Objective Assessment of Parkinson's Disease using Machine Learning

John Prince

Wolfson College

University of Oxford

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

October 2018

Supervised by:
Prof. Maarten De Vos
Prof. David Clifton

Objective Assessment of Parkinson's Disease using Machine Learning

John Prince
Doctor of Philosophy Thesis
Wolfson College
Trinity 2018

Abstract

Neurodegenerative disorders, such as Parkinson's disease (PD), are infamously heterogeneous in regards to onset age, symptom prevalence, and severity progression rate. The current 'gold standard' techniques currently in use to assess the highly debilitating motor and non-motor symptoms are subjective, infrequent, and often ineffective with up to 20% of cases going undiagnosed until post-mortem. As such, these traditional and well established clinical assessment techniques are now starting to be fused with data-driven approaches in a bid to improve diagnosis and severity monitoring through the identification of objective disease biomarkers. Digital sensors present the opportunity to extract quantitative measures that are representative of disease presence and severity. However, existing studies using digital sensors to perform disease quantification are restricted by small cohorts, inconsistent experimental protocols, and purely cross-sectional analyses. The objective of this thesis is to provide novel insights as to how digital sensors can be further leveraged alongside prior clinical knowledge in order to improve the way Parkinson's disease is assessed in both clinical and remote environments.

This thesis focuses on two large datasets that are juxtaposed in data quality, collection environment, and data type. Firstly, a clinical based approach to disease quantification is performed wherein an extensive network of wearable sensors is introduced into routine clinical care. Using this clinical dataset, the ability of wearable sensors to detect digital biomarkers distinctive of PD is investigated including the use of these biomarkers to perform automatic disease classification and severity prediction. Due to the longitudinal nature of data collection, new insights are revealed pertaining to symptom progression and the benefit of including longitudinal symptom variation into classification tasks. Secondly, the efficacy of performing disease assessment entirely in a remote environment using smartphones is investigated. Data was collected relating to many areas of disease manifestation and was often contributed daily by participants on a longitudinal basis. However, the remotely collected data suffers from a large degree of missingness, poor participant retention rate, and variable environmental conditions presenting new challenges during its analysis. High-frequency longitudinal analyses are performed and identify previously unseen motor and non-motor symptom progression characteristics. In order to perform disease classification using this dataset, a novel methodology is presented that compensates for the large quantities of source-wise missing data which, when combined with a state-of-the-art convolutional neural network, subsequently improves classification accuracy from 73.1% to 82.0%. Finally, a consistent analysis protocol is implemented on both datasets whilst simulating source-wise missing data; enabling a comprehensive comparison of missing data strategies for the purpose of disease classification.

This thesis presents findings that highly support the hypothesis that digital sensors can successfully perform objective disease assessment at both the individual and population level.

Acknowledgements

Firstly, I would like to thank Maarten for all of the time he has given me over the last three years. When the project direction was unclear, Maarten went far beyond the call of a supervisor whilst searching for viable alternatives and always seemed to have an endless stream of interesting projects up his sleeve. Once Parkinson's had been settled on as a final project, Maarten always had the right balance of letting me freely explore my interests whilst guiding me towards clinically orientated objectives.

Secondly, a thanks to David Clifton for providing me with additional project guidance - especially when the thesis direction was unclear. David's clear cut plan for my transfer of status helped tremendously during a stressful time whilst also letting me sample another potential thesis direction. My involvement in the Vietnamese studies gave me some truly unique field work experience in a high-impact affordable healthcare project.

I'd also like to thank the CIBIM lab. Thanks to Fernando for selflessly reading and commenting on the vast majority of this thesis and for contributing to my publications. Additional thanks go to Ollie, Kirubin, Navin, and Andy for making work seem less like work.

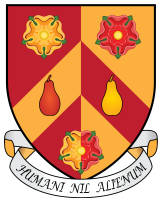
Thank you to my clinical collaborators, namely Chrystalina Antoniades, James Fitzgerald, and the whole OxQUIP team. My involvement in this study provided me the opportunity to work on the front-line of clinical disease assessment and has contributed significantly to this thesis.

Thank you to all members of the StAR team including David Springer, Andrew Farmer, and Kirsty Bobrow. During my involvement in the study I have worked on a diverse range of tasks that kept me constantly learning new technical and clinical skills in a completely different research area to my thesis. It has been hugely satisfying working on such a large scale project with such a supportive, diverse, and flexible team.

I would also like to acknowledge the support of the RCUK Digital Economy Programme and the generous support from Wolfson College through the Oxford-Wolfson-Marriott Graduate Scholarship in Healthcare Innovation.

In *crohn*ological order of appearance I'd like to thank Tom McBride, Anton Stalker, George Sloan, and Eddie Marks.

Finally, thank you to my family who have always shown/convincingly feigned an equal level of interest and enthusiasm about my work as I have. Their unconditional support is without a doubt what has gotten me through this degree and is therefore who this thesis is dedicated to.



Associated Publications

1. J. Prince, S. Arora, M. De Vos, "*Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes*", Physiological Measurement, vol 39, no 4, p 0044005, April 2018. [1]

2. J. Prince and M. De Vos, "*A Deep Learning Framework for the Remote Detection of Parkinson's Disease Using Smart-phone Sensor Data*" in Engineering in Medicine and Biology Society (EMBC), July 2018, Hawaii, IEEE [2]

3. J. Prince, F. Andreotti, M. De Vos, "*Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-wise Missing Data*", IEEE Transactions on Biomedical Engineering, August 2018. [3]

4. J. Prince, F. Andreotti, M. De Vos, "*Effects of Source-wise Missing Data Strategies on Classifier Performance*" in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, Brighton, (Invited Submission: Submission 29$^{\text{th}}$ October 2018) [4]

5. J. Prince, J. Fitzgerald, C. Antoniades, M. De Vos, "*Differentiation of Parkinsonian Disorders Using Wearable Sensors and Machine Learning*", Movement Disorders (Manuscript in Preparation).

Recurring Abbreviations

AFT	Alternating Finger Tapping
CNN	Convolutational Neural Network
DNN	Deep Neural Network
FOG	Freezing of Gait
HC	Healthy Controls
IMU	Inertial Measurement Unit
k-NN	k-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
L-PT	Learning-Phenotype
LR	Logistic Regression
MAE	Mean Absolute Error
MDS-UPDRS	Movement Disorder Society's-Unified Parkinson's Disease Rating Scale
M-TL	Multi-Task Learning
OxQUIP	The Oxford study of Quantification in Parkinsonism
PD	Parkinson's Disease
RF	Random Forest
RMSE	Root Mean Squared Error
SG-LASSO	Sparse Grouped-LASSO
SRTT	Serial Reaction Timed Test
STD	Standard Deviation
SVM	Support Vector Machine
TUG	Timed-Up-and-Go test
UPDRS	Unified Parkinson's Disease Rating Scale
YHC	Young Healthy Controls

Contents

List of Figures

List of Tables

1 Introduction

1.1 Clinical Motivation

Parkinson's disease (PD) is the second most common neurodegenerative disease, secondly only to Alzheimer's disease, and affects an estimated seven million people globally [5, 6]. The reported prevalence of PD varies due to environmental and genetic factors with the European prevalence rates ranging from 65.6 to 300 per 100,000 within the entire population [7]. In the United Kingdom, an estimated 127,000 people suffer from PD corresponding to a prevalence rate of 200 per 100,000 [8]. However, as a progressive neurodegenerative disease, PD mostly afflicts the elderly with prevalence rates ranging from 700 to 1500 per 100,000 in people over the age of 50 [7].

Parkinson's disease is considered a movement disorder and presents a multitude of motor symptoms that significantly impact quality of life. Most notably, motor symptoms pertaining to tremor, rigidity, akinesia (loss of control of movements), and postural instability are considered the cardinal symptoms and whose presence govern the clinical diagnosis decision [9]. However, non-motor symptoms are also widely reported particularly in relation to diminished cognitive ability, depression, and sleep disorders [10]. Instances of dementia are particularly pronounced in the later stages of the disease, occurring in an estimated 40% of patients [11]. There exists no cure for PD therefore all treatment types are orientated towards lessening symptom severity and decreasing the rate of progression [12]. If undetected and untreated, disease progression is rapid, often leading to complete loss of independence within eight to ten years [13]. Nonetheless, even when treated, the life expectancy of patients is reduced [14, 15]. Pulmonary conditions, including pneumonia, contribute

significantly to mortality risk of PD with roughly 100,000 deaths occurring in the global PD population in 2013, demonstrating a 28.2% increase since 1990 [16]

The current requirement for PD diagnosis and treatment to be performed entirely in a clinic environment incurs a substantial cost for healthcare services with annual service costs in the UK being estimated at over £13,000 per person per year [17, 18]. Therefore, with the projected number of people with PD to increase dramatically over the next decade, the ability to reliably perform assessments without the requirement of a clinical visit is highly desirable [19].

The current 'gold standard' assessment technique is a clinically administered rating scale wherein patients are visually assessed by a movement disorder specialist roughly once every three months. These rating scales are widely accepted to suffer from subjectivity as well as inter- and intra-rater variability [20, 21]. These limitations of the scoring system make diagnosis and monitoring extremely difficult to perform reliably, often leading to high misdiagnosis rates [22]. Further, longitudinal disease progression is poorly monitored as the variable nature of the rating scale often masks changes in symptom variation [23, 24].

The diagnostic challenges, in addition to the shortcomings of the severity scoring systems, have led to an active area of research identifying objective biomarkers capable of assessing PD though the use of digital sensors. The benefit of using digital sensors is that they enable the objective measurement of symptoms, thus removing the subjectivity of the current scoring systems. Such biomarkers can be used to perform disease classification or to predict the severity of symptoms. Most recently, the multiple sensors within smart-phones have been used to remotely collect data intended to measure symptom severity on a much more frequent basis.

The objective of this thesis is to determine the ability of digital sensors to be leveraged alongside prior clinical knowledge in order to improve the techniques in which Parkinson's disease is assessed in both clinical and remote environments.

1.2 Research Objectives of the Thesis

This thesis approaches the task of PD quantification from two directions. As the current means of disease assessment is performed in a clinical environment, the first approach entails the use of digital sensors as part of routine clinical care. The second approach entails transitioning the quantification techniques from a clinical environment into a remote environment. This therefore provides a 'bottom-up' approach via developing and validating techniques in a controlled setting and then increasing their complexity as needed to be suitable on less-controlled data in order to perform the same tasks. This thesis focuses on answering the following broad research objectives:

1. Can wearable sensors be incorporated into routine clinical care to identify objective clinical measures capable of disease quantification?

2. Can objective clinical measures successfully perform disease classification and severity prediction on a diverse population and, if so, which clinical tests and clinical measures are most capable of doing so?

3. Can objective clinical measures also be used to perform longitudinal disease monitoring during routine follow up clinical visits? If so, does the longitudinal behaviour of the measures differ between disease groups and can this be used to improve disease classification and severity prediction?

4. Can features be extracted from remotely collected data on a longitudinal basis to detect high-frequency symptom variation? If so, can personalised and population based conclusions be drawn that relate to disease presence and severity progression?

5. Can features extracted from multiple test-types in a remote environment be used to perform remote disease classification? If so, do the most pertinent features agree with those detected in the clinical setting when a different sample size is used for feature selection?

6. What is the effect of implementing an identical analysis protocol on two vastly different datasets for the purpose of disease classification in the presence of missing data?

1.3 Structure of the Thesis

The following chapter provides a succinct background to PD and its current assessment techniques followed by a literature review detailing the progress made in using digital sensors for objective disease assessment thus far. Chapter 3 details the technical methods utilised during later chapters including the definition of statistical procedures, machine learning models, and evaluation techniques. Chapter 4 introduces the two datasets with respect to their data types and participant demographics through a series of exploratory analyses which culminates in a discussion as to their primary differences. Chapters 5 to 8 are the main results chapters of this thesis. Chapter 5 utilises a clinically measured dataset to perform cross sectional and longitudinal disease assessment including classification and regression tasks. Chapter 6 pivots the focus of analysis to a remote environment via performing a longitudinal investigation on tests performed on smart-phones. Chapter 7 continues to investigate the remotely collected data but for the purpose of disease classification. This requires substantially more sophisticated dataset manipulation and classification techniques due to the nature of the remotely collected dataset. During chapter 8, the final results chapter, a series of techniques for overcoming missing data are employed on both the clinical and remotely acquired datasets for the purpose of disease classification. Finally, chapter 9 concludes the thesis by tying together the findings of each chapter whilst highlighting limitations and directions for future works.

1.4 Summary of Contributions

This section summarises the primary contributions made by each of the results chapters in this thesis:

- *Chapter 5*
  - The use of clinical features derived from wearable sensors to perform cross-sectional disease classification and severity prediction on a large and diverse clinical cohort.
  - The identification of novel longitudinal symptom behaviour as determined though objective clinical markers.
  - The implementation of longitudinal disease classification and severity tracking which incorporate longitudinal feature fluctuation.
- *Chapter 6*
  - The use of high-frequency longitudinal monitoring of symptom progression using data collected remotely on smart-phones on the largest cohort to date.
  - The identification of previously unseen longitudinal symptom behaviour in a motor and non-motor task which subsequently show differences between healthy controls and the Parkinson's population.
  - The first investigation of the relationship between the longitudinal behaviour of motor and non-motor symptom progression as measured in a remote environment.
- *Chapter 7*
  - The development of a novel methodology to overcome source-wise missing data that enables participants with missing data to be utilised in the development of machine learning models.
  - The application of this methodology, in conjunction with a series of convolutional neural networks, to a remotely collected dataset for the purpose of disease classification; proving highly superior in comparison to traditional missing data compensation techniques.

  - An investigation into the effect of sample size on the feature selection process on a dataset known to be heavily corrupted by various forms of noise.

- *Chapter 8*

  - A comparison of many missing data techniques, including the novel methodology and a deep neural network autoencoder, for the purpose of disease classification on two datasets.
  - Source-wise missing data is simulated such as to facilitate the study of each technique under variable degrees of missingness.
  - A quantitative analysis of the errors induced by imputation and autoencoding techniques which subsequently reveal the applicability of each technique in relation to the type of dataset being studied.

2

Parkinson's Disease: A Physiological & Assessment Background

This chapter provides the reader with a succinct background of the essential characteristics and assessment techniques for Parkinson's Disease (PD). Firstly, a brief history and introduction to PD is provided. This is intended to give the reader an insight into the pathophysiology, symptoms, and treatment challenges associated with the disease that will be essential for understanding the motivation of later chapters in this thesis. Secondly, the current 'gold standard' techniques for disease assessment are outlined with a focus on the clinical procedures and scoring systems. As will be discussed further, these systems are widely acknowledged to suffer from many limitations which are likely to lead to sub-optimal patient care and disease assessment. These limitations motivate the development of novel disease assessment techniques that form the backbone of this thesis. Finally, a detailed literature review is provided detailing the state-of-the-art disease assessment techniques wherein the use of digital sensors are employed. This encompasses both in-clinic and remote assessment tools for a wide array of disease symptoms. Through this literature review, the current landscape of objective PD assessment is summarised facilitating the identification of knowledge gaps in the literature thus motivating future research directions.

2.1 Introduction to Parkinson's Disease

Parkinson's disease (PD) first entered medical literature through James Parkinson's conjectures within his 1817 work "An essay on the shaking palsy" [25]. Although Parkinson only provided observational details for six cases, he provided a 'descriptive

and vivid definition' of the symptoms leading to a study in 1876 coining the disease as 'maladie de Parkinson' [26, 27]. It was not until 1961 that the diminished activity of dopamine within the brain was linked with the disease when the administration of dopaminergic therapies found to heavily alleviate symptoms – a finding that would lead to the author being awarded the Nobel Prize in Medicine in 2000 [28]. This finding paved the way for many of the modern areas of research into PD including the identification of more complex pathogenic mechanisms and treatment strategies [29].

Although the pathophysiology of PD is now linked to the gradual degeneration of neurons within the substantia nigra which causes a reduction of dopamine uptake in the basal ganglia, the underlying cause of the disease is unknown in the vast majority of cases [30–32]. The basal ganglia plays a vital role in muscle control and whose communication with the rest of the brain is facilitated through the neurotransmitter dopamine. As gradual neuron death occurs in the substantia nigra, less dopamine uptake occurs in the basal ganglia causing a loss of communication between the brain and the muscles – giving rise to the dysfunctional motor symptoms that define the disease [33].

The most common symptoms of PD are often abbreviated as TRAP: Tremor at rest, Rigidity, Akinesia, and Postural instability [27]. Indeed, it is the presence of these symptoms that are used to make initial disease diagnosis on purely clinical grounds [34–36]. However, there is no definitive test for PD with these symptoms commonly being used alongside medical history to form a diagnosis; often leading to high misdiagnosis rates (with diagnosis confirmed by mid-brain autopsy) [9, 31, 37]. Additional and more subtle motor symptoms may present throughout disease progression including swallowing [38], vocal [39–41], and ocular impairments [21, 42, 43]; although these are not integrated to the diagnosis decision. Perhaps the most disconcerting characteristic of the disease, and that which presents the most significant diagnostic challenge, is that roughly 60% of dopaminergic cells are lost in the substantia nigra by the time the first symptoms appear [31].

Contrary to James Parkinson's statement that "the senses and intellects" were "uninjured" in sufferers of shaking palsy, it is now widely accepted that many non-

motor symptoms occur in PD. The most frequently seen non-motors symptoms in the PD population include cognitive impairment, depression, and sleep disorders [44–46]. Not only do non-motor symptoms cause significant morbidity during the later stages of disease progression but the vast majority of people with PD report the presence of at least one non-motor symptom, with many reporting that the impact of non-motor symptoms have a larger affect on their quality of life than their motor counterparts [10, 47, 48]. Indeed, whereas motor-symptoms tend to present and deteriorate, non-motor symptoms such as memory impairment can be highly predictive of the later development of additional non-motor symptoms such as dementia [49].

