674–1683, 2011.

[21] Andrew G B Thompson, Jessica Lowe, Zoe Fox, Ana Lukic, Marie-claire Porter, Liz Ford, Michele Gorham, Gosala S Gopalakrishnan, Peter Rudge, A Sarah Walker, John Collinge, and Simon Mead. The Medical Research Council Prion Disease Rating Scale: a new outcome measure for prion disease therapeutic trials developed and validated using systematic observational studies. *Brain : a journal of neurology*, 136:1116–1127, 2013.

[22] Marc Manix, Piyush Kalakoti, Miriam Henry, Jai Thakur, Richard Menger, Bharat Guthikonda, and Anil Nanda. Creutzfeldt-Jakob disease: updated diagnostic criteria, treatment algorithm, and the utility of brain biopsy. *Neurosurgical Focus*, 39(November):1–11, 2015.

[23] M K Sandberg, H Al-Doujaily, B Sharps, A R Clarke, and J Collinge. Prion

propagation and toxicity in vivo occur in two distinct mechanistic phases. *Nature*, 470(7335):540–542, 2011.

[24] Andreas Schroter, Inga Zerr, Karten Henkel, Henriette J Tschampa, Michael Finkenstaedt, and Sigrid Poser. Magnetic Resonance Imaging in the Clinical Diagnosis of Creutzfeldt-Jakob Disease. *Journal of American Medical Association*, 57:1751–1757, 2000.

[25] Inga Zerr and Sigrid Poser. Clinical diagnosis and differential diagnosis of CJD and vCJD With special emphasis on laboratory tests. *Acta Pathologica, Microbiologica et Immunologica Scandinavica*, 110:88–98, 2002.

[26] Federico Caobelli, Milena Cobelli, Claudio Pizzocaro, Marco Pavia, Silvia Magnaldi, and Ugo Paolo Guerra. The Role of Neuroimaging in Evaluating Patients Affected by Creutzfeldt-Jakob Disease: A Systematic Review of the Literature. *Journal of neuroimaging : official journal of the American Society of Neuroimaging*, 6:1–12, 2014.

[27] D A Collie, R. J. Sellar, M. Zeidler, A. C. F. Colchester, R. Knight, and R. G. Will. MRI of Creutzfeldt-Jakob Disease : Imaging Features and Recommended MRI Protocol. *Clinical Radiology*, 56:726–739, 2001.

[28] D P Barboriak and J M Provenzale. MR diagnosis of Creutzfeldt-Jakob disease : signifiance of high signal intensity of the basal ganglia. *America Journal of Radiology*, 162:137–140, 1994.

[29] A Uemura, T O'uchi, T Sakamoto, and N Yashiro. High signal of the striatum in sporadic Creutzfeldt-Jakob disease: Sequential change on T2-weighted MRI. *Neuroradiology*, 44(4):314–318, 2002.

[30] Eiji Matsusue, Toshibumi Kinoshita, Shuji Sugihara, Shinya Fujii, Toshihide Ogawa, and Eisaku Ohama. White matter lesions in panencephalopathic type of Creutzfeldt-Jakob disease: MR imaging and pathologic correlations. *American Journal of Neuroradiology*, 25(6):910–918, 2004.

[31] E De Vita, B H Ridha, N C Fox, J S Thornton, and H R Jager. Voxel-based analysis of high- and standard b-value diffusion weighted imaging , and voxel based morphometry , in Alzheimer disease. In *ISMRM: Clinical Application of Diffusion Tensor Imaging III*, Quebec, 2011.

[32] Enrico De Vita, Harpreet Hyare, Gerard Ridgway, Marie-claire Porter, Andrew Thompson, Chris Carswell, Ana Lukic, Rolf Jager, Diana Caine, Tarek Yousry, John Collinge, Simon Mead, and John Thornton. Longitudinal VBM of regional progression in human prion disease. In *ISMRM*, page 3961, 2013.

[33] Enrico De Vita, Gerard R Ridgway, Mark J White, Marie-Claire Porter, Diana Caine, Peter Rudge, John Collinge, Tarek A Yousry, Hans Rolf Jager, Simon Mead, John S Thornton, and Harpreet Hyare. Neuroanatomical correlates of prion disease progression - a 3T longitudinal voxel-based morphometry study. *NeuroImage: Clinical*, 13:89–96, 2017.

[34] Enrico De Vita, Marie-claire Porter, Ivor Simpson, Zoe Fox, Gerard Ridgway, Sebastien Ourselin, Peter Rudge, Diana Caine, Rolf Jager, Tarek Yousry, John Collinge, Simon Mead, Harpreet Hyare, and John S Thornton. Cross Sectional and Longitudinal Magnetisation transfer Ratio in Prion disease at 3 Tesla Introduction . Human prion diseases are progressive and uniformly fatal neurodegenerative disorders caused by abnormally folded. In *International Society for Magnetic Resonance in Medicine*, volume 133, page 4272, 2015.

[35] Enrico De Vita, Harpreet Hyare, Gerard R Ridgway, Nicolas Toussaint, Ivor Simpson, Peter Rudge, Diana Caine, Rolf H Jager, Tarek Yousry, and John Collinge. Effectiveness of DWI acquisition at different spatial resolution to reveal prion disease pathology. In *ISMR*, volume 34, page 7902, 2013.

[36] Takaki Murata, Yusei Shiga, Shuichi Higano, Shoki Takahashi, and Shunji Mugikura. Conspicuity and evolution of lesions in Creutzfeldt-Jakob disease at diffusion-weighted imaging. *American Journal of Neuroradiology*, 23(7):1164–1172, 2002.

[37] K Kallenberg, W J Schulz-Schaeffer, U Jastrow, S Poser, B Meissner, H J Tschampa, I Zerr, and M Knauth. Creutzfeldt-Jakob disease: comparative analysis of MR imaging sequences. *AJNR. American journal of neuroradiology*, 27(7):1459–62, 8 2006.

[38] Geoffrey S Young, Michael D Geschwind, Nancy J Fischbein, Jennifer L Martindale, Roland G Henry, Songling Liu, Ying Lu, Stephen Wong, Hong Liu, Bruce L Miller, and William P Dillon. Diffusion-Weighted and Fluid-Attenuated Inversion Recovery Imaging in Creutzfeldt-Jakob Disease : High

Sensitivity and Specificity for Diagnosis. *American Journal of Neuroradiology American Society of Neuroradiology*, 26(July):1551–1562, 2005.

[39] Y Shiga, K Miyazawa, S Sato, R Fukushima, S Shibuya, Y Sato, H Konno, K Doh-ura, S Mugikura, H Tamura, S Higano, S Takahashi, and Y Itoyama. Diffusion-weighted MRI abnormalities as an early diagnostic marker for Creutzfeldt-Jakob disease. *Neurology*, 63(3):443–449, 2004.

[40] Philippe Demaerel, Raf Sciot, Wim Robberecht, Ren Dom, Dirk Vandermeulen, Frederik Maes, and Guido Wilms. Accuracy of diffusion-weighted MR imaging in the diagnosis of sporadic Creutzfeldt-Jakob disease. *Journal of Neurology*, 250(2):222–225, 2003.

[41] Henriette J Tschampa, Petra Mürtz, Sebastian Flacke, Sebastian Paus, Hans H Schild, and Horst Urbach. Thalamic involvement in sporadic Creutzfeldt-Jakob disease: a diffusion-weighted MR imaging study. *AJNR American journal of neuroradiology*, 24(5):908–915, 2003.

[42] Donald A Collie, David M Summers, Robin J Sellar, James W Ironside, Sarah Cooper, Martin Zeidler, Richard Knight, and Robert G Will. Diagnosing variant Creutzfeldt-Jakob disease with the pulvinar sign: MR imaging findings in 86 neuropathologically confirmed cases. *American Journal of Neuroradiology*, 24(8):1560–1569, 2003.

[43] Ryutarou Ukisu, Tamio Kushihashi, Takashi Kitanosono, Hidefumi Fujisawa, Hiroki Takenaka, Yoshimitsu Ohgiya, Takehiko Gokan, Hirotsugu Munechika, Kitanosono T et al. Ukisu R, Kushihashi T, and Received. Serial Diffusion-Weighted MRI of Creutzfeldt-Jakob Disease. *American Journal of Neuroradiology*, 184:560–566, 2005.

[44] Henriette J Tschampa, K Kallenberg, H A Kretzschmar, B Meissner, M Knauth, H Urbach, and I Zerr. Pattern of cortical changes in sporadic Creutzfeldt-Jakob disease. *American Journal of Neuroradiology*, 28(6):1114–1118, 2007.

[45] Bettina Meissner, K Kallenberg, P Sanchez-Juan, A Krasnianski, U Heinemann, D Varges, M Knauth, and I Zerr. Isolated cortical signal increase on MT imaging as a frequent lesion pattern in sporadic Creutzfeldt-Jakob Disease. *American Journal of Neuroradiology*, 29(8):1519–1524, 2008.

[46] Robert K Fulbright, C Hoffmann, H Lee, A Pozamantir, J Chapman, and I Prohovnik. MR imaging of familial Creutzfeldt-Jakob disease: A blinded and controlled study. *American Journal of Neuroradiology*, 29(9):1638–1643, 2008.

[47] Taro Shimono, Takahiro Tsuboyama, Masatomo Kuwabara, Sung-woon Im, Yukinobu Yagyu, Izumi Imaoka, Ryuichiro Ashikaga, Makoto Hosono, and Takamichi Murakami. Discordance of motion artifacts on magnetic resonance imaging in Creutzfeldt-Jakob disease: comparison of diffusion-weighted and conventional imaging sequences. *Radiation Medicine*, 26(3):151–155, 2008.

[48] A Krasnianski, K Kallenberg, D A Collie, B Meissner, W J Schulz-Schaeffer, U Heinemann, D Varges, D M Summers, H A Kretzschmar, T Talbot, R G Will, and I Zerr. MRI the classical MM1 and the atypical MV2 subtypes of sporadic CJD: An inter-observer agreement study. *European Journal of Neurology*, 15(8):762–771, 2008.

[49] D N Manners, P Parchi, C Tonon, S Capellari, R Strammiello, C Testa, G Tani, E Malucelli, C Spagnolo, P Cortelli, P Montagna, R Lodi, and B Barbiroli. Pathologic correlates of diffusion MRI changes in Creutzfeldt-Jakob disease. *Neurology*, 72(16):1425–1431, 2009.

[50] Harpreet Hyare, J Thornton, J Stevens, S Mead, P Rudge, J Collinge, T A Yousry, and H R Jäger. High-b-value diffusion MR imaging and basal nuclei apparent diffusion coefficient measurements in variant and sporadic Creutzfeldt-Jakob disease. *American Journal of Neuroradiology*, 31(3):521–526, 2010.

[51] Sabrina D Talbott, Brian M Plato, Ronald J Sattenberg, John Parker, and Jens O Heidenreich. Cortical restricted diffusion as the predominant MRI finding in sporadic Creutzfeldt-Jakob disease. *Acta radiologica (Stockholm, Sweden : 1987)*, 52(3):336–339, 2011.

[52] P Vitali, E MacCagnano, E Caverzasi, R G Henry, A Haman, C Torres-Chae, D Y Johnson, B L Miller, and M D Geschwind. Diffusion-weighted MRI hyperintensity patterns differentiate CJD from other rapid dementias. *Neurology*, 76(20):1711–1719, 2011.

[53] H Hyare, S Wroe, D Siddique, T Webb, N C Fox, J Stevens, J Collinge,

T Yousry, and J S Thornton. Brain-water diffusion coefficients reflect the severity of inherited prion disease. *Neurology*, 74(8):658–665, 2010.