As an incurable neurodegenerative disease, treatments for PD are directed at alleviating symptoms and slowing disease progression. Although the use of dopaminergic therapy for the alleviation of motor symptoms was first demonstrated in 1961, it took almost a decade of clinical trials before it was proved a widely successful treatment [50–52]. Dopaminergic therapies, specifically levodopa (L-Dopa) remain the most widely utilised treatment type and is considered the default treatment due to its very high therapeutic power and affordability [53]. However, the use of L-Dopa has been likened to 'a double-edged sword' on account of its ability to alleviate some symptoms, whilst exacerbating others (levodopa-induced dyskinesia) and having reduced beneficial effects over long time frames [54–56]. It was found that between 30% to 50% of patients using L-Dopa developed levodopa-induced dyskinesia within five years of usage and almost all patients experiencing negative effects during the entire course of treatment [57, 58]. The use of neurostimulation of the subthalamic nucleus, commonly referred to as Deep Brain Stimulation (DBS) is a surgical alternative to dopaminergic therapies that has shown particular efficacy with advanced stages of PD [59, 60]. Although the use of DBS sidesteps levodopa-induced dyskinesia, its usage must be balanced against the surgical risk of the procedure [61].

As a brief aside on treatment, it is worthy to note that the term 'Parkinson's disease' is usually a shorthand expression for idiopathic Parkinson's disease, which falls under the parkinsonian disease family. Idiopathic Parkinson's disease accounts

for around 75% of parkinsonism whilst other disorders, grouped as *atypical parkinsonism*, present neuron death in the substantia nigra in addition to cell death in the central and peripheral nervous systems [31]. As such, Idiopathic and atypical parkinsonism present very similar symptoms and yet the use of L-Dopa in atypical parkinsonism has little to no benefit to symptoms [62].

This introduction to PD is finalised by considering some of the longitudinal disease progression characteristics and their implications. During the early stages of the disease (often before diagnosis and the starting of treatment) motor symptoms deteriorate rapidly [63]. Only after the emergence and deterioration of symptoms will a patient typically seek a clinical diagnosis. Due to the affordability and improvement of motor symptoms the vast majority of people with PD are given dopaminergic treatment. The use of L-Dopa dramatically slows the rate at which motor symptoms deteriorate although also induces the aforementioned negative effects after extended usage [14]. Unfortunately, L-Dopa specifically targets the motor symptoms, resulting in the continuation of non-motor symptom deterioration [64]. However, the natural progression of PD is found to be highly heterogeneous and non-linear across the population dependent on a myriad of factors including age of onset, genetics, gender, and medical history [13, 65]. The presence and severity of both motor and non-motor symptoms is highly variable for patients at all stages of disease progression and is further complicated by the presence of treatment induced complications [66]. Not only does symptom deterioration dramatically impact quality of life, it also incurs a substantial cost to healthcare services and roughly doubles the mortality ratios compared to the healthy population [13, 67].

2.2 Disease Clinical Assessment Tools

The challenges associated with the diagnosis and progression monitoring of PD were highlighted in the previous section. Attention is now turned to describing the current clinical techniques used for implemented disease assessment.

Although several scoring systems exist for PD assessment, such as the Hoehn-Yahr [68, 69] scale and the Schwab and England scale [70], by far the most

widely employed assessment tool is the Unified Parkinson's Disease Rating Scale (UPDRS); specifically the Movement Disorder Society sponsored revision version (MDS-UPDRS) [71]. The scale consists of four main sections intended to assess a wide range of disease characteristics:

- *Part I*: inspects 13 disease characteristics pertaining to "non-motor experiences of daily living".
- *Part II*: inspects 13 disease characteristics pertaining to "motor experiences of daily living".
- *Part III*: is commonly called the motor-UPDRS and contributes the largest number of tests inspecting 33 disease characteristics and forms the "motor examination".
- *Part IV*: contributes six scores which pertain to "motor complications"

Each of the 65 disease characteristics are assigned a score between 0 (normal) to 4 (severe). The total MDS-UPDRS score is simply a summation of each individual score, spanning 0 (completely normal) to 260 (severely disabled). The majority of the Part I questions and all of Part II have been specially designed such as to allow the patient or their caregiver to complete without the help of a movement disorder specialist. Conversely, Parts III and IV necessitate the presence of a movement disorder specialist to assign the scores.

Parts I and II ask the patient or caregiver to rate their ability to complete everyday tasks and to rate the influence of certain non-motor symptoms (such as daytime fatigue, urinary problems, or cognitive impairment) or motor symptoms (such as eating, writing, or swallowing) on their severity.

Predictably, the motor-UPDRS (Part III) carries the most significance in the scoring system. During the motor examination, patients undergo examinations of their speech, rigidity, finger and toe tapping ability, gait characteristics, and tremors. These are often repeated for both the left and right hand side of the body

as asymmetry is inherent in disease manifestation [72]. Part IV focuses on the fluctuation of motor symptoms and the degree to which dyskinesia exists.

Although the MDS-UPDRS wears the 'gold standard' belt for disease assessment, it is widely accepted to suffer from many limitations. Firstly, as a movement disorder specialist is required to complete the most important sections of the examination, the MDS-UPDRS is required to be performed in a clinical environment. Combined with the fact that the full MDS-UPDRS examination lasts roughly one hour, it would be extremely expensive and intrusive to the daily life of the patient to undergo frequent disease assessment. As such, a patient will usually undergo a full examination once every three months at most. During their infrequent clinical visit, a patient may undergo the full examination whilst off their medication ('OFF' state) followed by an additional examination whilst on their medication ('ON' state); resulting in a full assessment lasting many hours. All reported movement disorder specialists for MDS-UPDRS completion are in tertiary healthcare environments. This type of travel can be highly inconvenient for many patients living in rural areas. Bluntly, to undergo assessment, an elderly patient with severe and disabling motor symptoms may have to travel a long distance whilst having not taken their medication.

Secondly, as would be expected from any rating system wherein symptoms are assessed by visual inspection, the MDS-UPDRS inevitably suffers from subjectivity [37, 73, 74]. This subjectivity comes in the form of both inter-rater (between scorers) and intra-rater (between the same scorer) adding unreliability to the system [23, 75–77]. In a recent study, the mean difference in inter-rater MDS-UPDRS variability was reported to be between 1.7 and 5.4, whereby nurses, residents and movement disorder specialists were found to consistently assign higher MDS-UPDRS scores as compared to senior movement disorder specialist [23]. Thus, a patient may be seen by a movement disorder specialist and assigned their total score. Three months later they may return and be examined by a different movement disorder specialist and receive a higher score. In this scenario it is extremely difficult to clarify whether the increase in score is due to symptom deterioration or due to the inter-rater subjectivity. If the score variability is treated as noise and is assumed to

vary around the true score, it would simply be a case of repeating the assessment at frequent intervals to compensate for the noise, but this is impossible due to the points of the previous paragraph.

2.3 Disease Assessment Using Digital Sensors

On account of the shortcomings of the clinical scoring systems highlighted in the previous section, the use of digital sensors to identify objective disease biomarkers is widely sought. Wearable sensors, such as Inertial Measurement Units (IMUs), have been extensively used for body motion analysis whereas microphones, eye-tracking cameras, and most recently smart-phones have been used to assess other motor and non-motor symptoms [78, 79]. Inertial measurement units are by far the most widely used digital sensors in the study of PD due to their versatility and affordability [21]. The usage of IMUs ranges from the collection of high quality signals from specific body locations in clinical environments, to the use of the smart-phone embedded IMU to conveniently collect signals 'in the wild'. As will be discussed later in this section, both of these usages of IMUs have advantages and limitations for the purpose of PD quantification.

In this section, a literature review is presented on the usage of digital sensors for the objective quantification of PD. This section begins by detailing the usage of specific digital sensors (excluding smart-phones) for a range of individual clinical tests followed by discussing the recent uptake of smart-phones at data collection platforms for PD assessment.

2.3.1 Gait & Postural Instability

As many of the motor-UPDRS questions are directly compared with specific gait and postural instability symptoms, it is the most studied area using digital sensors [80]. The use of digital sensors for the quantification of gait was first published in 1958 and included the use of pressure sensors and video cameras [81]. Today, these same sensors (though technologically more advanced) are still considered the 'gold-standard' for motion analysis. However, they are often very expensive

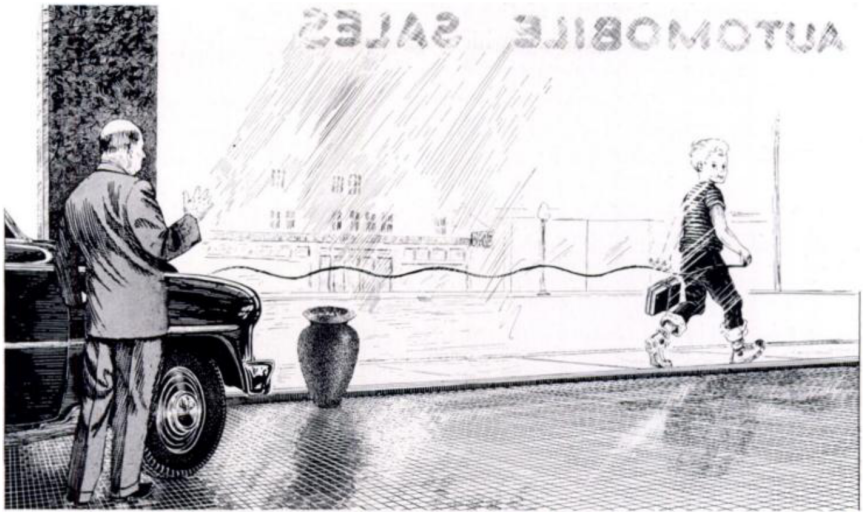


Figure 2.1: A crude gait measurement system. Figure reproduced with permission from Wolters Kluwer Health, Inc. and Copyright Clearance Center, Inc. under License Number 4423650476764. First published in 1953 by [91].

and are therefore only available in specific research laboratories [82, 83]. The number of studies that utilise such technologies have decreased notably over the last decade due to the availability of alternative and affordable digital sensors. The studies that continue to use the 'gold standard' technologies almost exclusively perform gait analyses [20, 84–87].

Although first conceived in 1783, it was not until 1961 that the accelerometer was first used to measure gait; but proved too expensive and impractical for widespread use [88, 89]. It was not until the last decade, with the reduction in cost of microelectromechanical systems (MEMS), that the popularity of IMUs for motion analysis increased dramatically [15, 90]. Figure 2.1 (first published in 1953 by [91]) provides a visualisation of the fundamental working principles of these types of devices. As the child leaves a trace of motion on the window, the oscillation can be observed and subsequently analysed for gait characteristics. For example, step length could be calculated by measuring the distance between two peaks.

Within the study of motion abnormalities in PD, the use of IMUs have predominantly focused on identifying features that either show statistical differences between people with PD and healthy controls (HC) or that correlate with clinically assigned severity scores such as the UPDRS [92–94]. The investigated features

tend to fall into one of two categories. In more clinically orientated studies, clinical specific features such as motion speed, cadence, motion variability, and symmetry are extracted from the signals [15, 95, 96]. Alternatively, in more technical studies, the features extracted pertain to the statistical, frequency, and spatiotemporal aspects of the signals [97–100]. Both approaches are valid and have shown capable of identifying a range of features showing statistical difference between people with PD and HCs. Often, the two feature sets are analogous to one another such as the clinical feature 'gait speed' being represented by the technical extraction of the IMUs signals dominant frequency. Indeed, the differences found between the groups match intuition with HCs showing a faster gait speed, longer stride length, and more symmetry [101, 102]. These same features (and their technical counterparts) demonstrated a generally strong ability to correlate with various aspects of the MDS-UPDRS with walking pace being the most widely studied variable; showing a correlation coefficient often exceeding 0.5 [93, 102–106].

Although the study and quantification of the Freezing-of-Gait (FOG) phenomenon is beyond the scope of this thesis, it is worthy of a brief note as it forms a substantial proportion of existing motion analysis literature. During FOG, a person with PD may become rigid and unable to move for up to several minutes [107]. Studies of FOG typically only recruit people with PD and attempt to find characteristics within the IMU signals that identify a freezing period [108–110]. The challenge in assessing FOG is that it usually presents in a later stage of the disease, and its occurrence can be unpredictable (though commonly induced by gait initiation or turns) [111]. Many of the studies investigating FOG suffer from low subject numbers, with an even smaller number presenting FOG episodes for assessment [20]. For a more exhaustive review on FOG quantification, the reader is directed to [110, 112–114].

The many studies objectifying gait features tend to have a high level of agreement with regards to which features show differences between PD and HCs. Conversely, the commonly used clinical Sway test produces many conflicting opinions in the literature when measured using digital sensors. Whereas some studies have suggested

the Sway test produces features distinct to PD [115–119], and even the potential to perform early diagnosis [120], other studies disagree or report uncertainty [121–123]. As the sway test is intended to assess postural instability (a cardinal symptom of PD), it would be expected to show differences between disease groups. In studies that discovered features showing significant differences, the jerkiness of posture, high frequency power, and total distance moved were the most commonly reported features, although not unanimously [20].

Due to versatility of IMUs, many studies have placed them at various locations on the body including the upper and lower limbs and on the torso [100, 123–125]. Studies utilising IMUs at multiple locations have shown the ability to detect asymmetric gait properties which can further differentiate people with PD from HC [96, 125, 126]. However, the variability of sensor placement has also been a hindrance in the generalisation of these results because studies extract different features (clinical or technical) from sensors at different body locations whilst completing tasks with differing experimental protocols [80]. As such, the optimal sensor arrangement for symptom objectification is yet to be identified as multiple confounding factors between studies are difficult to account for.

Surprisingly few studies have gone beyond traditional hypothesis testing and correlation analyses to perform classification or regression tasks using gait and sway tests, particularly with the use of IMU data. The use of Linear Discriminant Analysis (LDA) alongside the gait data collected from seven PD and seven HC participants using a high resolution video system yielded classification accuracy of 95.5% [127]. The features used in this study were entirely technical having been extracted from the pixel properties of each video frame. As this is an expensive sensor setup deriving abstract features from a small number of participants, this study makes for a good example of un-generalisable results. A k-Nearest Neighbour classifier was implemented on a cohort of 20 participants (10 HC: 10 PD) achieved a classification accuracy of 89% wherein IMUs were attached to the left and right leg during a period of gait [128]. However, all diseased participants were receiving DBS treatment therefore their symptoms were likely highly pronounced, biasing

the classification accuracy. On a similarly sized cohort (17 HC: 10 PD), [129] implemented a simple decision boundary classifier yielding an accuracy of 96.3% using data collected from a single waist mounted IMU during gait. However, this very high accuracy is here attributed to the nature of the small cohort. Indeed, the clinical features of this cohort showed no meaningful correlations with the MDS-UPDRS, contrary to the majority of existing literature [102, 104, 105, 130]. A linear Support Vector Machine (SVM) classifier was implemented during a gait test by [98] on a cohort of 34 PD participants and 4 HC participants yielding mis-classification rates of 22.5%. On such a limited and imbalanced dataset this result is highly unsuccessful. The most exhaustive use of gait IMU data for disease classification is that presented by [131]. A large cohort of 135 PD and 66 HC participants underwent classification using data arising from IMUs at six body locations. Using a clinical feature set, logistic regression was implemented achieving an Area Under the Receiver Operator Curve (AUROC) of 0.82. Furthermore, this study also performed classification using data from the Sway test using the same experimental set-up and cohort. Although the sway test yielded poorer classification results in comparison to the gait test (AUROC = 0.75), it still demonstrates on a large cohort that the sway test shows some level of disease discrimination. In a similar fashion, [115] found that in their cohort (12 HC: 13 PD) the majority of PD participants showed little visual postural disturbances, but significant sway impairment (as measured by a waist mounted IMU) leading to an AUROC of 0.93.

2.3.2 Tremor & Bradykinesia

In order to suitably quantify PD severity using tremor, it is firstly necessary to distinguish tremors that are indicative of PD, including Resting Tremor (RT) and Postural Tremor (PT), from tremors that result from other disorders, such as Essential Tremor (ET) [35]. Tremor distinction plays a vital role in guiding the appropriate course of treatment. Although many studies exist that have attempted this, tremor distinction continues to be a major challenge often leading to high misdiagnosis rates [132–134]. Nonetheless, differentiation between PD

and HC participants has still shown much promise due to the nature of PD tremor development usually being asymmetric and occurring when the limbs are stationary [135].

As for gait and postural instability assessment, there exists a set of questions within MDS-UPDRS Part III dedicated to tremor assessment, which is performed visually by a movement disorder specialist. Inertial Measurement Units are again the most common type of digital sensor used for tremor assessment, followed by smart-phones – whose usage will be discussed later in this section [80]. As tremors can occur on either the upper and lower limbs, IMUs are particularly suitable for measurement due to their unobtrusiveness and ability to be placed at multiple locations. Similarly, the MDS-UPDRS Part III contains three sections dedicated to bradykinesia assessment: finger tapping, hand opening and closing, and hand pronation/supination. A range of sensors have been used to quantify these tasks including IMUs, Musical Instrument Digital Interface (MIDI) keyboards, contact-pads, smart-gloves, and smart-phones [136–140].

Unlike the feature sets extracted for gait quantification, there exists no clinical feature set for tremor quantification. Rather, all features used to quantify tremor are technical/signal based. From the temporal domain, IMU signal features such as entropy, root mean square, and amplitude variation have shown to correlate highly with the tremor related sub-sections of the MDS-UPDRS [141]. The majority of signal based features showing correlation and disease differentiation however originate from the frequency domain [142, 143]. Through inspection of the frequency spectrum it has been shown that tremor is characterised by an increase of power at higher frequency bands [144–147]. Specifically, RT has a larger presence in the 4 - 6 Hz band whereas PT has a larger presence in the 4 – 9Hz band [148]. Alternatively, the use of discrete wavelet transformations have also shown capable of producing disease differentiating features from the wavelet-decomposed frequency spectra [149].

For the quantification of bradykinesia, both clinical and signal features are extracted from digital sensors. The most common clinical feature which is highly interpretable is the total number of taps a person makes in a defined period

of time [138, 150, 151]. This feature is widely reported to correlate with the associated MDS-UPDRS sub-section score as well as the motor-UPDRS whilst showing differentiating properties between disease groups [152, 153]. Clinical features peripheral to the total number of taps, such as tapping speed, rhythm, and fatigue are also commonly reported as showing disease differentiating ability [99, 154]. The signal based features originate from both the temporal and frequency domain and include amplitude (interpreted as opening and closing ability) and dominant frequency (interpreted as of speed tapping) [99, 155–157]. Indeed, both feature sets have even shown to correlate significantly with MDS-UPDRS measures whilst simultaneously producing higher intra-test reliability than clinician based scoring [158]. Similar to the gait tests, in a bid to quantify bradykinesia different studies have adopted different experimental protocols and sensor arrangements and thus restricting the generalisation of results [1].