[54] H Hyare, E De Vita, C Carswell, A Thompson, A Lukic, T Yousry, P Rudge, S Mead, J Collinge, and J Thornton. Cerebral diffusion tensor imaging in prion diseases : voxelwise analysis and comparison with VBM. In *ISMRM: Clinical Application of Diffusion Tensor Imaging II*, Quebec, 2011.

[55] Harpreet Hyare, Enrico De Vita, Marie-claire Porter, Ivor Simpson, Ged Ridgway, Simon Mead, Peter Rudge, John Collinge, and John Thornton. Putamen radial diffusivity is an independent predictor of prion disease severity. In *ISMR*, volume 57, page 7334, 2013.

[56] E De Vita, G R Ridgway, R I Scahill, D Caine, P Rudge, T A Yousry, S Mead, J Collinge, H R Jäger, J S Thornton, and H Hyare. Multiparameter MR imaging in the 6-OPRI variant of inherited prion disease. *American Journal of Neuroradiology*, 34:1723–1730, 2013.

[57] Leo H. Wang, Robert C. Bucelli, Erica Patrick, Dhanashree Rajderkar, Enrique Alvarez, Miranda M. Lim, Gabriela Debruin, Victoria Sharma, Sonika Dahiya, Robert E. Schmidt, Tammie S. Benzinger, Beth A. Ward, and Beau M. Ances. Role of magnetic resonance imaging, cerebrospinal fluid, and electroencephalogram in diagnosis of sporadic Creutzfeldt-Jakob disease. *Journal of Neurology*, 260(2):498–506, 2 2013.

[58] F. Clarençon, F. Gutman, C. Giannesini, A. Pénicaud, D. Galanaud, K. Kerrou, B. Marro, and J. N. Talbot. MRI and FDG PET/CT findings in a case of probable Heidenhain variant Creutzfeldt-Jakob disease. *Journal of Neuroradiology*, 2008.

[59] Andrew G B Thompson, Jessica Lowe, Zoe Fox, Ana Lukic, Marie Claire Porter, Liz Ford, Michele Gorham, Gosala S Gopalakrishnan, Peter Rudge, A Sarah Walker, John Collinge, and Simon Mead. The medical research council prion disease rating scale: A new outcome measure for prion disease therapeutic trials developed and validated using systematic observational studies. *Brain*, 136(4):1116–1127, 2013.

[60] Florence I Mahoney and Dorothea W Barthel. Functional Evaluation: The Barthel Intex. *Maryland State Medical Journal*, 14:56–61, 1965.

[61] Graham Teasdale and Bryan Jennett. Assessment of Coma and Impaired Consciousness. A Practical Scale. *The Lancet*, 304(7872):81–84, 1974.

[62] Saima Rathore, Mohamad Habes, Muhammad Aksam Iftikhar, Amanda Shacklett, and Christos Davatzikos. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 2017.

[63] Stefan Klöppel, Ahmed Abdulkadir, Clifford R. Jack, Nikolaos Koutsouleris, Janaina Mourão-Miranda, and Prashanthi Vemuri. Diagnostic neuroimaging across diseases, 2012.

[64] John Ashburner. VBM Tutorial. Technical report, 2010.

[65] Christos Davatzikos, Ahmet Genc, Dongrong Xu, and Susan M. Resnick. Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage*, 2001.

[66] Nick C Fox and MRCP A Peter Freeborough. Brain Atrophy Progression Measured from Registered Serial MRI: Validation and Application to Alzheimer's Disease. *JMRI*, 7(6), 1997.

[67] Bradford C. Dickerson, Akram Bakkour, David H. Salat, Eric Feczko, Jenni Pacheco, Douglas N. Greve, Fran Grodstein, Christopher I. Wright, Deborah Blacker, H. Diana Rosas, Reisa A. Sperling, Alireza Atri, John H. Growdon, Bradley T. Hyman, John C. Morris, Bruce Fischl, and Randy L. Buckner. The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral Cortex*, 19(3):497–510, 3 2009.

[68] Brian Patenaude, Stephen M. Smith, David N. Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 2011.

[69] M Jorge Cardoso, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C Fox, and Sebastien Ourselin. STEPS: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation. *Medical image analysis*, 17(6):671–684, 2013.

[70] Owen T. Carmichael, Howard A. Aizenstein, Simon W. Davis, James T. Becker, Paul M. Thompson, Carolyn Cidis Meltzer, and Yanxi Liu. Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 2005.

[71] Liang Zhan, Jiayu Zhou, Yalin Wang, Yan Jin, Neda Jahanshad, Gautam Prasad, Talia M. Nir, Cassandra D. Leonardo, Jieping Ye, and Paul M. Thompson. Comparison of nine tractography algorithms for detecting abnormal structural brain networks in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7(APR), 2015.

[72] Chong Yaw Wee, Pew Thian Yap, Wenbin Li, Kevin Denny, Jeffrey N. Browndyke, Guy G. Potter, Kathleen A. Welsh-Bohmer, Lihong Wang, and Dinggang Shen. Enriched white matter connectivity networks for accurate identification of MCI patients. *NeuroImage*, 2011.

[73] Martin Dyrba, Michel Grothe, Thomas Kirste, and Stefan J Teipel. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Human Brain Mapping*, 36(6):2118–2131, 2015.

[74] Hong Ying Zhang, Shi Jie Wang, Bin Liu, Zhan Long Ma, Ming Yang, Zhi Jun Zhang, and Gao Jun Teng. Resting brain connectivity: Changes during the progress of Alzheimer disease. *Radiology*, 256(2):598–606, 2010.

[75] D Araújo, A Dria Neto, A Martins, and J Melo. Comparative study on dimension reduction techniques for cluster analysis of microarray data. In *The 2011 International Joint Conference on Neural Networks*, pages 1835–1842, 2011.

[76] I. T. Jolliffe. *Principal Component Analysis.* Springer, New York.

[77] Xinzheng Xu, Tianming Liang, Jiong Zhu, Dong Zheng, and Tongfeng Sun. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, 2019.

[78] R. Hamer and F. Young. *Multidimensional Scaling.* Psychology Press, New York, 1987.

[79] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York.

[80] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319, 12 2000.

[81] H Greenspan, B van Ginneken, and R M Summers. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 5 2016.

[82] Yann Lecun and Yoshua Bengio. Convolutional Networks for Images, Speech, and Time-Series. In Michael A. Arbib, editor, *The handbook of brain theory and neural networks*, pages Arbib, Michael A. MIT Press, 1998.

[83] Yvan Saeys, Iaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics, 10 2007.

[84] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers and Electrical Engineering*, 2014.

[85] Saima Rathore, Mohamad Habes, Muhammad Aksam Iftikhar, Amanda Shacklett, and Christos Davatzikos. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 2017.

[86] Isabelle Guyon and Andr Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[87] Kelly Peterson. *Personalized Gaussian Process-Based Machine Learning Models for Forecasting Alzheimer's Disease Progression.* PhD thesis, MIT, 2019.

[88] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 8 2005.

[89] Chong Yaw Wee, Pew Thian Yap, and Dinggang Shen. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. *Human Brain Mapping*, 34(12):3411–3425, 12 2013.

[90] Xiaoying Tang, Yuanyuan Qin, Jiong Wu, Min Zhang, Wenzhen Zhu, and Michael I. Miller. Shape and diffusion tensor imaging based integrative analysis of the hippocampus and the amygdala in Alzheimer's disease. *Magnetic Resonance Imaging*, 2016.

[91] Christos Davatzikos, Yong Fan, Xiaoying Wu, Dinggang Shen, and Susan M. Resnick. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of Aging*, 2008.

[92] Yong Fan, Dinggang Shen, Ruben C. Gur, Raquel E. Gur, and Christos Davatzikos. COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, 1 2007.

[93] Josien P.W. Pluim, J. B.A.Antoine Maintz, and Max A. Viergever. Mutual-information-based registration of medical images: A survey, 8 2003.

[94] Igor O. Korolev, Laura L. Symonds, and Andrea C. Bozoki. Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLoS ONE*, 11(2), 2 2016.

[95] Hongmei Mi, Caroline Petitjean, Bernard Dubray, Pierre Vera, and Su Ruan. Robust feature selection to predict tumor treatment outcome. *Artificial Intelligence in Medicine*, 64(3):195–204, 2015.

[96] Iman Beheshti, Hasan Demirel, and Hiroshi Matsuda. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in Biology and Medicine*, 2017.

[97] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422, 2002.

[98] E Romero and J M Sopena. Performing Feature Selection With Multilayer Perceptrons. *IEEE Transactions on Neural Networks*, 19(3):431–441, 2008.

[99] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer, 2nd edition edition, 2008.

[100] Probal Chaudhuri, Anil K. Ghosh, and Hannu Oja. Classification based on hybridization of parametric and nonparametric classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1153–1164, 2009.

[101] Richard O. Duda, David G. Stork, and Peter E.Hart. *Pattern Classification.* Wiley-Interscience, New York, 2000.

[102] Annette J Dobson. *An introduction to generalized linear models.* 2002.

[103] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* New York, 8 2006.

[104] Anil Rao, Ying Lee, Achim Gass, and Andreas Monsch. Classification of Alzheimer's Disease from structural MRI using sparse logistic regression with optional spatial regularization. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2011.

[105] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, and Jussi Tohka. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, 104:398–412, 2015.

[106] Sven A. Forner, Leonel T. Takada, Brianne M. Bettcher, Iryna V. Lobach, Maria Carmela Tartaglia, Charles Torres-Chae, Aissatou Haman, Julie Thai, Paolo Vitali, John Neuhaus, Alan Bostrom, Bruce L. Miller, Howard J. Rosen, and Michael D. Geschwind. Comparing CSF biomarkers and brain MRI in the diagnosis of sporadic Creutzfeldt-Jakob disease. *Neurology: Clinical Practice*, 5(2), 2015.

[107] Yun Wang, Chenxiao Xu, Ji Hwan Park, Seonjoo Lee, Yaakov Stern, Shinjae Yoo, Jong Hun Kim, Hyoung Seop Kim, and Jiook Cha. Diagnosis and prognosis of Alzheimer's disease using brain morphometry and white matter connectomes. *NeuroImage: Clinical*, 2019.

[108] Christopher J C Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Technical report, 1998.

[109] Michael E.Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, (1):211–244, 2001.

[110] Peter M. Rasmussen, Lars K. Hansen, Kristoffer H. Madsen, Nathan W. Churchill, and Stephen C. Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, 2012.

[111] Jinfeng Zhuang, Ivor W. Tsang, and Steven C.H. Hoi. A family of simple non-parametric Kernel learning algorithms. *Journal of Machine Learning Research*, 12:1313–1347, 2011.

[112] Carl Rasmussen and Christopher Williams. *Gaussian processes for machine learning.*, volume 14. MIT Press, 2004.

[113] David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B Tenenbaum, and Zoubin Ghahramani. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. *Advances in Neural Information Processing Systems*, 28, 2013.

[114] David Duvenaud, Hannes Nickisch, and Carl Edward Rasmussen. Additive Gaussian Processes. *Advances in Neural Information Processing Systems*, 24:1–9, 2011.

[115] Liane S. Canas and, Benjamin C. Yvernault, David M. Cash, Erika Molteni, Tom Veale, Tammie L. Benzinger, Sebastien Ourselin, Simon Mead, and Marc Modat. Gaussian Processes with optimal kernel construction for neuro-degenerative clinical onset prediction. In *Medical Imaging 2018: Computer-Aided Diagnosis of SPIE*, volume 10575, Houston, 2018.