Many studies have performed classification tasks using tremor and bradykinesia data. A simple threshold decision system achieved a sensitivity and specificity of 88.7% and 89.5% respectively at identifying wrist tremor on a cohort of 85 PD patients [159]. This study however collected data continuously over the course of a day and focused on the time a tremor was present, as opposed to specific tremor types. On a smaller cohort of 18 PD patients and five HCs, an impressive 87% accuracy is reported for predicting severity prediction when using a hidden Markov Model (HMM) [160]. This study utilised an extensive sensor network at both wrists, both ankles, chest, and lumbar. Using an unspecified classification technique, the power spectral density and entropy of accelerometer signals achieved a sensitivity of 80% and specificity of 90% at differentiating the tremor of 20 PD patients from the essential tremor of 20 participants [149]. There are fewer investigations into disease classification using bradykinesia tasks, and those that do exist tend to show a slightly weaker disease classification ability than tremor tasks. Using a multi-centre cohort (10 HC: 95 PD) an average disease classification accuracy of 86% was achieved through the use of a simple finger tapping task utilising a personal handheld computer [161]. In contrast, very mixed ability to perform disease

classification using tapping IMU data was reported using a SVM wherein sensitivity and specificity ranged from 0.95 and 0.83 to 0.57 and 0.79 respectively [162]. A secondary use of a SVM for classification utilised a large sensor network throughout a range of motor tests for bradykinesia [163]. The severity of bradykinesia was predicted with an average estimation error of 2.2% on a cohort of 12 PD patients when compared to clinically assigned MDS-UPDRS Part III sub-sections.

2.3.3 Speech

As people with PD predominantly exhibit visible motor impairments such as postural instability and bradykinesia, its manifestation in speech is somewhat less intuitive. However, *dysarthria* is a motor speech disorder which is symptomatic of several neurodegenerative diseases including PD [164, 165], multiple sclerosis [166], and Huntington's disease [167]. The pathophysiology of dysarthria in the case of PD stems from degradation of neurons within the substantia niga, leading to a reduction of control of the cranial nerves (specifically the vagus nerve) which control the oscillation of the vocal cords [168]. Dysarthria, since first noted in PD in 1963, has been widely reported at many stages of disease progression with a prevalence of between 70 – 90% [169, 170]. As audio recording technologies and analysis techniques have become widespread and affordable, the quantification of dysarthria through objective biomarkers has become an active area of research [86].

The primary characteristics of dysarthria within PD include a reduction in voice volume, a breathy voice, poor articulation, jerky or arrhythmic speech, and the tendency to talk in a monotonous manner [165, 171]. The standard protocol for the assessment of vocal impairment involves a patient holding a sustained vowel phonation, typically /a/ ("ahh"), for as long as possible.

A wide array of signal processing techniques have been proposed so as to extract quantitative dyphonia measures that identify dysarthria. One of the central principals to voice analysis involves determining the vocal fundamental frequency (frequency of vocal fold oscillation). Indeed, variable fundamental frequency and variable amplitude were two of the first recognised objective measures to

be observed in PD patients [169, 172]. These features were further investigated and validated using larger sample sizes and more technologically advanced recording techniques [173–175]. Subsequently, studies comparing these characteristics between PD patients and HCs discovered a significant reduction in the variability of fundamental frequency and its amplitude [176]. Variable fundamental frequency corresponds to jerky and arrhythmical speech whereas the variable amplitude commonly demonstrates a reduction in volume [176].

The techniques employed to estimate fundamental frequency from raw audio signals range from temporal domain procedures including autocorrelation [177, 178], cross-correlation [179], and peak detection [180] and frequency domain procedures such as discrete Fourier transform [181, 182], sub harmonics ratios [183], and discrete wavelet transformations [184]. Due to the availability of many fundamental frequency estimation algorithms, the use of ensemble approaches to fuse their estimates has also been proposed [168].

Upon the estimation of fundamental frequency, numerous methods for quantifying its variability have been proposed. A brief summary of the most commonly found algorithms are given below alongside references to recommended further reading:

- *Jitter* – quantifies the cycle-to-cycle duration variation of the fundamental frequency [185]
- *Shimmer* – quantifies the cycle-to-cycle amplitude variation of the fundamental frequency [180]
- *Harmonics to Noise Ratio* and *Noise to Harmonics Ratio* – quantifies the quantity of noise within the vocal recording relative to the fundamental frequency. The presence of noise in a vocal recording can be symptomatic of incomplete vocal fold closure [186]
- *Mel Frequency Cepstral Coefficients* – is an abstract frequency domain technique intended to estimate the source of the fundamental frequency which may be interpreted as studying the organs peripheral to the vocal cords. This tends to involve performing an inverse frequency domain procedures [168, 170]

Many studies have utilised these features for the purpose of disease classification and severity monitoring. The seminal work in the field achieved a classification accuracy of 91% utilising a SVM with a radial basis function on a cohort of 31 (8 HC: 23 PD) [187]. The success of this study is partially attributed to the novel feature selection strategies employed as well as the definition of a novel dysphonia feature. The dataset used in this study was subsequently made openly available and has undergone classification by other researchers. The use of fuzzy c-means clustering yields an accuracy of 96% whilst alternative SVMs, k-NN, and artificial neural networks also commonly yield accuracies above 90% [188–192]. The clear limitation of all these results is the use of a highly imbalanced dataset in which 20 of the 23 PD participants have a Y&H score that is considered 'Mild/Severe' or above. An alternative classification study using a larger dataset with a wider range of severities noted a large decrease in accuracy to 79% when utilising a Gaussian mixture model on a cohort of 23 PD participants and 23 HCs [193]. Interestingly, the only overlapping feature used by [193] and [187] were measures of the fundamental frequency (not including its variable nature). Severity prediction has also been performed using several techniques on different datasets. Using 42 participants with PD, [39] performed longitudinal regression using a random forest over the course of six months and achieved total MDS-UPDRS errors on average of 7.5 points different from those assigned by a clinician. Similarly, the authors continue to apply their methodology to a cohort of 52 PD patients and achieve errors of only 6.9 MDS-UPDRS points [184]. On another dataset of 99 PD patients, a mean absolute error of 4.6 was achieved using an almost identical feature set and a linear SVM [175].

2.3.4 Memory

Cognitive and memory impairments have been reported in all stages of PD progression and have been shown as a preceding symptom of Parkinson's disease dementia [194]. Procedural memory impairments are associated with damage to the basal ganglia and are therefore often pronounced in people with PD [195]. Procedural learning, otherwise known as implicit memory, governs the ability to learn a task

through practice [196]. Often, the task being learnt will relate to both motor and non-motor skills whose implementation requires more than theoretical learning [197].

The most common test used to assess implicit memory is the Serial Reaction Timed Test (SRTT) [198]. During the SRTT, participants are asked to respond as quickly as possible when presented with a sequence of visual stimuli. Commonly, there will be four stimuli whose positions vary spatially on a screen, and upon being presented with each stimuli the participant must respond by pressing one of four corresponding keys. Unbeknown to the participants is that the sequence of the stimuli is the same for each implementation. As such, after repeating the SRTT multiple times, one would be expected to memorise the sequence and subsequently decrease the reaction time to each stimuli. Via measuring the time taken by a participant to react to each stimuli over the course of multiple implementations, it is possible to quantify the rate of learning. Indeed, using such an approach has repeatedly demonstrated that people with PD show a significant impairment in memorising the sequence when compared to healthy controls [199–204]. In the largest meta-analysis of the SRTT utilising 13 recent studies (220 HC : 185 PD), the reaction time for HCs was found to decrease by an average of 99.1 milliseconds compared to an average decrease of 76.6 milliseconds in the PD population [197]. All 13 studies included in [197] systematically reported higher reaction times in the PD population after multiple implementations of the SRTT. All of the studies investigating implicit memory deficits in PD have focused on identifying statistically significant differences in reaction times, with none continuing to assess the ability of memory deficits to predict disease presence or severity.

2.3.5 The Capability of Smart-phones

The current state-of-the-art technique for disease quantification, which has purposely been saved for last in this review, is the usage of smart-phones. Whereas the previously discussed studies focused on assessing one aspect of PD using a single sensor type, such as an IMU *or* an audio recording device, the embedment of multiple sensors within smart-phones facilitates multiple aspects of the disease to

be assessed using a single device. Furthermore, over 3.3 billion smart-phones are in use globally with 70.8% of the United Kingdom population owning a smart-phone, a proportion of the population that has monotonically increased since records began [205, 206]. As such, leveraging the widespread usage of smart-phones presents the opportunity to provide affordable healthcare on a previously unseen scale [207]. Indeed, smart-phones are already widely used for assessing a myriad of medical disorders including mental health [208, 209], cardiac health [210, 211], and diabetes [212, 213]. It therefore comes as no surprise that smart-phones are now being explored as a potential platform for objective PD assessment; proving applicable to both the academic and industrial agendas [214].

Smart-phones have been used for multiple types of PD assessment in both clinical and remote environments. Many of the clinical based studies focused on validating the efficacy of smart-phones to replace traditional IMUs for the assessment of gait, tremor, and bradykinesia. An emerging field of work, and that which is most novel and relevant to this thesis, entails the use of smart-phones in remote environments. As the state-of-the-art disease assessment sensor, smart-phones are less widely used and therefore occupy less of the literature than the traditional sensors outlined in the previous sections. Here, an overview of smart-phone usage in clinical environments and remote environments is provided.

Of the studies utilising smart-phones for the purpose of gait quantification in a clinical environment, study objectives were either focused on finding correlations between the features extracted by smart-phones and by those extracted by traditional IMUs [215–218] or on the classification of FOG [216, 219–221]. A cohort of 49 participants performed the common clinical Timed-Up-and-Go test whilst wearing a waist mounted IMU and a smart-phone [216]. Despite the smart-phone having a smaller sampling frequency, the features showed a high level of intra-sensor reliability [0.92]. Throughout these gait clinical studies, the smart-phone placement varies and is a research interest in its own right with the most common locations being the lumbar, front trouser pocket, or ankles [222–225].

The use of smart-phones for tremor assessment in clinical environments is commonplace, often validating the ability of the IMU within smart-phones to collect temporal and frequency based features that correlate with motor-UPDRS sub-sections [226–228]. Disease differentiating features have been validated [219] wherein PD tremor presented significantly higher power at higher frequencies when compared to essential tremor, consistent with the findings of traditional IMU usage. Tremor classification has also been shown possible; on a cohort of 14 PD participants and 18 HCs, Parkinsonian tremor was differentiated from essential tremor with an accuracy of 96% using smart-phones [229]. Bradykinesia studies follow suit, presenting a range of studies successfully replacing traditional IMUs with smart-phones for identifying disease differentiating features (such as total number of taps) [230, 231] and for features which correlate with clinically assigned UPDRS [138, 139, 156, 228].

In the context of PD assessment, the true power and potential of smart-phones lies in their ability to be used remotely. This presents the opportunity for users to complete a range of tests on a much more frequent (often a daily) basis. Whereas the majority of clinical based smart-phone studies have focused on a specific aspect of the disease (such as gait), the new wave of remotely based studies collect and fuse multiple data types. Most commonly, the IMU will be used for gait, postural instability, tremor, and bradykinesia alongside the microphone being used to assess dysphonia. Furthermore, novel test types intended to assess non-motor disease symptoms have been incorporated into smart-phones. Memory and reaction tasks have been introduced wherein the touch-screen presents a suitable means of assessing short-term memory and the commonly used Serial Reaction Time Test (SRTT) [199]. Several of the data collection applications also enable users to report demographic data, medication usage, and complete sub-sections of the MDS-UPDRS Parts I and II; providing an alternative metric to the clinically assigned MDS-UPDRS.

As data from multiple test types is collected (for example gait, tremor, and voice), many papers report the classification or regression results when utilising all data types together; often without reporting the results of each test individually

[79, 232, 233]. Unlike their clinical counterpart studies, the features extracted from remotely performed tests are almost entirely focused on technical/signal based features [2, 79, 220, 234, 235]. Most striking about the results of studies using remotely collected data is the highly variable nature of their conclusions. Several remotely conducted studies report the correlations between signal features and self-assigned MDS-UPDRS as well as the ability of these features to differentiate between disease groups that are consistent with the clinical findings [228, 236]. Whilst fusing features collected remotely from gait, postural instability, voice, bradykinesia, and reaction tests a classification accuracy of 98.5% is reported in a pilot study utilising a cohort of 20 participants (10 HC and 10 PD) [99]. On the other hand, other remotely conducted studies observe a dramatic decrease in correlation strength and classification ability, with [237] reporting classification accuracies as low as 50% on a larger cohort of 46 (23 HC and 23 PD). However, the same study reported postural instability as yielding a higher classification accuracy ($\sim$60%) when compared to the gait test ($\sim$50%) – a finding that is in direct contradiction to the corresponding clinical findings. These conflicting conclusions have also been seen during the remote identification of medication response [79, 238] as well as FOG [239, 240].

The most notable benefit of the studies utilising remotely collected smart-phone data is the inclusion of high frequency longitudinal data [78, 79, 223]. This previously unseen form of data has given rise to the first personalised modelling analyses. Thus far, these studies have been entirely focused on medication response using data from the voice, tapping, and walking activities [237, 241].

More recently, smart-phones have been used for *passive* data collection [79, 223, 242, 243]. This requires no active involvement by participants; the smart-phone application continuously collects data even when the application is not actively being used. Studies utilising such data have focused on performing Human Activity Recognition (HAR) tasks, such as differentiating periods of walking from sitting or lying down, as opposed to disease classification [244, 245]. However, the vast majority of this data in unlabelled and suffers from a large degree of noise, making the extraction of disease specific features challenging [1, 79]

2.4 Discussion

Digital sensors have unequivocally demonstrated the ability to perform objective assessment of numerous PD symptoms. This chapter has outlined their usage in clinical and remote environments and highlighted their current successes and shortcomings whilst introducing the future trends and possibilities. This discussion summarises these findings with the intention of revealing the knowledge gaps needing to be addressed in order to advance the techniques of objective, robust, and clinically relevant disease assessment.

The primary and recurring limitation of studies utilising digital sensors in a clinical environment is small cohort sizes. Indeed, a recent review of studies utilising wearable sensors discovered that 70% of 136 studies had a cohort size below 30, whereas another review found 65% of 848 studies had a cohort size below 30 [20, 80]. Furthermore, many cohorts lacked a diverse range of participants in relation to disease severity; with a disproportionate number of PD participants having mild or severe symptoms [80, 187]. Due to the many studies utilising digital sensors presenting results from small cohorts with developed symptoms whilst using different sensor arrangements and experimental protocols, there is yet to be a consensus on the optimum way to measure specific symptoms. As such, it is important to perform more comprehensive studies consisting of participants at all stages of disease progression. Furthermore, in the case of body motion analysis, future studies need to utilise larger IMU networks capable of measuring symptoms at multiple locations of the body. Via inspecting which sensors are capable of robustly measuring specific symptoms, conclusions can be drawn regarding the most efficient and reproducible sensor arrangement. Finally, as clinical based studies utilising digital sensors are almost entirely cross-sectional, a significant knowledge gap exists regarding longitudinal development and variation of objective measurements. The studies that have performed longitudinal analyses consist of infrequent measurements over long time frames and have suggested changes in objective measures can be detected. The currently emerging datasets that utilise larger cohorts and sensor networks should aim to perform regular follow

up measurements with the aim of detecting objective measurements of disease development. Overcoming such limitations in clinical disease assessment are the primary objectives of the analysis performed in Chapter 5.

The use of smart-phones in clinical environments has also proven highly capable of objective disease assessment. However, their potential role for disease assessment in remote environments is still shrouded in uncertainty due to conflicting results and lack of test-specific conclusions. Many researchers have proposed that environmental noise is responsible for the decrease in disease classification accuracies and yet do not define the source of noise nor which test is most susceptible to noise. Furthermore, new and non-clinically validated tests have been incorporated into the data collection platforms which have yet to undergo examination and whose efficacy for disease assessment is unknown. An additional source of uncertainty in remotely collected datasets is the use of self-reported severity scores. Although the MDS-UPDRS Parts I and II have been designed to be completed by patients or caregivers, this has only been validated on small sample sizes and is likely to further contribute inter-rater subjectivity to the already inherent variable nature of the MDS-UPDRS scoring system. Nonetheless, participant numbers are vastly larger in remotely collected datasets with the largest such dataset containing over 8,000 participants; many of whom contribute multiple types of test data on a longitudinal basis [78]. This presents the opportunity to determine the relationships between tests and their respective symptoms, whilst tracking their longitudinal development in a non-invasive and affordable manner. The efficacy of using remotely collected data to perform these tasks is addressed in Chapters 6 and 7. Finally, the similarities and discrepancies between the quality of data collected in clinical and remote environments is yet to be formally assessed and has thus far only been subject to conjecture. The ability to transition a clinically validated objective assessment technique is yet to be implemented on a remote dataset. As such, the aim of chapter 8 is to perform a quantitative comparison between a clinically collected and a remotely collected dataset.

Although many clinical and data science related challenges exist in remotely collected datasets, their richness and diversity enable new types of disease assessment (including personalised monitoring) to be performed which may lead to the identification of novel disease characteristics.

3

Outline of Technical Methodologies

The aim of machine learning is to automate the process of pattern recognition for the purpose of making predictions. Traditionally, machine learning attempts to learn the association between a set of measured data with a corresponding set of outcomes. The measured data is composed of *observations* and *features*. From each observation, a consistent set of p features, $\mathbf{x} \in \mathbb{R}^{1\times p}$, is measured alongside the corresponding outcome, y. When data is collected from multiple observations, the measured data is assembled into a *design matrix*, commonly denoted as $\mathbf{X} \in \mathbb{R}^{N\times p}$ where N is the number of observations. Similarly, if a single outcome is measured for each observation, all outcomes are assembled into a response vector, denoted as $\mathbf{y} \in \mathbb{R}^{N\times 1}$. It is the role of machine learning to learn the function - or *model* - $f(\cdot)$ that solves $\mathbf{y} = f(\mathbf{X})$. This scenario, wherein the outcomes of each observation have been measured, is known as a *supervised learning* algorithm. Alternatively, during *unsupervised learning* the outcomes of the observations are unknown.

3.0.1 Single and Multi-Source Datasets

Machine learning in its most traditional form entails the mapping of one set of features onto one set of outcomes as described in the previous passage. However, recent technological advances have dramatically altered the techniques in which data is collected. Consequently, vastly greater quantities of disparate data types can now be collected from the same set of observations. This has given rise to a family of machine learning algorithms that utilise multiple types of measured data in order to make predictions.

In a single source dataset (Figure 3.1(a)), a single set of data, $\mathbf{X} \in \mathbb{R}^{N\times p}$, has been collected with the intention of predicting a single outcome vector, $\mathbf{y} \in \mathbb{R}^{N\times 1}$.

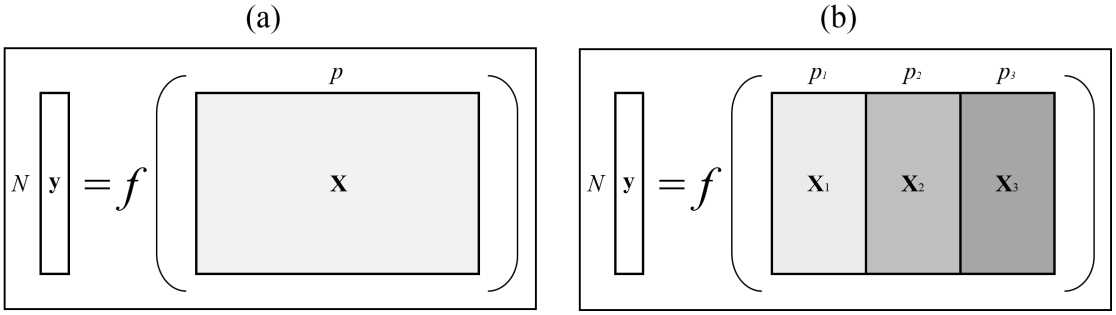


Figure 3.1: Visualisations of (a) a single source dataset composed of N observations and p features and (b) a multi-source dataset composed of N observations and three separate sources of data contributing a total of $\sum_{i=1}^{3} p_i$ features.

This is the most common form of dataset wherein the features within $\mathbf{X}$ have been collected from a single source or test. As above, the form of the machine learning task when utilising a single source dataset is $\mathbf{y} = f(\mathbf{X})$.

Conversely, a multi-source (also known as a multi-modal) dataset (Figure 3.1(b)) is one in which multiple sets of data have been collected from the *same observations* with the intention of predicting a single common response vector. Each source provides a different set of features for each observation. An example of a multi-source dataset would be the data collected during a hospital visit wherein a separate set of features comes from blood, urine, electrocardiogram, and x-ray tests. The purpose of the machine learning task utilising a multi-source dataset is to learn a single model using all of the data sources. Thus, the form of the machine learning task when utilising a multi-source dataset would be $\mathbf{y} = f(\mathbf{X}_1,\mathbf{X}_1,...,\mathbf{X}_S)$ where S is the total number of sources or tests contributed by each observation. Alternatively, it is viable to consider a multi-source dataset as a set of S single-source datasets wherein a separate $f(\cdot)$ is determined for each source individually such as $\mathbf{y} = f_1(\boldsymbol{X}_1)$.

3.1 Statistical Approaches to Data Analysis

There are two stages of statistically analysing a dataset: exploratory and confirmatory. During exploratory analyses, the underlying characteristics of the data are revealed in order to motivate the future direction of analysis. This commonly is

performed through visualisation of the distributions within and the relationships between different components of a dataset. Once a research objective and hypothesis has been formulated (based on the results of an exploratory analysis), a confirmatory analysis is used to either prove or disprove the hypothesis; usually through the use of a statistical hypothesis test.

3.1.1 Correlation Analysis

Correlation analyses are used for determining the association or dependence between sets of random variables and is considered an exploratory analysis. This relationship is usually quantified through the correlation coefficient which determines the joint variability of the random variables. Most frequently reported for continuous-valued data is Pearson's correlation coefficient [246], wherein the covariance of the two measurements is scaled by their standard deviations:

$$\mathrm{Corr}(X,Y)_{Pearson} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sum_n^{i=1}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_n^{i=1}(x_i-\bar{x})^2\sum_n^{i=1}(y_i-\bar{y})^2}} \tag{3.1}$$

where X and Y are the two sets of random variables, and $\bar{x}$ and $\bar{y}$ are their mean values respectively. Pearson correlation coefficients are always within the range [-1 +1]. A correlation coefficient of +1 indicates a perfectly proportional and linear relationship between the measurements whereas a correlation coefficient of -1 indicates a perfectly inversely proportional and linear relationship. A correlation coefficient of 0 indicates no covariance exists between the measurements and they are independent.

Alternatively, Spearman's rank correlation coefficient determines the monotonic relationship between two measurements [247]. This correlation approach is therefore less concerned with the similarity in value of two random variables, and more concerned with whether an increase in one measurement corresponds to an increase in the other measurement. Formally:

$$\mathrm{Corr}(X,Y)_{Spearman} = 1 - \frac{6\sum_{i=1}^{n}(x_i-y_i)^2}{n(n^2-1)} \tag{3.2}$$

The range of the Spearman rank correlation coefficient is also [-1 +1] with positive coefficients demonstrating a proportional relationship and a negative coefficients demonstrating an inversely proportional relationship.

3.1.2 Hypothesis Testing

Hypothesis tests are performed in order to statistically confirm or reject a hypothesis regarding a set of collected data and therefore is a form of confirmatory analysis[248]. A *null hypothesis* is defined concerning an expected finding within the collected data, such as whether two samples originate from the same distribution [249]. Performing a hypothesis test produces a significance value, referred to as a *p-value*. A hypothesis test result is considered to be *significant* if it has been deemed to have not occurred due to chance, as governed by a threshold *p-value* known as the significance level. Commonly, the highest significance level used is 0.05, representing a probability of 5% of getting the observed value of the test statistic, or a value with greater evidence against the null hypothesis.

There are a myraid of hypothesis testing techniques, all of which are accompanied by a specific set of assumptions regarding both the data and the hypothesis. However, hypothesis tests can be divided into two key types: parametric and non-parametric [250].

Parametric approaches are the most widely used hypothesis testing techniques and inherently assume an underlying distribution to the data. Examples of parametric hypothesis tests include the independent t-test and the paired t-test [251]. The former of these methods would be suitable to test whether two random variables (both of which are assumed to be Gaussian distributed) are likely to belong to the same population distribution. Alternatively, the paired t-test is more applicable to testing whether a distribution has changed significantly; such as whether the distribution of a random variable is altered due to a external factor.

Conversely, non-parametric hypothesis tests do not assume a prior knowledge of the distribution of the data. Examples of non-parametric hypothesis tests include the Kolmogorov-Smirnov test and the use of bootstrap sampling [252,

253]. The former of these methods directly compares the similarity between the cumulative distribution functions of two random variables, making no prior assumption regarding their distribution. Bootstrap sampling entails randomly sampling each measurement set separately (with replacement) and calculating a test statistic, commonly referred to as a bootstrap statistic. Consequently, when this is repeated many times new distributions are formed of the bootstrap samples enabling the use of either parametric or non-parametric hypothesis tests on the new distributions. Bootstrap sampling is particularly well suited for hypothesis tests wherein few (under 30) samples are present [254].

3.2 Standard Machine Learning Models

Supervised machine learning tasks can be divided into two types depending on the form of the outcome measure. If the outcome measure is discrete (such as male/female or diseased/not diseased) the problem is one of *classification* where the machine learning task attempts to correctly categorise each of the observations. Conversely, if the form of the outcome measure is continuous (such as height or blood pressure) the problem is one of *regression* where the machine learning task attempts to estimate the value of the continuous outcome. In unsupervised learning, where no outcomes exist, techniques such as clustering are commonly employed such as to group similar sets of observations.

3.2.1 Linear Regression

To introduce the model notation used throughout this work, the linear regression model is presented as the simplest traditional regression technique [255]. A general linear regression model containing p dependent variables is defined as:

$$\begin{aligned} \hat{y} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \\ &= \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta} \end{aligned} \tag{3.3}$$

where $\hat{y}$ is the model prediction, $\boldsymbol{x} = [x_1, x_2, ..., x_p]$ are the feature values, $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]$ are the model weightings of each feature, and β_0 is the model intercept.

The model weights and intercept are easily optimized via the minimization of the residual sum of squares between the model's estimated response, $\hat{y}$, and the true response, y:

$$\arg\min_{\boldsymbol{\beta},\beta_o} \sum_{i=1}^{N} (\hat{y}_i - y_i) \tag{3.4}$$

which when substituting in Equation 3.3 becomes:

$$\arg\min_{\boldsymbol{\beta},\beta_o} \sum_{i=1}^{N} (\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta} - y_i) \tag{3.5}$$

which, when assuming that the number of observations is larger than then number of features, can be solved algebraically.

3.2.2 Logistic Regression

Logistic Regression (LR) defies its namesake in that it is a classification technique. It is a probabilistic discriminative model that determines the probability of a feature vector as belonging to a discrete outcome. The key difference between logistic regression and linear regression is the choice of prior placed on the model output. Via the use of the logistic function, the output of logistic regression is bounded between 0 and 1. The logistic function (otherwise known as the Sigmoid function) is defined as:

$$\sigma(x) = \frac{e^x}{1 + e^x} \tag{3.6}$$

for the independent variable x [256]. Formally, we can define a logistic regression model as:

$$\sigma(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}) = p(y = 1|\boldsymbol{x}, \beta_0, \boldsymbol{\beta}) = \frac{e^{\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}}} \tag{3.7}$$

where β_0 is the model intercept and $\boldsymbol{\beta}$ are the feature weightings.

The $\boldsymbol{\beta}$ coefficients are determined using Maximum Likelihood Estimation (MLE) of the training data. The i^{th} training point has a label, y_i and a corresponding feature vector, $\mathbf{x}_i$. As such, the probability of all N training points given a set of coefficients is given by:

$$L(\beta) = p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i)^{y_i}(1 - p(\boldsymbol{x}_i))^{1-y_i}) \tag{3.8}$$

For mathematical convenience the coefficients that optimize the log-likelihood are found via taking the natural logarithm of Equation 3.8 and substituting in Equation 3.7, thus yielding:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}) - \ln(1 + \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}) \tag{3.9}$$

which can be solved through finding the $\boldsymbol{\beta}$ which causes the first derivative to be equal zero:

$$\frac{\delta\mathcal{L}(\boldsymbol{\beta})}{\delta\boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{x}_i(y_i - p(\boldsymbol{x}_i)) = 0 \tag{3.10}$$

As Equation 3.10 is transcendental, its solution is approximated numerically.

3.2.3 Regularisation & Feature Selection

Regularisation is widely employed across many ill-conditioned ($p > N$) machine learning applications so as to reduce the degree to which a model 'overfits' to the training data. Overfitting occurs when a model's parameters are highly sensitive to the training data which often leads to poor generalisation and model performance when applied to new test data. Addition of a regularisation term to a model's loss function during the determination of the coefficients results in a less complex and more general model. Furthermore, regularisation can serve a second purpose in the form of feature selection. Feature selection is an essential component to the successful development of classification and regression models and often improves the performance and interpretability of a model. The inclusion of redundant features during model development is associated with several detrimental factors including overfitting and increasing the computational complexity.

In this section, two regularisation techniques that are commonly utilised during the development of linear regression and logistic regression models are outlined alongside their respective capabilities of performing feature selection.

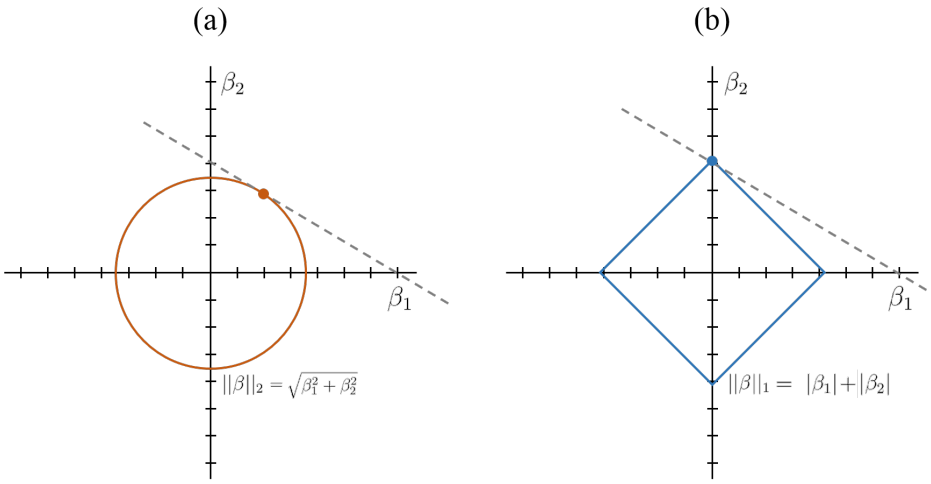


Figure 3.2: A schematic of the ℓ_1 and ℓ_2 regularisation approaches. The grey dotted line corresponds to an isobar of constant model loss as achieved at each combination of β_1 and β_2. The model parameters that correspond to this loss, whilst constrained by the ℓ_1 and ℓ_2 regularisation terms, are shown by the red and blue circles respectively. It is evident that in plot (a) sparsity is not induced as is the case in plot (b) where $\beta_1 = 0$.

Tikhonov Regularisaion

Tikhonoc Regularisation, commonly referred to as Ridge regularisation, utilises an ℓ_2-norm regularisation term [257, 258]. Figure 3.2(a) demonstrates how the geometric interpretation of the ℓ_2-norm corresponds to the Cartesian distance from the origin of two parameter values. The level of regularisation introduced to the model is governed by the *shrinkage parameter* variable, usually denoted as r. During optimisation, the sum of absolute error penalty ($\sum_i^n (y_i - \hat{y}_i)^2$) is minimised subject to the shrinkage constraint $||\boldsymbol{\beta}||_2 \leq r$. Ridge regression, expressed in Lagrangian form is

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\mathbf{y} - \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}||_2^2 + \lambda||\beta||_2 \tag{3.11}$$

where λ is the Lagrangian representation of the shrinkage constraint. As such, via increasing the value of λ, the model parameters are subject to a larger regularisation constraint leading to larger model coefficients of more flexibility [259].

Least Absolute Shrinkage and Selection Operator

Conversely, whereas Ridge regularisation utilised the ℓ_2-norm, Least Absolute Shrinkage and Selection Operator (LASSO) regularisation utilises a ℓ_1-norm regularisation term. Not only is LASSO a popular technique for regularisation, but it is also found to induce sparsity into model parameters and therefore inherently performs feature selection [260–262]. Figure 3.2(b) demonstrates how the geometric interpretation of the ℓ_1-norm corresponds to the absolute distance from the origin of two parameters. Again, a shrinkage parameter is used to determine the amount of regularisation and sparsity imposed on the model parameters. The LASSO, expressed in Lagrangian form is:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\mathbf{y} - \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}||_2^2 + \lambda||\beta||_1 \tag{3.12}$$

where λ is the Lagrangian representation of the shrinkage constraint. As sparsity is promoted by use of the ℓ_1-norm, there will be values within $\boldsymbol{\beta}$ that are zero. Feature selection can therefore be performed via extracting the dense feature set (all features whose $\beta \neq 0$).

3.2.4 Random Forests

A Random Forest (RF), more formally known as an ensemble of decision trees, is a non-parametric classification and regression technique whose theory is based on the traditional decision tree [263]. A decision tree aims to divide a feature space into a pre-specified number of sub-spaces via recursively finding the best split between feature values given their corresponding labels. Each split is formally referred to as a node, with each node determining an optimal threshold for splitting a single feature. If using a decision tree for classification, each of the final sub-spaces are assigned a binary label based the majority class labels of the training observations present in the sub-space. Conversely, if a decision tree is being used for regression purposes, the sub-spaces are assigned the average of the continuous responses of the training observations present in the sub-space.

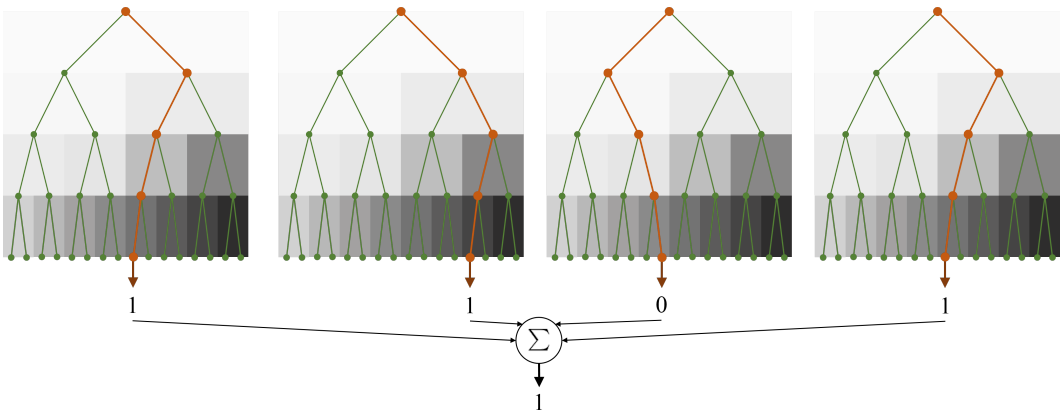


Figure 3.3: A schematic of the Random Forest classification procedure. Multiple decision trees are trained using random subsets of features and observations with the branches of each tree dividing the original feature space into a pre-defined number of sub-spaces. Finally, the binary predications made by each tree are aggregated to yield the final classification prediction.

However, RFs are populated with multiple decision trees, each using a random subset of features and a random subset of observations from the original training set (Figure 3.3). During the training of a RF, a random subset of observations and features will be selected to train a single decision tree or a *base learner*. Multiple base learners will be trained with each using a different subset of observations and features. Upon completing the training of all the base learners, their responses are aggregated or *ensembled*. Ensemble learning entails fusing the responses of multiple predictions in the hope that multiple predictions will be more reliable than a single prediction. Ensemble learning, and a number of its specific techniques, will be discussed in further detail later in this chapter (§3.4).