[116] Michael E Tipping. The Relevance Vector Machine. In *Neural Information Processing Systems*, 2000.

[117] Janaina Mourão-Miranda, Karl J. Friston, and Michael Brammer. Dynamic discrimination analysis: A spatial-temporal SVM. *NeuroImage*, 2007.

[118] Stefan Klöppel, Cynthia M. Stonnington, Carlton Chu, Bogdan Draganski, Rachael I. Scahill, Jonathan D. Rohrer, Nick C. Fox, Clifford R. Jack, John Ashburner, and Richard S.J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 3 2008.

[119] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, and Dinggang Shen. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 2011.

[120] Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2):895–907, 2012.

[121] J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. In *Information Sciences*, volume 237, pages 59–72, 7 2013.

[122] Olfa Ben Ahmed, Jenny Benois-Pineau, Michelle Allard, Gwnalle Catheline, and Chokri Ben Amar. Recognition of Alzheimer's disease and Mild Cognitive Impairment with multimodal image-derived biomarkers and Multiple Kernel Learning. *Neurocomputing*, 2017.

[123] Katja Franke, Gabriel Ziegler, Stefan Kloppel, and Christian Gaser. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, 2010.

[124] Andrew Y. Ng; and Michael I. Jordan. On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. In *Neural Information Processing Systems*, 2002.

[125] Thomas P Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[126] Malte Kuss, Carl Edward Rasmussen, and Carl@tuebingen Mpg De. Assessing Approximate Inference for Binary Gaussian Process Classification Malte Kuss. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

[127] Jaakko Riihimäki, Pasi Jylänki, and Aki Vehtari. Nested Expectation Propagation for Gaussian Process Classification with a Multinomial Probit Likelihood. *Journal of Machine Learning Research*, 14:75–109, 2013.

[128] Jonathan Young, Marc Modat, Manuel J. Cardoso, Alex Mendelson, Dave Cash, and Sebastien Ourselin. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2013.

[129] Alzheimers Disease Neuroimaging Initiative. Alzheimers Disease Neuroimaging Initiative, 2017.

[130] Edward Challis, Peter Hurley, Laura Serra, Marco Bozzali, Seb Oliver, and Mara Cercignani. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage*, 112:232–243, 5 2015.

[131] Wolfgang Fruehwirt, Pengfei Zhang, Matthias Gerstgrasser, Dieter Grossegger, Reinhold Schmidt, Thomas Benke, Peter Dal-Bianco, Gerhard Ransmayr, Leonard Weydemann, Heinrich Garn, Markus Waser, Michael Osborne, and Georg Dorffner. Bayesian Gaussian process classification from event-related brain potentials in Alzheimers disease. In *Artificial Intelligence in Medicine*, volume 10259 LNAI, pages 65–75. Springer Verlag, 2017.

[132] Tong Tong, Katherine Gray, Qinquan Gao, Liang Chen, and Daniel Rueckert. Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. *Pattern Recognition*, 2017.

[133] Katherine R. Gray, Paul Aljabar, Rolf A. Heckemann, Alexander Hammers, and Daniel Rueckert. Random forest-based similarity measures for multimodal classification of Alzheimer's disease. *NeuroImage*, 2013.

[134] Alessia Sarica, Antonio Cerasa, and Aldo Quattrone. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Frontiers in aging neuroscience*, 9:329, 2017.

[135] Murat Bilgel, Jerry L Prince, Dean F Wong, Susan M Resnick, and Bruno M Jedynak. A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *NeuroImage*, 134:658–670, 2016.

[136] Bruno M Jedynak, Bo Liu, Andrew Lang, Yulia Gel, and Jerry L Prince. A computational method for computing an Alzheimer's disease progression score; experiments and validation with the ADNI data set. *Neurobiology of Aging*, 36(S1):178–184, 2015.

[137] Mahesh N. Samtani, Nandini Raghavan, Gerald Novak, Partha Nandy, and Vaibhav A. Narayan. Disease progression model for Clinical Dementia Rating-Sum of Boxes in mild cognitive impairment and Alzheimer's subjects from the Alzheimer's Disease Neuroimaging Initiative. *Neuropsychiatric Disease and Treatment*, 2014.

[138] Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease. *NeuroImage*, 10 2017.

[139] Jonathan D Rohrer, Jennifer M Nicholas, David M Cash, John van Swieten, Elise Dopper, Lize Jiskoot, Rick van Minkelen, Serge A Rombouts, M Jorge Cardoso, Shona Clegg, Miklos Espak, Simon Mead, David L Thomas, Enrico De Vita, Mario Masellis, Sandra E Black, Morris Freedman, Ron Keren, Bradley J MacIntosh, Ekaterina Rogaeva, David Tang-Wai, Maria Carmela Tartaglia, Robert Laforce, Fabrizio Tagliavini, Pietro Tiraboschi, Veronica Redaelli, Sara Prioni, Marina Grisoli, Barbara Borroni, Alessandro Padovani, Daniela Galimberti, Elio Scarpini, Andrea Arighi, Giorgio Fumagalli, James B Rowe, Ian Coyle-Gilchrist, Caroline Graff, Marie Fallström, Vesna Jelic, Anne Kinhult Ståhlbom, Christin Andersson, Hkan Thonberg, Lena Lilius, Giovanni B Frisoni, Michela Pievani, Martina Bocchetta, Luisa Benussi, Roberta Ghidoni, Elizabeth Finger, Sandro Sorbi, Benedetta Nacmias, Gemma Lombardi, Cristina Polito, Jason D Warren, Sebastien Ourselin, Nick C Fox, and Martin N Rossor. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis. *The Lancet Neurology*, 14(3):253–262, 2015.

[140] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus Robert Müller. Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399, 2011.

[141] Jorge L. Bernal-Rusiel, Douglas N. Greve, Martin Reuter, Bruce Fischl, and Mert R. Sabuncu. Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. *NeuroImage*, 2013.

[142] Geert Verbeke; and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2009.

[143] David A Harville. Maximum Likelihood Approaches to Variance Component Estimation and to Related. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[144] José Pinheiro, José C Bates, and Douglas M. Model Building for Nonlinear Mixed-Effects Models. Technical report.

[145] Jorge L. Bernal-Rusiel, Martin Reuter, Douglas N. Greve, Bruce Fischl, and Mert R. Sabuncu. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. *NeuroImage*, 2013.

[146] Sheng Luo and Jue Wang. Bayesian hierarchical model for multiple repeated measures and survival data: An application to Parkinson's disease. *Statistics in Medicine*, 33(24):4279–4291, 2014.

[147] Mert R Sabuncu, Jorge L Bernal-rusiel, Martin Reuter, Douglas N Greve, and Bruce Fischl. NeuroImage Event time analysis of longitudinal neuroimage data. *NeuroImage*, 97:9–18, 2014.

[148] Jonathan D Rohrer, Jennifer M Nicholas, David M Cash, John van Swieten, Elise Dopper, Lize Jiskoot, Rick van Minkelen, Serge A Rombouts, M Jorge Cardoso, Shona Clegg, Miklos Espak, Simon Mead, David L Thomas, Enrico De Vita, Mario Masellis, Sandra E Black, Morris Freedman, Ron Keren, Bradley J MacIntosh, Ekaterina Rogaeva, David Tang-Wai, Maria Carmela Tartaglia, Robert Laforce, Fabrizio Tagliavini, Pietro Tiraboschi, Veronica Redaelli, Sara Prioni, Marina Grisoli, Barbara Borroni, Alessandro Padovani, Daniela Galimberti, Elio Scarpini, Andrea Arighi, Giorgio Fumagalli, James B Rowe, Ian Coyle-Gilchrist, Caroline Graff, Marie Fallström, Vesna Jelic, Anne Kinhult Ståhlbom, Christin Andersson, Hkan Thonberg, Lena Lilius, Giovanni B Frisoni, Michela Pievani, Martina Bocchetta, Luisa Benussi, Roberta Ghidoni, Elizabeth Finger, Sandro Sorbi, Benedetta Nacmias, Gemma Lombardi, Cristina Polito, Jason D Warren, Sebastien Ourselin, Nick C Fox, and Martin N Rossor. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Fron-

totemporal dementia Initiative (GENFI) study: a cross-sectional analysis. *The Lancet Neurology*, 14(3):253–262, 2015.

[149] J. B. Schiratti, S. Allassonnière, A. Routier, O. Colliot, and S. Durrleman. A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[150] R. Guerrero, A. Schmidt-Richberg, C. Ledig, T. Tong, R. Wolz, and D. Rueckert. Instantiated mixed effects modeling of Alzheimer's disease markers. *NeuroImage*, 142:113–125, 11 2016.

[151] Peter Karcher and Yuedong Wang. Generalized nonparametric mixed effects models. *Journal of Computational and Graphical Statistics*, 10(4):641–655, 2001.

[152] Michael C. Donohue, Hlne Jacqmin-Gadda, Mlanie Le Goff, Ronald G. Thomas, Rema Raman, Anthony C. Gamst, Laurel A. Beckett, Clifford R. Jack, Michael W. Weiner, Jean Franois Dartigues, and Paul S. Aisen. Estimating long-term multivariate progression from short-term data. *Alzheimer's and Dementia*, 2014.

[153] Adrian Dalca, Ramesh Sridharan, Mert R Sabuncu, and Polina Golland. Predictive Modeling of Anatomy with Genetic and Clinical Data. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, volume 9351, pages 519–526, 2015.

[154] Tian Ge, Thomas E Nichols, Debashis Ghosh, Elizabeth C Mormino, Jordan W Smoller, and Mert R Sabuncu. A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. *NeuroImage*, 109:505–514, 2015.

[155] Alexander Schmidt-Richberg, Ricardo Guerrero, Christian Ledig, Helena Molina-Abril, Alejandro F. Frangi, and Daniel Rueckert. Multi-stage Biomarker Models for Progression Estimation in Alzheimers Disease. In Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin, and M. Jorge Cardoso, editors, *Information Processing in Medical Imaging*, volume 9123

of *Lecture Notes in Computer Science*, Cham, 2015. Springer International Publishing.

[156] Alexander Schmidt-Richberg, Christian Ledig, Ricardo Guerrero, Helena Molina-Abril, Alejandro F Frangi, and Daniel Rueckert. Learning Biomarker Models for Progression Estimation of Alzheimer ' s Disease. *PloS one*, 11(4):1–27, 2016.

[157] Zitao Liu and Milos Hauskrecht. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial Intelligence in Medicine*, 65(1):5–18, 2015.

[158] Jung Won Hyun, Yimei Li, Chao Huang, Martin Styner, Weili Lin, and Hongtu Zhu. STGP: Spatio-temporal Gaussian process models for longitudinal neuroimaging data. *NeuroImage*, 134:550–562, 2016.

[159] Marco Lorenzi, Maurizio Filippone, Giovanni B. Frisoni, Daniel C. Alexander, and Sebastien Ourselin. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in Alzheimer's disease, 2019.

[160] Jonathan D F Wadsworth, Andrew F Hill, Jonathan A Beck, and John Collinge. Molecular and clinical classification of human prion disease. *British Medical Bulletin*, 66:241–254, 2003.

[161] Marc Modat, David M Cash, Pankaj Daga, Gavin P Winston, John S Duncan, and Sbastien Ourselin. Global image registration using a symmetric block-matching approach. *Journal of Medical Imaging*, 1(2):24003, 2014.