The primary benefit of using a RF over a traditional decision tree is the decreased chance of model overfitting; as decision trees are prone to, especially in the case of noisy data [264]. However, the number of trees and the number of features utilised by each tree are hyper-parameters specific to each RF model. Nonetheless, attempts to generalise the number of features within each tree have been made with $\sqrt{p}$ features having shown to produce generally competitive results [265]. An additional benefit of RF is their inherent ability to perform feature selection [266]. Each feature is assigned a level of importance depending on their contribution to successfully splitting a

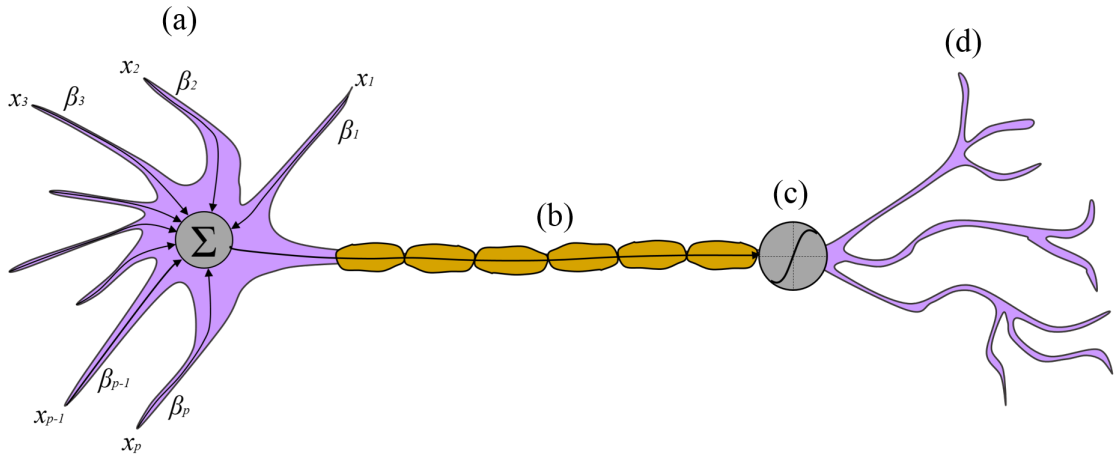


Figure 3.4: The essential biological components of a single neuron demonstrating each of their corresponding mathematical operations. (a) dendrites/inputs, (b) axon/transmitter, (c) activation function, and (d) output.

node. Further, as each tree utilises a different subset of features, this enables a comprehensive comparison of importance between many combinations of features.

3.3 Neural Network Models

3.3.1 Fundamental Neuron Theory

One would be remiss to not acknowledge the irony of applying neural networks (NNs) during the study of a neurodegenerative disease. The foundation of NN theory can be interpreted as replicating the behaviour of biological neurons. The most basic building block of a NN model is a single neuron - or *unit*. Figure 3.4 shows a schematic that relates the biological components of a single neuron to their equivalent mathematical operations. The four key components, as labelled in Figure 3.4 are summarised as:

a *The dendrites* provide the inputs to the neuron which are combined in the cell nucleus. In the mathematical neuron model each input receives a weight and the nucleus represents a summation operation for all the weighted inputs.

b *The axon* acts an an electrical transmitter along which the weighted summation of all inputs is transferred from the nucleus to the activation function.

c *The activation function* receives the weighted sum of the inputs and determines whether the neuron will 'fire' i.e. produce an output. The activation function governs the magnitude of the output as a function of the weighted sum of the neuron inputs.

d *The output* of the neuron, as determined by the activation function, can then either be fed into further neurons or can be the final model output.

This process is denoted mathematically as:

$$\hat{y} = \varphi(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}) \tag{3.13}$$

where $\varphi(\cdot)$ is the activation function. One should note that if the logistic function is used as the activation function in a single neuron model, the NN model reduces to Equation 3.7 and is a simple logistic regression model.

When first conceived in the 1940's the use of NNs was restricted by the low level of computational power available and the restricted availability of large data sets. Accordingly, the dramatic increase of available computational power has been accompanied by an rise in NN usage. This recently found popularity has given rise to many new NN models being proposed, most notably including the invent of *deep learning* . Deep learning is a broad term used to describe a range of a NN architectures. Deep neural networks (DNNs) and Convolutional Neural Networks (CNNs) are two state-of-the-art deep learning architectures which are used throughout this thesis and are described in the following sections. The reader should also be aware that alternative deep learning architectures are also commonly used such as Recurrent Neural Networks (RNNs) and Deep Belief Networks (DBNs). However, these are beyond the scope of this thesis and the reader is directed to [267–270] for a description and comparison of these architectures.

3.3.2 Deep Neural Networks

Deep Neural Networks, often referred to as deep feedforward networks or multilayer perceptions, are the foundation and simplest form of deep learning [269–271]. They

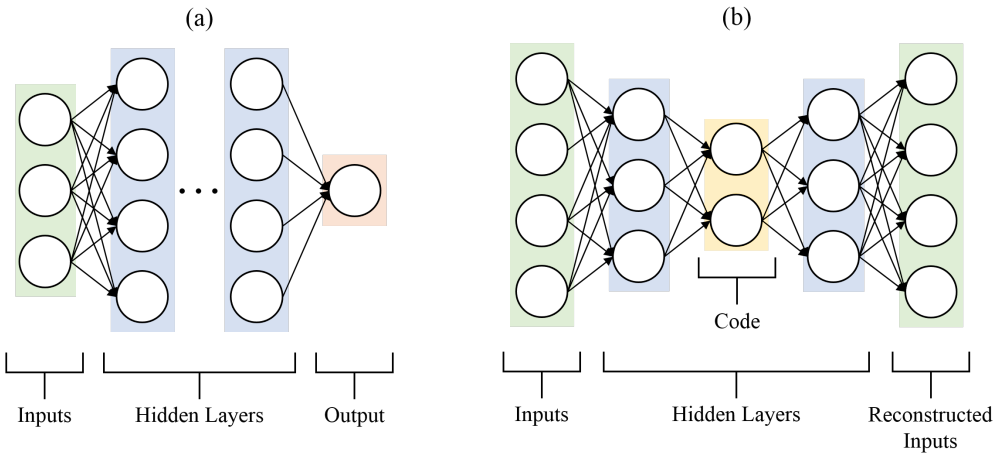


Figure 3.5: Example architectures of (a) a general deep neural network for the purpose of classification and (b) a deep autoencoder with three hidden layers. Each white circle represents a single neuron as depicted in Figure 3.4.

are a direct extension from artificial neural networks, differing only in the number of hidden layers present in the network. Figure 3.5(a) shows the three key components of a deep neural network architecture. A neural network is formed via stacking multiple layers of connected units. In a fully connected network, such as that in Figure 3.5(a), the inputs of the output layer and hidden layers are the outputs from all units from the previous layer. A network is described as 'deep' if there are many, usually more than three, hidden layers of neurons between the inputs and outputs.

The inputs to the network, in a traditional deep neural network, will be a feature vector. Upon passing through the first hidden layer, these features are transformed into a more abstract representation, commonly in a non-linear fashion through the use of a non-linear activation function. The use of multiple hidden layers has shown to allow a large amount of flexibility for the network to map a set of inputs to an output.

3.3.3 Neural Network Autoencoders

A special usage of a DNN is that of deep neural network autoencoders. A deep neural network autoencoder is an unsupervised technique primarily used for performing non-linear dimensionality reduction [271–273]. Autoencoders firstly utilise an encoding

network which condense the network inputs, $\boldsymbol{X}$, into a smaller representation referred to as *the code*, denoted as $\boldsymbol{Z}$. Mathematically,

$$Z = \varphi(\boldsymbol{b} + \boldsymbol{W}_e \boldsymbol{X}) \tag{3.14}$$

where $\varphi(\cdot)$ is an activation function[1]. Secondly, a decoding network is used to reconstruct the input space using the encoded representation or *code*.

$$X' = \varphi(\boldsymbol{b} + \boldsymbol{W}_d \boldsymbol{Z}) \tag{3.15}$$

Thus, an autoencoder is a simply a symmetrical deep neural network whose central hidden layer is of a lower dimensionality than the input layer and whose number of input units are equal to the number of output units (Figure 3.5(b)). During training, the reconstruction error between the inputs and the autoencoders reconstruction of the inputs is minimised.

Once the autoencoder has been optimised such that the reconstuction error is at its minima, dimensionality reduction is performed via replacing the input feature values with their corresponding code vectors.

Furthermore, deep autoencoders have also been widely used for removing noise from raw signals and feature sets [274]. When presented with multiple signals, some of which contain noise, deep autoencoders will reconstruct all signals using a set a joint model weights that attempt to best preserve the shared signal characteristics most present in all of the signals.

3.3.4 Convolutional Neural Networks

The defining idiosyncrasy of Convolutional Neural Networks (CNNs) is that they replace at least one traditional matrix multiplication with a convolutional operation [267]. The convolutional operation most commonly employed in CNNs is the discrete convolution operator, defined as:

[1]For neural network models, the weighting matrix/tensor traditionally is notated as $\boldsymbol{W}$ as opposed to the weighting vector $\boldsymbol{\beta}$ for a single neuron model.

$$s(t) = \sum_{a=-L}^{L} x(a)w(t-a) = (x * w)(t) \tag{3.16}$$

where x is the input data of length $2L$ and w is a weighting function or *kernel*. A convolution operation is identical to determining the cross-correlation between a signal and a kernel except for that in convolution the kernel has been time-reversed. However, the input used in Equation 3.16 is a simple 1-dimensional signal. This convolutional operation can be scaled up to function on multi-dimensional inputs such as images and videos [267, 275, 276]. During the training of a CNN model, the weights that form each kernel are learnt simultaneously with the weights of the other model parameters such as the weights of additional neurons in a DNN. As such, CNNs facilitate the use of raw data as an input to a neural network model without the prior need to manually define and extract features, as is required for the vast majority of other machine learning models.

Whereas traditional matrix multiplication procedures, such as those utilised in a fully connected DNN, require an interaction between all input units with all output units (Figure 3.5), CNNs possess the ability to circumvent this inefficient procedure via the use sparse parameter representations. Sparse parameter representations are achieved via using a kernel whose dimensionality is smaller than that of the raw input. Consequently, each input unit will only interact with a number of output units equal to the size of the kernel [277, 278]. This trait allows CNNs to learn considerably fewer parameters. Although this sparsity results in fewer direct connections between units, when multiple convolutional layers are utilised the number of indirect connections increase as shown in Figure 3.6.

Furthermore, as a kernel is utilised at all positions of the input, each weighting of the kernel will be multiplied by each of the inputs elements. Not only does the use of the same kernel on all input elements give rise to highly efficient parameter storage through parameter sharing, it also gives rise to one of the CNNs most powerful characteristics - *translational invariance* [267, 279]. Translational invariance allows the outputs of a CNN to vary in a consistent manner when presented with different inputs whose elements are in different orders. As such, a kernel trained for identifying

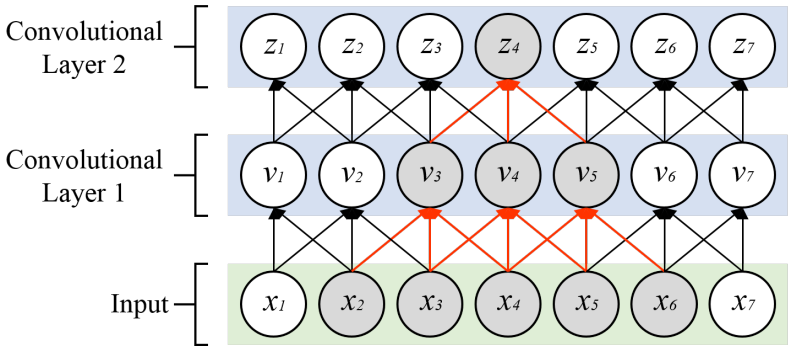


Figure 3.6: Visualisation of the sparse parameter representation and indirect connections of a convolutional neural network. In this example a kernel of width three is used. The sparse parameter representation is demonstrated by each of the input units connecting to only three units in the first convolutional layer. The indirect connection between node z_4 in the second convolutional layer and the input nodes x_2 to x_6 are shown in red. Figure motivated by [267].

a specific aspect of an input will be suitable to use on an input wherein the specific aspect appears at a different location.

Finally, an additional concept widely employed in CNN architectures is pooling. As its name implies, pooling entails combining multiple outputs from a convolutional layer into a single output. This is achieved through the use of a pooling function; for which many exist such as max-pooling, mean-pooling, or weighted average pooling. The use of pooling further contributes to the translational invariance of the convolutional weights whilst also reducing the number of parameters needing to be learnt during the training phase [267, 280].

3.3.5 Key Neural Network Concepts

There are several key concepts that must be carefully considered when implementing the aforementioned NN architectures.

Activation Functions

Activation functions are primarily used in neural networks to induce non-linearity into each of the individual neurons [281]. Many different types of activation functions exist and are chosen according to the specific type of network, layer, and desired output.

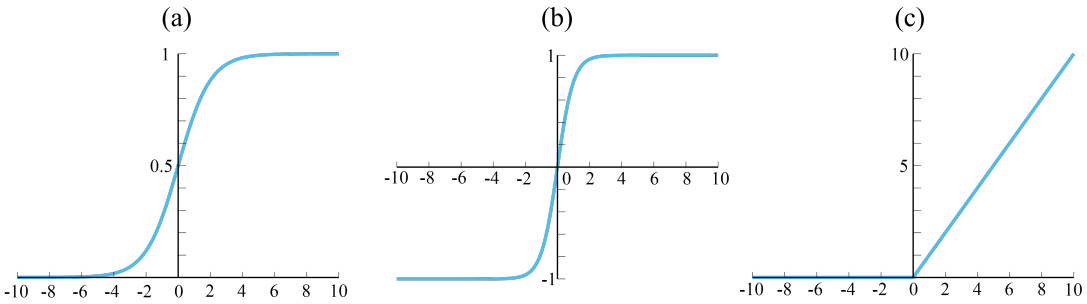


Figure 3.7: The three most commonly used neural network activation functions (a) Logistic, (b) Tanh, and (c) Rectified Linear Unit (ReLU).

Three of the most popular and non-linear activation functions are provided in Figure 3.7. Figure 3.7(a) depicts the logistic (also known as the softmax) activation function which is historically the most popular. However, the primary limitation of the logistic activation function is that at extreme values the gradient tends to zero; a trait that proves problematic during network optimization [282]. Figure 3.7(b) depicts the hyperbolic tangent function ($tanh(x)$). Unlike the logistic function, the hyperbolic tangent is capable of producing negative values, making it a popular choice in deep autoencoders and recurrent neural networks. For example, if one were to use a deep denoising autoencoder utilising a logistic activation to de-noise a corrupted sine wave, the autoencoder could only represent the sine wave values that fall above 0. Conversely, the use of the *tanh* activation function would enable the autoencoder to output both the positive and negative de-noised outputs. However, the *tanh* suffers from the same vanishing gradient limitation as the logistic function. Recently, the Rectified Linear Unit (ReLU) has become the most popular activation function, defined as $g(z) = max(0, z)$, and is depicted in Figure 3.7(c). The popularity of ReLU lies in its usage dramatically decreasing learning times of deep networks whilst also reducing the likelihood if inducing a vanishing gradient [283].

Dropout Regularization

A unique form of regularisation exists for neural network algorithms which has proven more successful than traditional regularisation techniques. Dropout regularisation is based on the principal that via repeatedly randomly omitting random units from

a neural network during different epochs of the training phase, each neuron becomes less sensitive to noise [276]. This process entails training many 'thinned' neural networks, each of which has a different set of randomly omitted units. At the end of the training process, the weights determined by each of the thinned networks are averaged and used in the original unthinned network [284]. This is found to produce smaller unit weights which have a strong ability to prevent over-fitting. The rate at which units are omitted varies, with the literature default being 50%. It should be noted that at a dropout rate of 50%, roughly double the number of iterations are needed for the model to converge during training [284].

Gradient-Based Optimization

The optimal weights of each unit within a neural network are determined through optimization of a loss function. This entails determining the optimal set of model weights that minimises the error between the output of the network and the true response presenting a highly non-convex and challenging problem. In the case of all neural network models, the optimisation procedure utilises the gradient descent algorithm.

Gradient descent is an iterative process which aims to find the minima of an objective function, denoted as $J(\boldsymbol{W})$ [285]. This is achieved via calculating the derivative of $J(\boldsymbol{W})$, with respect to the model weights, and updating the model weights such that the objective function decreases. Formally:

$$\boldsymbol{W}' = \boldsymbol{W} - \eta \nabla_{\boldsymbol{W}} J(\boldsymbol{W}) \tag{3.17}$$

where $\boldsymbol{W}$ are the weights of the current model iteration, $\boldsymbol{W}'$ are the updated weights to be used in the following model iteration, $\nabla_{\boldsymbol{W}} J(\boldsymbol{W})$ is the derivative of the objective function using the current model weights, and η is a specified scalar referred to as the learning rate which governs the amount the model weights are altered between each iteration.

The objective function is formed via determining the total model error through the use of a loss function:

$$J(\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} L[y_i, f(\boldsymbol{x}, \boldsymbol{W})] \tag{3.18}$$

where $L[y_i, f(\boldsymbol{x}, \boldsymbol{W})]$ is determining the loss between the model prediction, $f(\boldsymbol{x}, \boldsymbol{W})$, and the true response, y_i, for the i^{th} observation. Many loss functions exist and are chosen depending on the type of model being developed, with the majority of recent neural networks utilising cross-entropy loss functions. When performing binary classification, the use of the binary cross-entropy loss function is the most widely employed and is defined as:

$$L[y, f(\boldsymbol{x}, \boldsymbol{W})] = y \cdot \log[f(\boldsymbol{x}, \boldsymbol{W})] - (1 - y) \cdot \log[1 - f(\boldsymbol{x}, \boldsymbol{W})] \tag{3.19}$$

However, large datasets have proven to be a double edged sword. Although large quantities of training data help create more general models that are less prone to over-fitting, they also require larger computational power to train and optimize. The latter of these points has proven particularly poignant in the use of traditional gradient descent; often rendering the technique excessively time consuming for model optimisation.

As such, alternative variations of gradient descent have been proposed to overcome these limitations with the most popular being that of stochastic gradient descent [285]. Stochastic gradient descent dramatically reduces training time via repeatedly approximating the optimal model on many subset samples, or *mini-batchs*, of the original dataset. An estimate of the gradient is produced via:

$$J(\boldsymbol{W}) = \frac{1}{m'}\nabla_{\boldsymbol{W}} \sum_{i=1}^{m'} L[y_i, f(\boldsymbol{x}, \boldsymbol{W})] \tag{3.20}$$

where m' is the number of observations in the subset. Consequently, the weights determined on each mini-batch are averaged and used in the final model. A particularly popular stochastic gradient descent algorithm is *Adam* on account of its capability of producing adaptive learning rates (Equation 3.17) whilst being memory efficient and therefore suitable for very large datasets [286].

3.4 Ensemble Learning

Ensemble learning, often referred to as classifier fusion, is a technique used for fusing the responses of multiple classifiers. Their popularity and strength lie in their ability to account for variability within each of the separate classifiers that form the ensemble. Accordingly, the outputs of classifier ensemble techniques are commonly found to outperform that of a single classifier in terms of both accuracy and error rate [287–289].

The underlying theory behind ensemble learning is that a consensus decision on multiple estimates is more likely to produce a correct estimate than if only a single estimate were used. Suppose a classifier is trained on a dataset containing a degree of noise; as such the classifier will posses some level of error. When presented with a test observation, this classifier will make a prediction which will be accompanied by the underlying error of the classifier; as in Figure 3.8(a). If multiple classifiers are subsequently trained using the same dataset, when presented with the same test observation, it is expected that the classifiers will possess some agreement or overlap in their decisions; as in Figure 3.8(b). Their area of agreement is typically smaller than each of their individual error rates, thus producing a combined prediction with a smaller chance of error.

The individual classifiers that are combined in an ensemble are referred to as the 'base' classifiers. Many ensemble procedures exist which can be used on the 'soft' responses of classifiers (i.e. the posterior probabilities) or on the 'hard' responses of classifiers (i.e. the discrete classes). The most widely used classifier ensemble strategies are majority voting and probability averaging.

Suppose B base classifiers have been trained and their responses for a single observation are to be ensembled. The ensemble learner $\mathcal{F}$ is defined as:

$$\mathcal{R} = \mathcal{F}(R_i, ..., R_B) \tag{3.21}$$

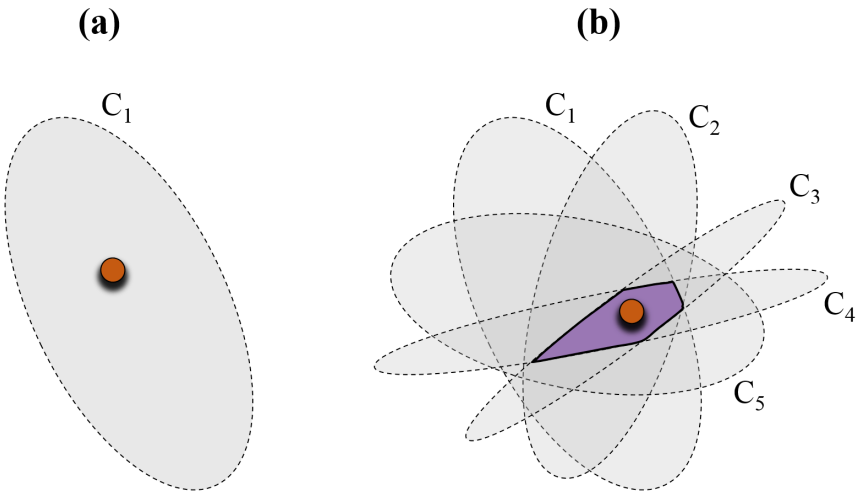


Figure 3.8: A schematic of the fundamental conept of ensemble learning wherein the overlapping predictions of multiple classifiers results in a smaller region of error.

where R_i is the response of the i^{th} classifier and $\mathcal{R}$ is the ensemble response. Majority voting is implemented on the hard responses of each base classifier and is defined as:

$$\mathcal{F} = \begin{cases} 1 & \text{if } \sum_{i=1}^{B} R_i \geq \frac{B}{2} \\ 0 & \text{if } \sum_{i=1}^{B} R_i < \frac{B}{2} \end{cases} \tag{3.22}$$

Majority voting therefore returns the discrete response that occurs most frequently between the base classifiers for a single observation.

Alternatively, the popular unweighted average probability ensemble is performed on the soft response of each classifier and is defined as:

$$\mathcal{F} = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^{B} P(R_i=1)}{B} \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3.23}$$

The mean probability ensemble therefore returns a response based on the average probability of all classifiers for a single observation.

Alternative and more complex ensemble strategies exist such as boosting (multiple classifiers training using resampling of the training data) and bagging (bootstrap aggregating). Recall from §3.2.4 that Random Forest classifiers are formally an ensemble of multiple decision tree classifiers often using either boosting or bagging, with each tree being a base learner that provides a single prediction.

3.5 Evaluation Approaches and Metrics

3.5.1 Performance Generalisation

Performance generalisation refers to the ability of a machine learning model to be successfully used on previously unseen observations. Cross validation is widely used in statistics and machine learning a means of assessing a model's generalization ability and whether over or under-fitting is occurring [290].

In k-fold cross-validation, the N observations in a dataset are randomly assigned into one of k approximately equal sized groups, herein referred to as folds. A model is then trained on $k-1$ folds, with the single fold not used during training being set aside against which the model performance can be assessed. This procedure is repeated such that each fold has been used as a validation set exactly once. As such, the metrics of each model can be reported alongside a measure of their variability. When this process is repeated many times, such as to maximise the level of statistical confidence, it is called repeated k-fold cross-validation

Several variations of k-fold cross-validation exist including Leave-one-out-cross-validation (LOOCV) and stratified cross-validation. In the former's case, the number of folds is set equal to the number of observations. This results in all but one observations being used for model training, leaving a single observation to be used for validation. This method is particularly popular in smaller datasets and yields a low bias but usually a high variance evaluation. In the latter's case, the k folds are created such that the balance of classes within each fold is approximately equal to the balance of classes in the entire dataset. This is particularly popular in imbalanced datasets and maintains a consistent level of model bias across all training procedures [291].

3.5.2 Classification Model Metrics

In order to evaluate the performance of a binary classification model, the model predictions must be compared against the corresponding true values. Many evaluation techniques exist with the most common and intuitive evaluation metric

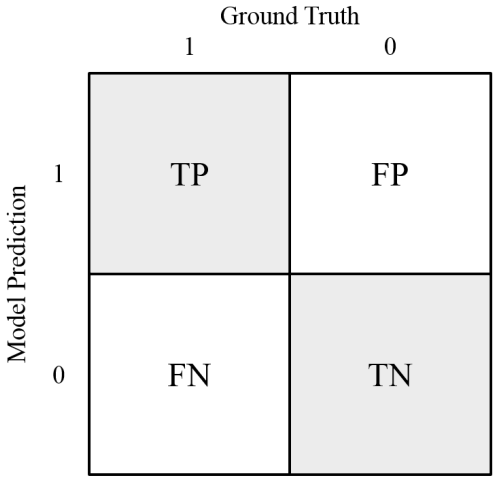


Figure 3.9: A confusion matrix demonstrating how the TP, TN, FP, and FN metrics are calculated between the predictions of a model and the true responses.

being the classification accuracy (Acc) which states the percentage of model predictions that are the same as the corresponding true values and is defined as:

$$\text{Acc} = 100 \times \frac{TP + TN}{TP + FP + FN + TN} \tag{3.24}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative rates respectively. Figure 3.9 depicts a confusion matrix demonstrating how each of these metrics are calculated.

Additional popular metrics are the sensitivity (Sen) and specificity (Spc) which are defined as:

$$\text{Sen} = \frac{TP}{TP + FN} \quad ; \quad \text{Spc} = \frac{TN}{TN + FP} \tag{3.25}$$

While sensitivity states the proportion of *positive* points correctly classified, the specificity states the proportion of *negative* points correctly classified.

Finally, the sensitivity and specificity metrics are commonly combined into a single metric called the F_1 score which is defined as:

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3.26}$$

3.5.3 Regression Model Metrics

Whereas classification evaluation techniques utilised discrete model predictions, the evaluation of regression models use continuous model predictions. A widely used regression model metric is the Mean Absolute Error (MAE) which calculates the mean magnitude of error between the predictions and the ground truth and is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.27}$$

An additional popular metric is the Root Mean Squared Error (RMSE) which instead operates on the squared magnitude of error between the predictions and the ground truth and is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{y}_i]^2} \tag{3.28}$$

The RMSE is always equal to or larger than the MAE and its final value is highly influenced by points possessing a large error.

4

Definition of Datasets and Data Types

4.1 Introduction

This thesis makes use of two datasets that are juxtaposed by collection environment, collection devices, and activity types. Firstly, a dataset collected entirely in a clinical environment as part of a longitudinal clinical trial will be introduced with respect to the participant demographics, activity types, and data quality. This study utilises a large array of Inertial Measurement Units (IMUs) as part of routine clinical care with the objective of identifying clinical body motion features capable of monitoring disease progression. The second dataset was collected entirely in a remote environment using smart-phones and is the largest open-access dataset currently available. This dataset collects data using the wide array of sensors embedded within smart-phones as a means of assessing multiple aspects of PD; many of which have undergone little study in remote environments. Finally, a discussion of each dataset is provided including a comparison with respect to their data quality, participant demographics, and limitations.

4.2 The OxQUIP Study

Study Background

The Oxford study of Quantification in Parkinsonism (OxQUIP), is a large clinical observational study being conducted at the John Radcliffe Hospital, Oxford. The study follows participants during their routine clinical care over the course of three years with the objective of identifying measures that can detect disease progression. In addition to detecting progressive disease measures, the OxQUIP study also aims to objectively evaluate the efficacy of different treatment types. This is achieved

via recruiting a large number of participants on different types of treatment and monitoring their symptoms in different treatment states. A unique characteristic of the OxQUIP participant cohort is the inclusion of atypical parkinsonian participants who have been diagnosed with Progressive Supranuclear Palsy (PSP). As such, the overarching aim of the study is to objectively monitor the progression and effect of treatment on symptoms within a cohort of typical and atypical parkinsonian participants on a longitudinal basis.

During routine clinical visits, test data is collected from participants using wearable sensors in the form of bracelets or anklets which wirelessly transmit accelerometer and gyroscopic waveforms to clinical analysis software which produces the objective measurements. The study utilises a network of wearable sensors at several locations on the body, facilitating a comprehensive assessment of movement behaviour.

The study began recruiting participants in November 2016 and is currently still active. The OxQUIP study has been reviewed and received ethical approval by the South West - Cornwall & Plymouth Research Ethics Committee, Clinical Trials and Research Governance (University of Oxford), and the Scientific committee of the Union Chimique Belge's early drug development unit.

Study Participants

Currently, 123 participants with a professionally diagnosed parkinsonian disorder and 39 healthy control participants have been enrolled. All healthy control participants are the spouses of the parkinsonian participants; providing an approximately age and gender matched control group.

Upon being enrolled, all parkinsonian participants are assigned to a sub-group by a movement disorder specialist based on the time since their diagnosis, the treatment being used, and the form of parkinsonism. These sub-groups are defined as:

- *Early Stage Not Medicated* (ESNM): Participants who received their diagnosis less than 8 years ago and are not receiving any treatment.

Table 4.1: Demographics of OxQUIP Participants

	n	Baseline MDS-UPDRS	Age	Male/Female %	Years Since Diagnosis	% Receiving Treatment
ESNM	15	26.0 ± 10.5	67.1 ± 5.9	42.9/57.1	2.7 ± 2.1	0
ESM	62	25.4 ± 14.7	66.2 ± 8.0	57.8/42.2	4.0 ± 1.9	100
SEV	20	28.7 ± 13.4	67.0 ± 16.3	54.5/45.5	13.3 ± 3.9	100
DBS	7	32.1 ± 5.4	58.7 ± 6.2	62.5/37.5	10.3 ± 5.7	100
PSP	25	49.0 ± 12.9	70.9 ± 9.2	53.8/46.2	2.6 ± 1.7	100
HC	39	5.6 ± 6.3	68.8 ± 8.4	52.1/47.9	N/A	0

- *Early Stage Medicated* (ESM): Participants who received their diagnosis less than 8 years ago and are receiving treatment.
- *Severe* (SEV): Participants who received their diagnosis more than 8 years ago and are receiving treatment.
- *Deep Brain Stimulation* (DBS): Participants who are currently receiving Deep Brain Stimulation treatment.
- *Progressive Supranuclear Palsy* (PSP): Participants who have been professionally diagnosed as having PSP - the atypical parkinsonian disorder.

Finally, all healthy control participants are assigned to a single sub-group. The demographics of all sub-groups are provided in Table 4.1. Further, the age distributions of all HC and PD participants are shown in Figure 4.1(a) whereas Figure 4.1(b) demonstrates the age distributions of each of the sub-groups individually.

Additionally, all participants undergo follow up examinations approximately every three months. The number of follow up visits of each sub-group is shown in Table 4.2. However, it should also be noted that multiple types of missing data are present. Firstly, the MDS-UPDRS scores for many participants are incomplete. For example, at baseline visit, although 162 participants underwent testing, 12 of their MDS-UPDRS scores were incomplete. Additionally, the data collection system was found to occasionally lose connection or for the data to be corrupted. Finally, if a participant was struggling to complete the tests the clinical staff allowed them to stop, which resulted in missing test data. As such, the number of participants who have all their data at all visits are given in the bottom row of Table 4.2. For

Table 4.2: The number of OxQUIP participants present at each follow up visit.

	Baseline	3 Months	6 Months	9 Months	12 months	15 months	Total
ESNM	13	13	12	11	7	4	60
ESM	62	56	56	52	46	37	309
SEV	20	14	14	14	14	6	82
DBS	3	7	0	0	0	0	10
PSP	25	15	13	8	6	3	70
HC	39	34	29	24	14	13	153
Total	162	139	124	109	87	63	684
Complete Data	145	111	95	73	32	13	N/A

example, all 95 participants who have complete data at the six month visit, are also present in the 111 participants with complete data at the three month follow up data and the 145 participants with complete data at baseline visit.

Study Severity Scoring

All severity scoring is performed by a movement disorder specialist during routine clinical visits. Severity scoring is performed via a full implementation of the MDS-UPDRS.

When the ESM participants make their second visit (three month follow up) they are asked to participate in the 'levodopa challenge'. This entails firstly undergoing examination in the 'OFF' medication state (usually between 9:00 and 12:00) followed by in the 'ON' medication state (usually between 13:00 and 17:00). A full MDS-UPDRS examination is performed in both states.

The distributions of total MDS-UPDRS scores for all HC and PD participants are given in Figure 4.1(c) for all 684 MDS-UPDRS instances. Additionally, Figure 4.1(d) demonstrates the total MDS-UPDRS scores of each of the sub-groups individually.

Study Activities

Whilst attending their routine MDS-UPDRS examination, participants are asked to complete additional walking, Timed-Up-and-Go (TUG), and sway tasks whilst wearing the APDM Mobility Lab^{TM} (`http://www.apdm.com/mobility/`) sensor network [292]. The APDM Mobility Lab system consists of multiple synchronised

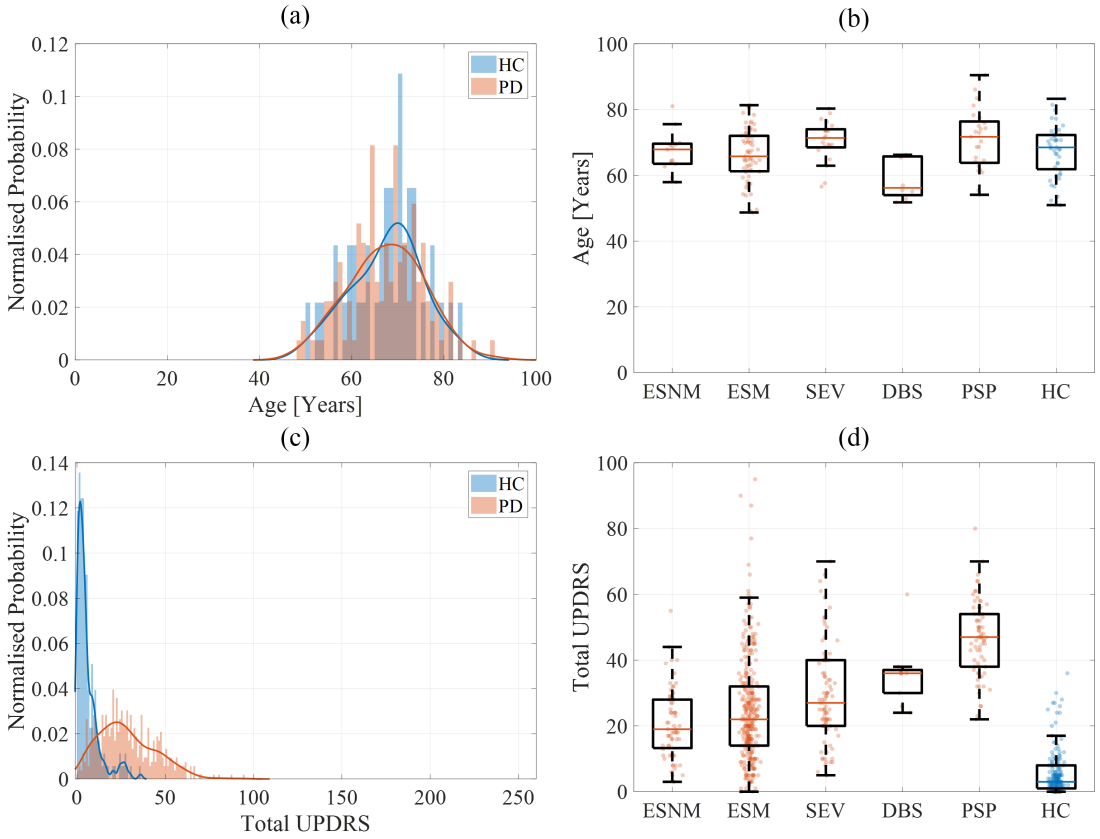


Figure 4.1: Normalised Histograms and density distributions (as determined through kernel density estimation) of the OxQUIP demographics for (a) the ages of all HC and PD participants (b) the ages of each participant sub-group (c) the MDS-UPDRS of all HC and PD participants and (d) the MDS-UPDRS of each participant sub-group.

Opal^{TM} IMUs that wirelessly transmit data to the Mobility Lab software running on a nearby laptop or personal computer.

During the OxQUIP study, the sensor network consisted of six IMUs attached at the lumbar, trunk, left and right wrists, and left and right feet (Figure 4.2). All sensors provided tri-axial accelerometer and tri-axial gyroscope signals at a frequency of 100Hz. This results in each activity providing 36 unique waveforms for each participant in each activity. The implementation of all activities was overseen and supervised by a movement disorder specialist.