[162] M Jorge Cardoso, Matthew J Clarkson, Gerard R Ridgway, Marc Modat, Nick C Fox, and Sebastien Ourselin. LoAd: a locally adaptive cortical segmentation algorithm. *NeuroImage*, 56(3):1386–1397, 2011.

[163] Carole Sudre, M Jorge Cardoso, Willem Bouvy, Geert Biessels, Josephine Barnes, and Sebastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *Medical Imaging, IEEE Trans. On (TMI)*, 34(c):1, 2015.

[164] Andrew Melbourne, Nicolas Toussaint, David Owen, Ivor Simpson, Thanasis Anthopoulos, Enrico De Vita, David Atkinson, and Sebastien Ourselin. Nifty-

Fit: a Software Package for Multi-parametric Model-Fitting of 4D Magnetic Resonance Imaging Data. *Neuroinformatics*, 2016.

[165] M Jorge Cardoso, Marc Modat, Robin Wolz, Andrew Melbourne, David Cash, Daniel Rueckert, and Sebastien Ourselin. Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. *IEEE Transactions on Medical Imaging*, 34(9):1976–1988, 2015.

[166] Y L Chung, A Williams, D Ritchie, S C Williams, K K Changani, J Hope, and J D Bell. Conflicting MRI signals from gliosis and neuronal vacuolation in prion diseases. *Neuroreport*, 10(17):3471–3477, 1999.

[167] Martin Zeidler, Robin J. Sellar, Donald A. Collier, Richard Knight, Gillian Stewart, Margaret Ann Macleod, James W. Ironside, Simon Cousens, Alan F.C. Colchester, Donald M. Hadley, and Robert G. Will. The pulvinar sign on magnetic resonance imaging in variant Creutzfeldt-Jakob disease. *Lancet*, 2000.

[168] Geoffrey McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004.

[169] Claudio Soto and Nikunj Satani. The intricate mechanisms of neurodegeneration in prion diseases, 2011.

[170] Denis Le Bihan, Jean-Franois Mangin, Cyril Poupon, Chris A Clark, Sabina Pappata, Nicolas Molko, and Hughes Chabriat. Diffusion Tensor Imaging: Concepts and Applications. *Journal of Magnetic Resonance Imaging*, 13, 2001.

[171] V S Fonov, A C Evans, R C McKinstry, C R Almli, and D L Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009.

[172] Ellen B Roecker. Prediction Error and Its Estimation for Subset-Selected Models. *Technometrics*, 33(4):459–468, 11 1991.

[173] Benson Mwangi, Tian Siva Tian, and Jair C. Soares. A review of feature reduction techniques in Neuroimaging, 2014.

[174] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference,and Prediction*, volume 27. Springer, 2 edition, 2009.

[175] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Cordinate Descent. *Journal of Statistical Software*, 58(9):1–22, 2014.

[176] Florentina Bunea, Yiyuan She, Hernando Ombao, Assawin Gongvatana, Kate Devlin, and Ronald Cohen. Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 2011.

[177] Gavin C Cawley and Nicola L C Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

[178] Joshua B Tenenbaum. Mapping a manifold of perceptual observations. In *Advances in Neural Information Processing Systems 10*, pages 682–688, 1998.

[179] Laurens Van Der Maaten, Eric Postma, and Jaap Van Den Herik. Dimensionality Reduction: A Comparative Review. Technical report, Tilburg University, 2009.

[180] Ik Soo Lim, P. de Heras Ciechomski, S. Sarni, and D. Thalmann. Planar arrangement of high-dimensional biomedical data sets by isomap coordinates. In *16th IEEE Symposium Computer-Based Medical Systems, 2003. Proceedings.*, 2003.

[181] L.S. Canas, B. Yvernault, C. Sudre, E. De Vita, M.J. Cardoso, J. Thornton, F. Barkhof, S. Ourselin, S. Mead, and M. Modat. Imaging biomarkers for the diagnosis of Prion disease. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, volume 10574, 2018.

[182] S Roberts, M Osborne, M Ebden, S Reece, N Gibson, and S Aigrain. Gaussian Processes for Timeseries Modelling. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, pages 1–27, 2012.

[183] David K Duvenaud, Hannes Nickisch, and Carl E Rasmussen. Additive Gaussian Processes. In J Shawe-Taylor, R S Zemel, P L Bartlett, F Pereira, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 226–234. Curran Associates, Inc., 2011.

[184] Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville

Tolvanen, and Aki Vehtari. GPstuff: Bayesian Modeling with Gaussian Processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013.

[185] Alain Rakotomamonjy;, Francis R. Bach;, Stéphane Canu;, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[186] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. SVM and Kernel Methods Matlab Toolbox, 2005.

[187] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 2009.

[188] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006.

[189] W Chu and Z Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

[190] Thomas P. Minka. Expectation Propagation for approximate Bayesian inference. 2013.

[191] John C Morris, Paul S Aisen, Randall J Bateman, Tammie LS Benzinger, Nigel J Cairns, Anne M Fagan, Bernardino Ghetti, Alison M Goate, David M Holtzman, William E Klunk, Eric McDade, Daniel S Marcus, Ralph N Martins, Colin L Masters, Richard Mayeux, Angela Oliver, Kimberly Quaid, John M Ringman, Martin N Rossor, Stephen Salloway, Peter R Schofield, Natalie J Selsor, Reisa A Sperling, Michael W Weiner, Chengjie Xiong, Krista L Moulder, and Virginia D Buckles. Developing an international network for Alzheimers research: the Dominantly Inherited Alzheimer Network. *Clinical Investigation*, 2(10):975–984, 10 2012.

[192] Neil P. Oxtoby, Alexandra L. Young, David M. Cash, Tammie L. S. Benzinger, Anne M. Fagan, John C. Morris, Randall J. Bateman, Nick C. Fox, Jonathan M. Schott, and Daniel C. Alexander. Disease progression models for dominantly-inherited Alzheimers disease. *Brain*, 141(5):1244–1246, 3 2018.