Walking Activity

The two minute walk test was performed once at each visit and all participants completed the test on the same straight and level surface.

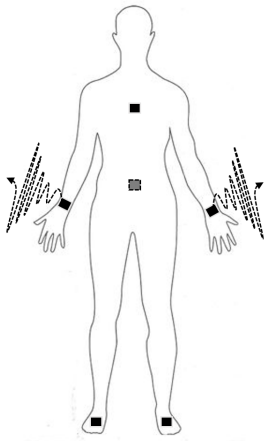


Figure 4.2: The locations of the six Opal IMUs during each activity in the OxQUIP Study.

Example waveforms from the lumbar, right arm, and right foot IMUs during a walking test are given for a HC and PD participant in Figures 4.3 and 4.4 respectively. These plots are annotated such as to demonstrate how each sensor is capable of detecting different types of motion. The x-axis gyroscope of the Lumbar sensor shows the ability of detecting turning periods yet a relatively noisy means of detecting individual steps. The x-axis gyroscope of the right arm sensor shows a weaker ability to detect turning periods whilst the accelerometer signals showing capable of detecting individual steps.

Figures 4.3 and 4.4 also enable a preliminary comparison of signal characteristics between disease groups. The magnitude of the lumbar x-acceleration from the PD participant is smaller than that of the HC thus demonstrating a reduced amount and velocity of motion. Additionally, the PD participant shows longer turning periods than the HC participant. The right arm acceleration and gyroscope signals from the HC participant clearly show each individual step. However, the individual steps in the corresponding waveforms from the PD participant are much more subtle and less pronounced. This indicates a smaller degree of arm movement by the PD participant; indicative of limb rigidity. The same trait is apparent in the right foot sensor wherein the magnitude of accelerations and rotation rates are dramatically decreased in the PD participant. Furthermore, the step velocity can be seen to be

lower in the PD participant as the distance between the peaks of the right foot y-axis accelerometer signals are further apart than in corresponding HC waveform.

Using the waveforms from all sensors, the Mobility Lab software performs automatic analysis such as to extract a wide range of clinical features pertaining to specific limb movements, left-right symmetry, and coronal and sagittal movements. A full list of these features and their descriptions are provided in Appendix A.1.

Sway Activity

The sway test was also performed once at each visit. During the sway test, participants were asked to stand upright and as still as possible for thirty seconds with their eyes closed on a firm surface. In all implementations of the sway test, a wooden template was placed on the floor so as to ensure all participants' feet were the same distance apart.

Example plots demonstrating the movement of two participants during the sway test are shown in Figure 4.5. These highlight the different degree of postural stability between a HC and a severe PD participant.

Using the waveforms from the lumbar sensor, the Mobility Lab software automatically extracts a range of clinical features pertaining to coronal and sagittal movements such as sway area, frequency dispersion, and path lengths. A full list of the sway test features and their descriptions are provided in Appendix A.1.

Timed-Up-and-Go Activity

The TUG test was repeated three times at each visit. Each test entailed the participant starting in a sitting position for three seconds, followed by standing up when instructed by the movement disorder specialist, walking forward seven meters, performing a 180 degree turn, walking back to the chair and sitting back down. The walking section of the test is completed on the same surface at the two minute walk test.

Example waveforms from the lumbar, right arm, and right foot IMUs during a single TUG test are given for a HC and PD participant in Figures 4.6 and 4.7 respectively. These figures demonstrate the TUG tests of the same participants

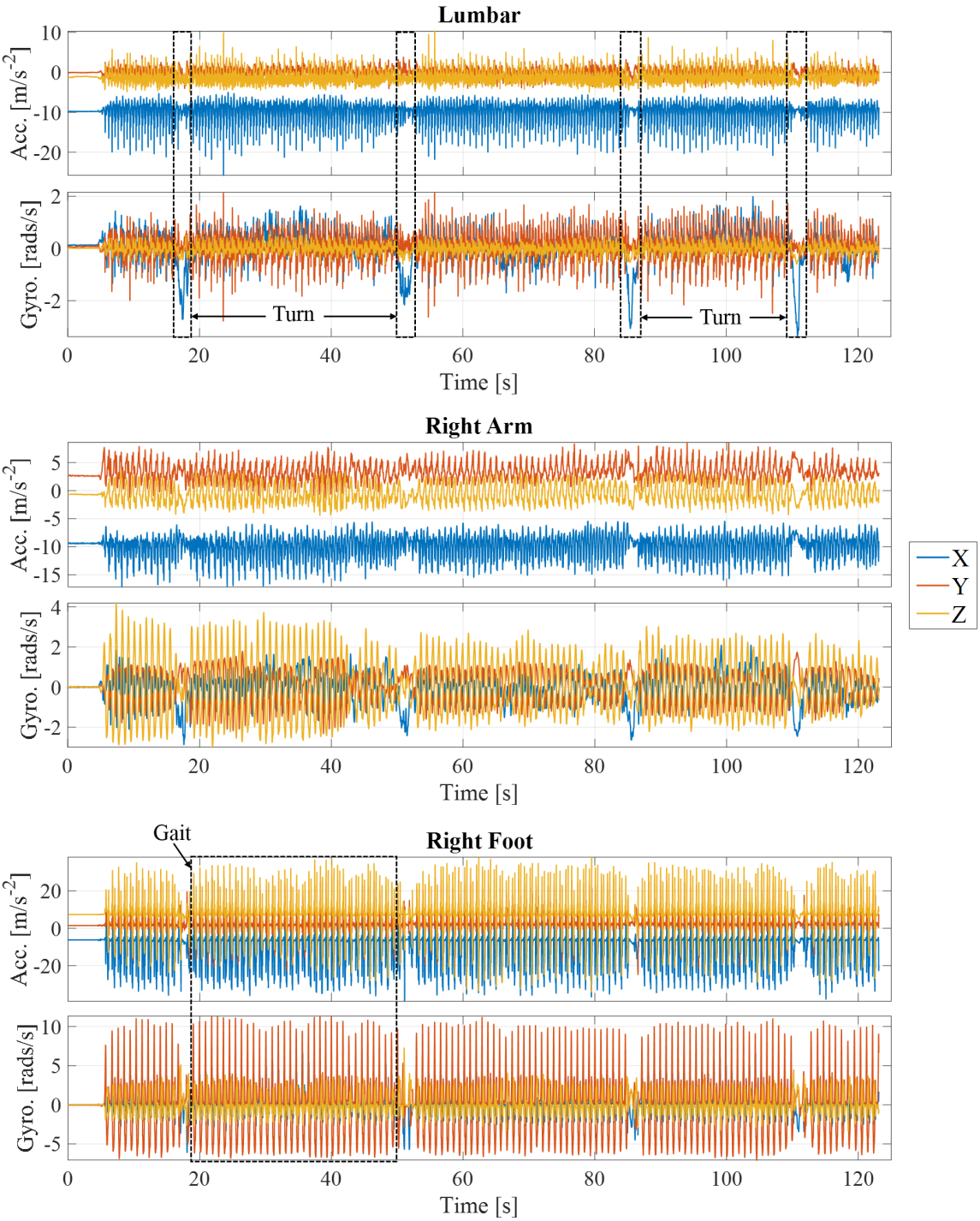


Figure 4.3: Example waveforms collected from a HC participant during a walking test using IMUs at the (a) Lumbar (b) Right Arm and (c) Right Foot

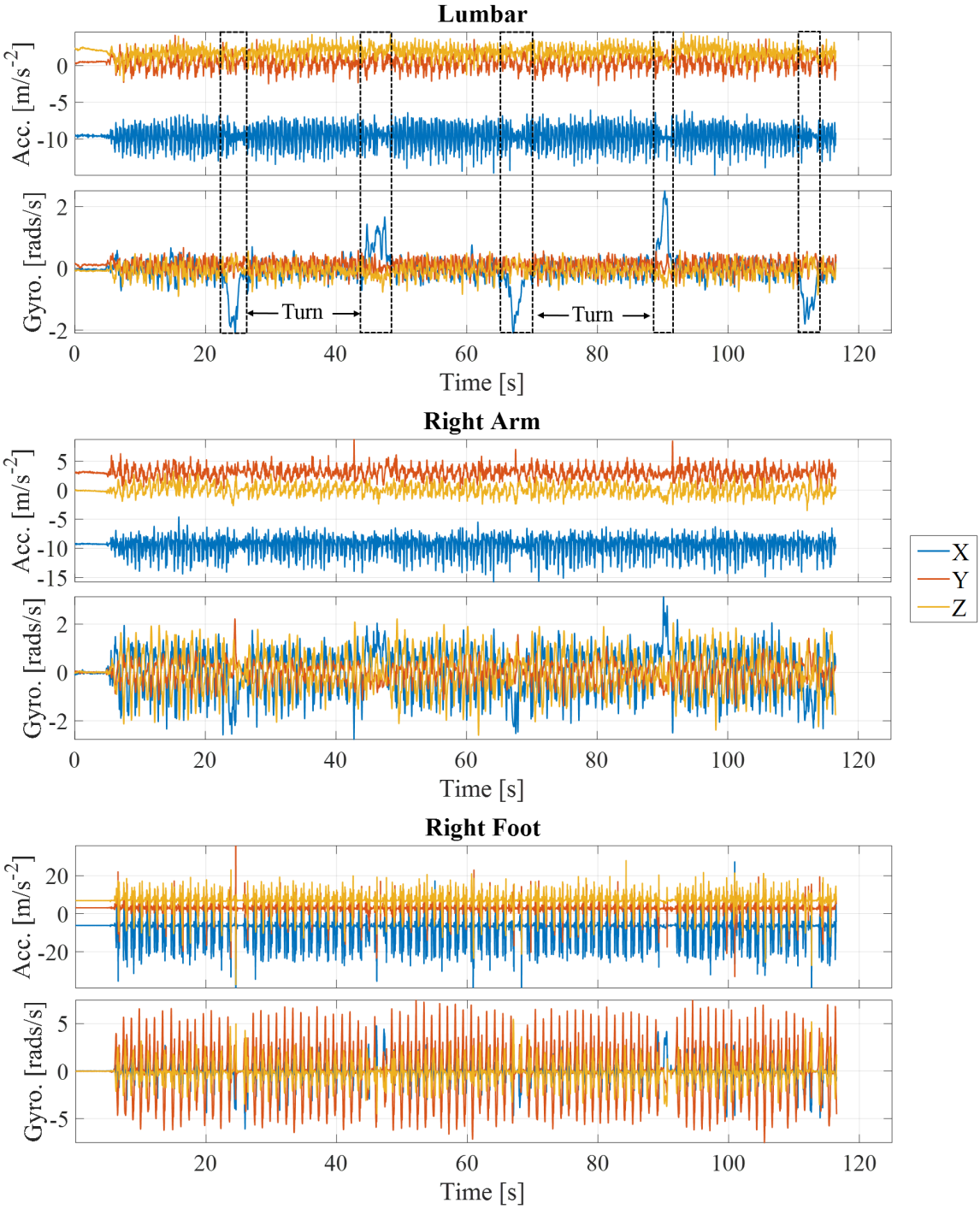


Figure 4.4: Example waveforms collected from a severe PD participant during a walking test using IMUs at the (a) Lumbar (b) Right Arm and (c) Right Foot

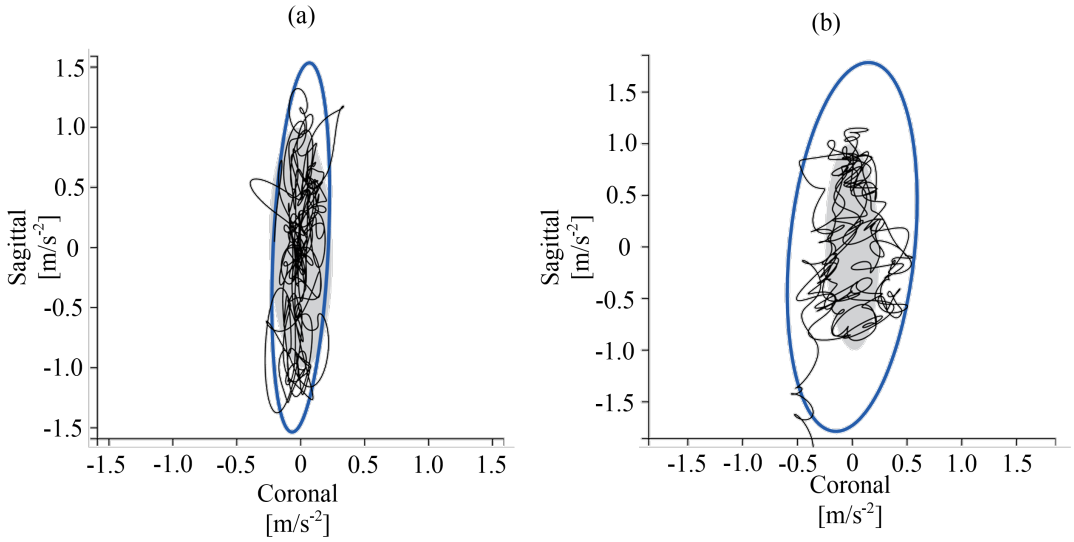


Figure 4.5: The coronal-sagittal stabilograms of (a) a Healthy Participant and (b) a severe PD Participant. The grey box represents a normative 95% Ellipse Sway Area whereas the blue box demonstrates the participants' 95% Ellipse Sway Area.

used in Figure 4.3 and 4.4. Again, these plots reveal how each sensor is capable of measuring different characteristics of motion. The lumbar sensor in both participants demonstrates the ability of the z-axis accelerometer and the y-axis axis gyroscope to detect sit-to-stand and stand-to-sit periods whilst the x-axis gyroscope can detect the turning period. Conversely, the right arm sensor cannot detect turning periods or transition movements yet can detect individual steps in some waveforms. Unlike in the walking test, the right foot sensor appears capable of only detecting individual steps.

Similarly to the walking activity, Figures 4.6 and 4.7 enable a preliminary comparison between disease groups during the TUG test. As in the walking activity, the range of accelerations shown by the PD participant is smaller than the HC participant representing a smaller degree of displacement during motion. Further, the duration of both the sit-to-stand and stand-to-sit periods are longer in the PD participant as shown by the accelerations from the lumbar sensor and the turning periods are smoother and shorted in the HC participant.

Using the waveforms from all sensors, the Mobility Lab software automatically extracts a range of clinical features specific to the TUG test. These features do

not include features from the walking period as these are encapsulated within the walking specific features. A full list of the TUG test features and their descriptions are provided in Appendix A.1.

4.3 The mPower Study

Study Background

The mPower study is an observational smart-phone based study of Parkinson's disease whose data was collected between March and September 2015 [78]. The mPower study presents the first example of an open-source dataset that was collected entirely in a remote environment. This dataset is intended to provide the research community a platform to assess the feasibility utilising remotely collected data to assess symptom fluctuation and the effects of medication whilst also facilitating remote disease classification and severity regression.

The study was conducted by Sage Bionetworks (Seattle, WA) who subsequently control the distribution of the dataset. A custom iPhone (Apple Inc., CA) application was developed using Apple's ResearchKit library whose source code is freely available (`https://github.com/Sage-Bionetworks/mPower`). The application was free for volunteers to download and was only available within the United States. The study utilised a novel enrolment strategy enabling participants to provide consent directly through the iPhone application. During the enrolment process, participants were given the option of sharing their data with the entire research community ('share widely') or exclusively with the Sage Bionetworks mPower study team ('share narrowly').

The dataset is stored using a data and analysis sharing service called Synapse, which was also developed by Sage Bionetworks. Access to the dataset is granted by Sage Bionetworks to Certified Synapse users upon completion of a policies and procedures test and the submission of a data usage statement. Retrieval of data from Synapse is achieved via the Synapse Application Programming Interface (API) which is available in the Python and R programming languages.

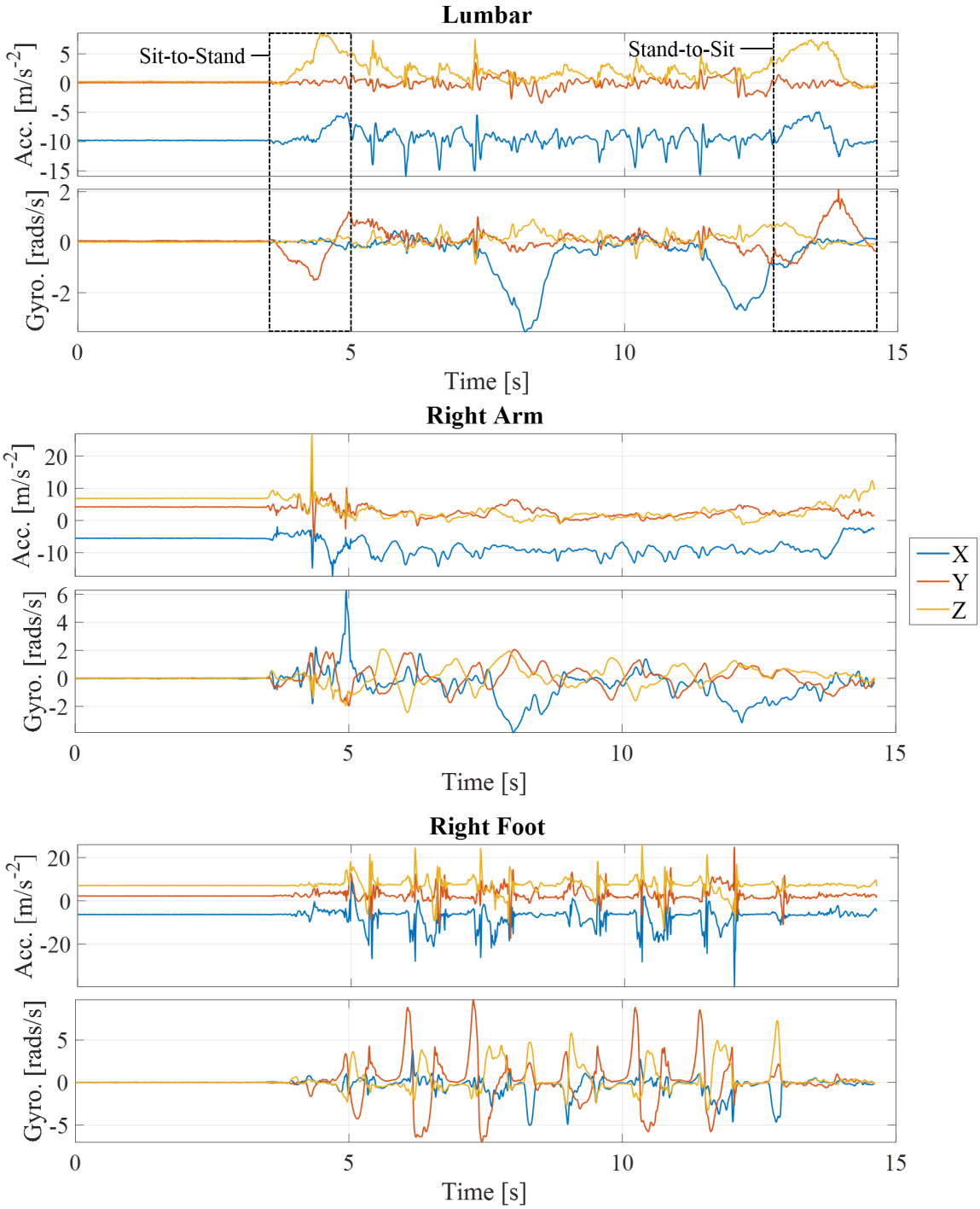


Figure 4.6: Example waveforms collected from a HC participant during a TUG test using IMUs at the (a) Lumbar (b) Right Arm and (c) Right Foot.