[193] Kan Li, Richard O'Brien, Michael Lutz, and Sheng Luo. A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data. *Alzheimer's and Dementia*, 14(5):644–651, 5 2018.

[194] Ron Brookmeyer and Nada Abdalla. Multistate models and lifetime risk estimation: Application to Alzheimer's disease. *Statistics in Medicine*, 38(9):1558–1565, 4 2019.

[195] Mengying Sun, Inci M. Baytas, Liang Zhan, Zhangyang Wang, and Jiayu Zhou. Subspace network: Deep multi-task censored regression for modeling neurodegenerative diseases. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2259–2268. Association for Computing Machinery, 7 2018.

[196] Zhigang Zhang and Jianguo Sun. Interval censoring, 2 2010.

[197] Orla M. Doyle, Eric Westman, Andre F. Marquand, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, Simon Lovestone, Steve C.R. Williams, and Andrew Simmons. Predicting progression of Alzheimer's disease using ordinal regression. *PLoS ONE*, 9(8), 8 2014.

[198] Marco Lorenzi, Xavier Pennec, Giovanni B. Frisoni, and Nicholas Ayache. Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. *Neurobiology of Aging*, 2015.

[199] Mingxia Liu, Jun Zhang, Pew Thian Yap, and Dinggang Shen. View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multimodality data. *Medical Image Analysis*, 2017.

[200] Alireza Farhangfar, Lukasz Kurgan, and Jennifer Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 2008.

[201] Lei Yuan, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 7 2012.

[202] Kim Han Thung, Pew Thian Yap, and Dinggang Shen. Multi-stage diagnosis of Alzheimers disease with incomplete multimodal data via multi-task deep learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[203] T. Schneider. Analysis of incomplete climate data: Estimation of Mean Values

and covariance matrices and imputation of Missing values. *Journal of Climate*, 2001.

[204] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing Missing Data for Gene Expression Arrays. Technical report, Stanford University, California, 1999.

[205] Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with Student-t likelihood. In Y. Bengio and D. Schuurmans and J. D. Lafferty and C. K. I. Williams and A. Culotta, editor, *Advances in Neural Information Processing Systems*, pages 1910–1918. Curran Associates, Inc., 2009.

[206] Jylänki Pasijylanki and Vanhatalo Jarnovanhatalo. Robust Gaussian Process Regression with a Student-t Likelihood Aki Vehtari. *Journal of Machine Learning Research*, 12:3227–3257, 2011.

[207] Joaquin Quiñonero-Candela and Lars Kai Hansen. Time Series Prediction Based on the Relevance Vector Machine with Adaptive Kernels. In *International Conference on Acoustics, Speech and Signal Processing*, 2002.

[208] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

[209] Nagarajan Natarajan, Nagarajn@microsoft Com, Pradeep Ravikumar, Machine Learning, and Ambuj Tewari. Cost-Sensitive Learning with Noisy Labels. *Journal of Machine Learning Research*, 18:1–33, 2018.

[210] Ricardo Henao and Ole Winther. PASS-GP: Predictive active set selection for Gaussian processes. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2010*, 2010.

[211] Patrick Dallaire, Camille Besse, and Brahim Chaib-Draa. Learning Gaussian process models from uncertain data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009.

[212] Neil D Lawrence. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. Technical report.

[213] Michalis K Titsias and Neil D Lawrence. Bayesian Gaussian process latent variable model. In *International Conference on Artificial Intelligence and Statistics*, Sardinia, 5 2010.

[214] GPy. GPy: A Gaussian process framework in python, 2012.

[215] James Hensman, Neil D Lawrence, and Magnus Rattray. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. Technical report, 2013.

[216] Pablo Moreno-Muñoz, Antonio Artés-Rodríguez, and Mauricio A Álvarez. Heterogeneous Multi-output Gaussian Process Prediction. In *Neural Information Processing Systems (NeurIPS 2018)*, Montréal, 2018.

[217] Francesco Carlo Morabito, Maurizio Campolo, Nadia Mammone, Mario Versaci, Silvana Franceschetti, Fabrizio Tagliavini, Vito Sofia, Daniela Fatuzzo, Antonio Gambardella, Angelo Labate, Laura Mumoli, Giovanbattista Gaspare Tripodi, Sara Gasparini, Vittoria Cianci, Chiara Sueri, Edoardo Ferlazzo, and Umberto Aguglia. Deep Learning Representation from Electroencephalography of Early-Stage Creutzfeldt-Jakob Disease and Features for Differentiation from Rapidly Progressive Dementia. *International Journal of Neural Systems*, 27(2), 2016.

[218] Meenal J. Patel, Alexander Khalaf, and Howard J. Aizenstein. Studying depression using imaging and machine learning methods, 2016.

[219] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer, Cham, 2018.

[220] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *International Conference on Intelligent Computing*, volume 3644, pages 878–887, 2005.

[221] Haibo He, Yang Bai, E A Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.

[222] Rok Blagus and Lara Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14(1):106, 2013.

[223] Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. Unsupervised Data Imputation via Variational Inference of Deep Subspaces. In *CVPR*, 3 2019.

[224] Liane S Canas, Benjamin Yvernault, Carole H Sudre, Jorge Cardoso, John Thornton, Frederik Barkhof, Sebastien Ourselin, Simon Mead, and Marc Modat. Multikernel Gaussian Processes for patient stratification from imaging biomarkers with heterogeneous patterns. In *Learning from Limited Labeled Data: Weak Supervision and Beyond, NIPS*, Long Beach, 2017.

[225] Yunseong Hwang, Anh Tong, and Jaesik Choi. Automatic Construction of Nonparametric Relational Regression Models for Multiple Time Series. *The 33rd International Conference on Machine Learning (ICML 2016)*, 48, 2016.

[226] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.