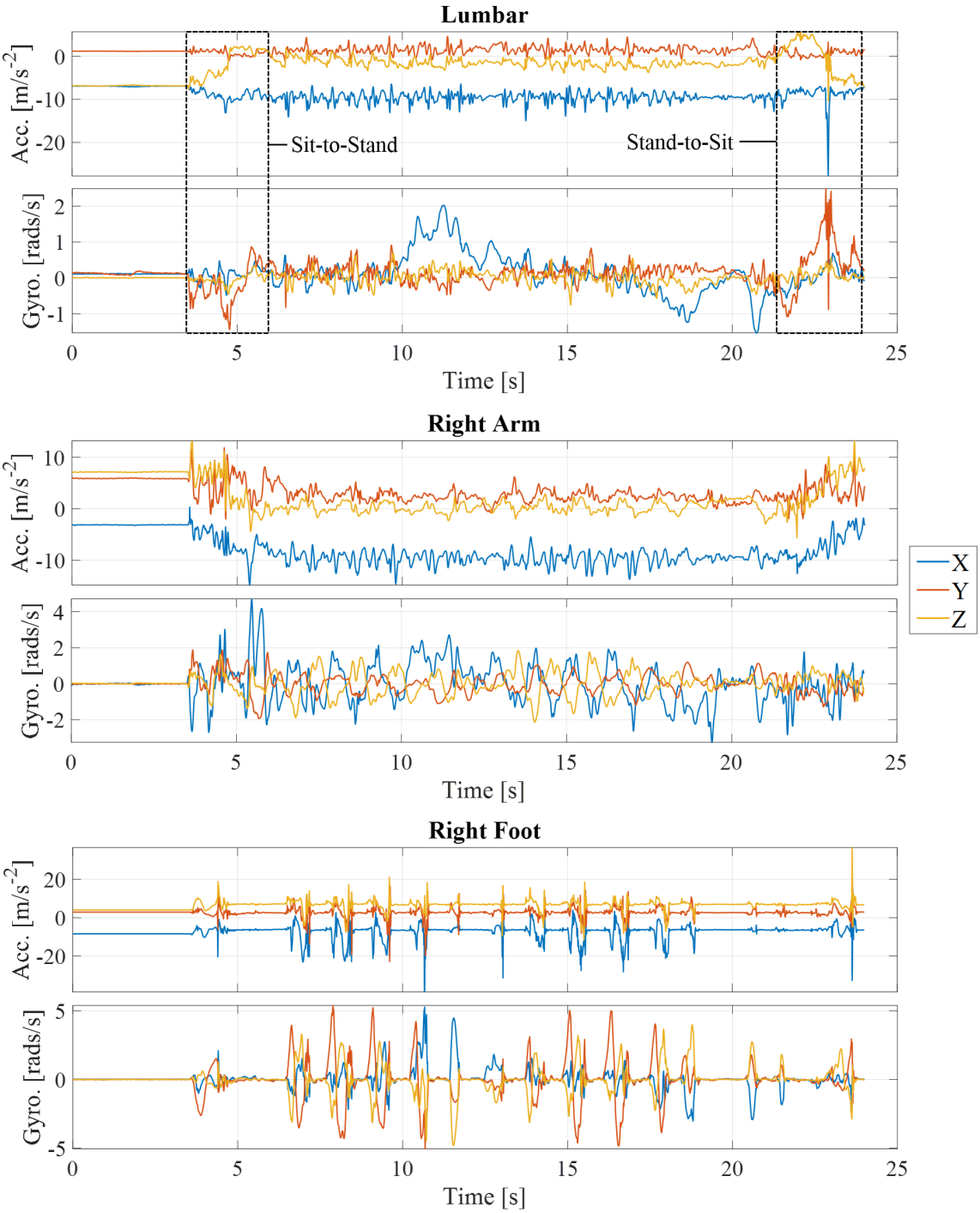


Figure 4.7: Example waveforms collected from a severe PD participant during a TUG test using IMUs at the (a) Lumbar (b) Right Arm and (c) Right Foot.

The iPhone application presented participants with a quick, intuitive, and non-invasive means of providing multiple forms of data including demographic data, three motor tests, one non-motor test, and several severity reporting questionnaires. The convenience of the application resulted in a very high number of participants who contributed multiple data types on a longitudinal basis.

Study Participants

All volunteers who downloaded the mPower application were asked to complete a brief eligibility test (confirming they are above 18 years old, reside in the United States, and are able to read and write using an iPhone) followed by providing electronic consent and data sharing preferences.

The mPower application was downloaded a total of 48,104 times during the first six months. However, many participants either did not enrol (31,519), did not provide a valid email address (1,901), or withdrew from the study prior to choosing how to share their data (2,483). Of the 12,201 participants who successfully completed the enrolment requirements, 9,520 participants chose to share their data with the entire research community.

Upon successfully enrolling in the study, participants were asked to contribute a one-time demographics survey. Of the 9,520 who opted to share their data widely, 6,805 participants completed the demographics survey with 173 participants refusing to answer whether they had received a professional diagnosis.

As the mPower study allowed participants of any age (18 or above) to enrol, the participants are divided into three sub-groups for the exploratory analyses of this chapter[1]. Firstly, Young Healthy Controls (YHCs) are participants who have not had a professional diagnosis and who are under the age of 50. Healthy Controls (HCs) are participants who have not had a professional diagnosis and who are 50 years old or above. Finally, all participants with a professional diagnosis are assigned to the PD sub-group. The demographics of each group is given in Table 4.3. Further, the distribution of ages for each of the sub-groups is shown in Figure 4.8(a).

[1]These sub-groups, using their same definitions, will be revisited during Chapter 6.

Table 4.3: Demographics of mPower participants. The mean ± standard deviation is given for distribution data.

	n	Baseline MDS-UPDRS	Age	Male/Female %	Years Since Diagnosis	% Receiving Treatment
YHC	4,946	5.6 ± 5.3	29.1 ± 8.8	75.2/24.8	N/A	0
HC	599	6.0 ± 5.6	60.9 ± 8.1	73.2/26.8	N/A	0
PD	1,087	13.5 ± 8.1	60.7 ± 10.3	61.2/38.8	10.0 ± 5.6	93.3%

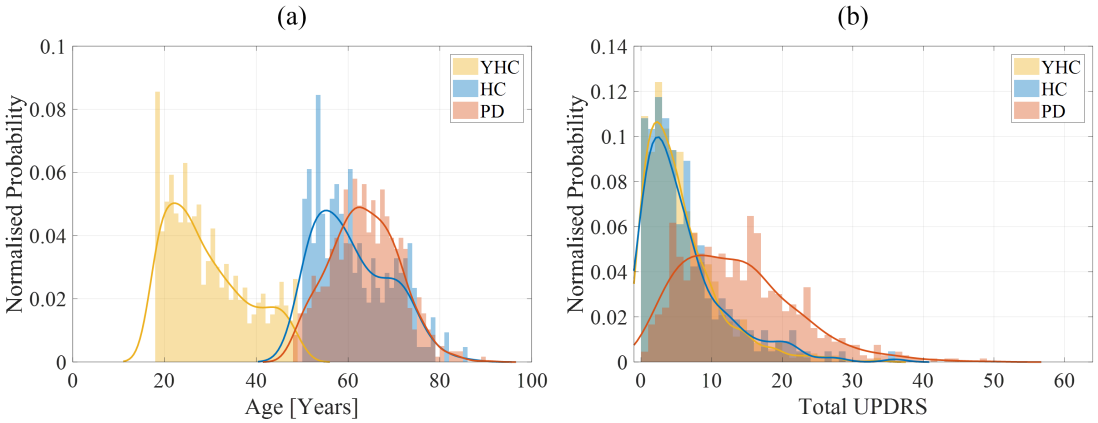


Figure 4.8: Normalised Histograms and density distributions (as determined through kernel density estimation) of the mPower demographics for (a) the ages of all HC and PD participants and (b) the MDS-UPDRS all HC and PD participants.

Study Severity Reporting

As discussed in Chapter 2, Parts I and II of the MDS-UPDRS were designed such as to allow a patient or caregiver to complete them without the aid of a movement disorder specialist. This trait was exploited by the mPower application as participants were able to complete a subset of MDS-UPDRS Parts I and II questions at a maximum frequency of once a month. All participants, regardless of whether they reported as having a professional diagnosis of PD were allowed to complete the MDS-UPDRS survey. As a total of 16 MDS-UPDRS questions are available to participants, the maximum possible severity is 64 as opposed to the maximum possible severity of 260 in the full MDS-UPDRS. The MDS-UPDRS survey was completed a total of 2,305 times by 2,204 participants. The total severity score distributions for all 2,305 instances is given in Figure 4.8(b).

Distribution of all MDS-UPDRS data is controlled by the International Parkinson and Movement Disorder Society. In order to access the MDS-UPDRS data, written

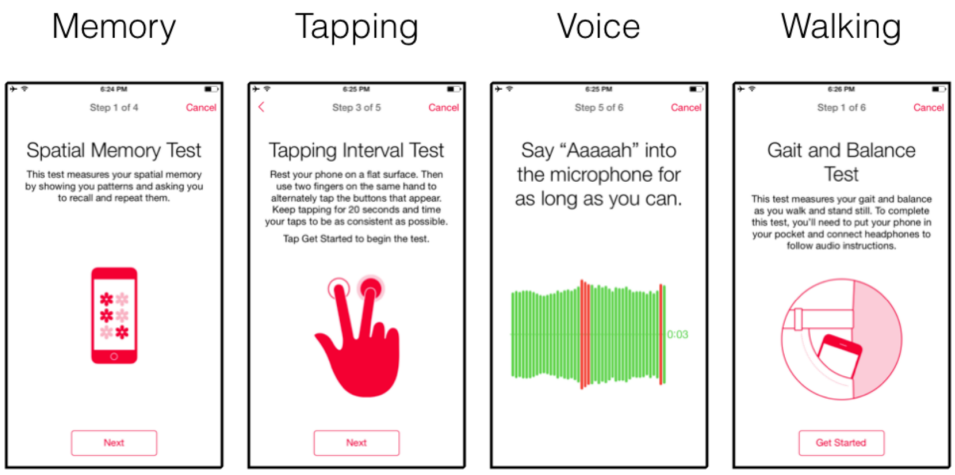


Figure 4.9: Screen-shots of the instructions given to participants in each of the four mPower activities.

approval is required from these governing bodies based on an additional data usage statement.

As an aside, eight questions from an additional and less commonly used severity scoring system, the Parkinson's Disease Questionnaire 8 (PDQ-8), are also available. However, these questions were completed by a substantially smaller number of participants and contain a large degree of overlap with the MDS-UPDRS subset of questions, and are therefore not further discussed in this research.

Study Activities

The mPower application provided four activities for the participants to complete which were intended to assess numerous disease symptoms. Screen-shots of each of the activities are provided in Figure 4.9. Prior to starting any given activity, participants were presented with a screen allowing them to report their current medication state as being either 'Before Medication', 'After Medication', or 'Another Time'. Alternatively, the participants could select 'I Do Not Take Parkinson's Medication'.

Walking Activity

The walking activity entailed a participant placing their iPhone into the front pocket of their trousers and performing a walking test. This activity was intended to collect data which would allow objective analysis of gait and postural instability

symptoms. The experimental protocol varied numerous times throughout the study. During the earlier application versions, participants were asked to walk 20 steps in a straight line. Later versions of the application asked the participants to walk 20 steps in a straight line, perform a 180 degree turn, and to stand still until the test is complete. The application provided audio instructions throughout the walking test as to when each of the separate test components should be performed. The walking activity was completed 35,410 times by 3,101 participants.

All data collected during each walking activity originates from the smart-phone embedded IMU including the tri-axial accelerometer and tri-axial gyroscope signals at a sampling frequency of 100Hz. Furthermore, attitude coordinates (otherwise known as Quaternions) were collected so as to enable the iPhone coordinate system to be aligned with the global coordinate system through the use of a quaternion transformation [293]. The use of a quaternion transformation accounts for participants placing their phone into their pocket at different orientations and thus ensures the signal calibration between all recordings is approximately equal [294].

To provide an introduction and visualisation of the walking activity data, three example recordings are provided in Figure 4.10. These examples are intended to demonstrate the highly variable nature of the type of data collected by the walking activity. In Figure 4.10(a), an example of an eighteen second recording is provided which consists of a clean gait signal. Alternatively, Figure 4.10(b) demonstrates a recording from a later version of the application that lasts approximately 30 seconds. Again, this recording shows a very clean section of gait followed by an approximately 15 second period of standing still. Finally Figure 4.10(c), demonstrates a noisy and improperly implemented walking test recording. The first 24 seconds of this recording are undefinable. The final six seconds however shows oscillatory behaviour indicative of walking with a period of no movement.

Tapping Activity

During the tapping activity, participants were asked to place their iPhone on a flat surface with the screen pointing upwards. Using two fingers from the same

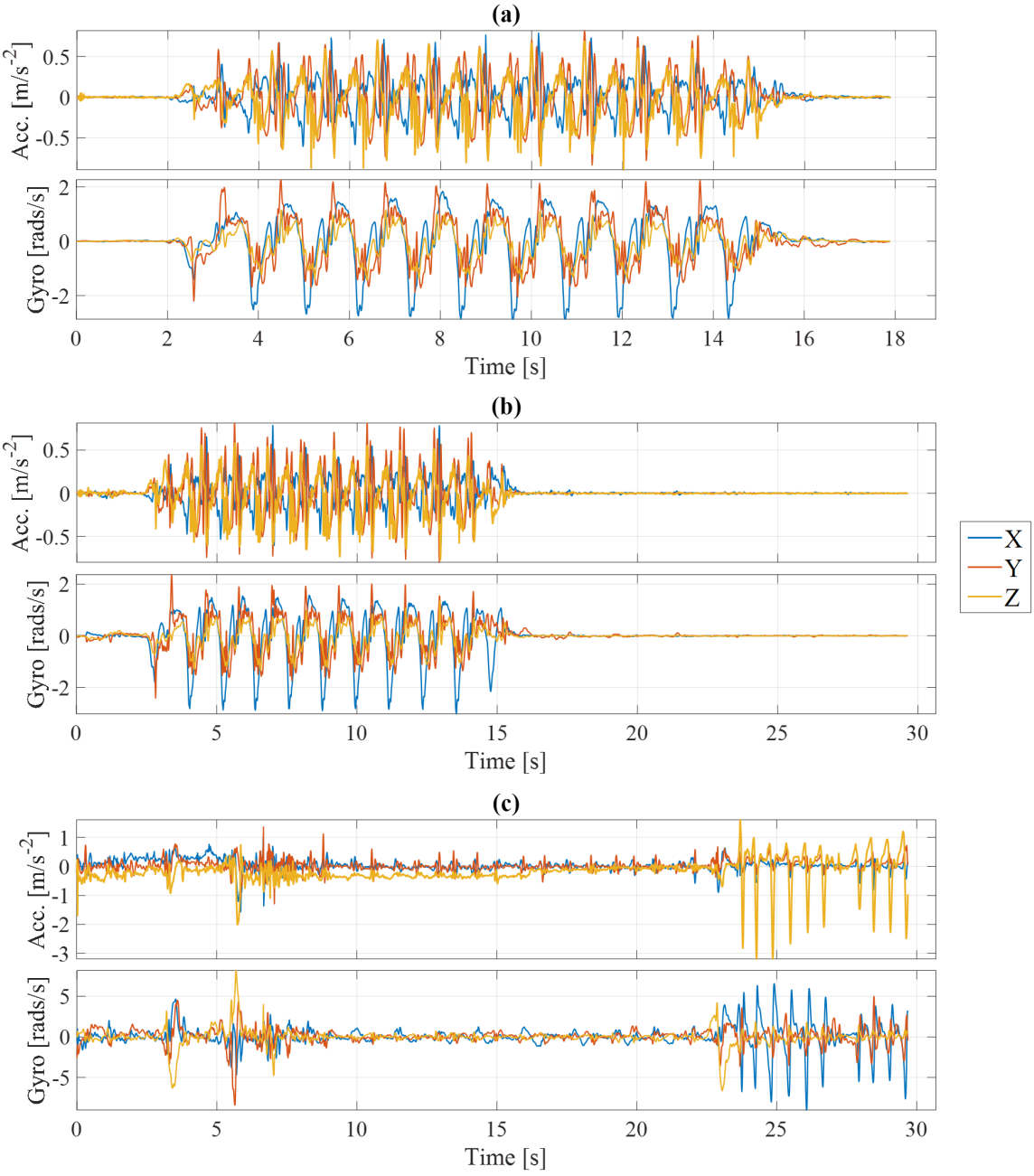


Figure 4.10: Example IMU waveforms from the mPower walking test demonstrating (a) a high quality gait recording (b) a high quality gait and standing recording from a HC participant and (c) a low quality from an improperly completed gait and standing test consisting of mostly noise with very little gait from a PD participant.

hand, the participants alternately tapped two on-screen buttons for 20 seconds. This test therefore replicates the common clinical Alternate Finger Tapping (AFT) test which is used to assess dexterity. The tapping activity was completed 78,887 times by 8,003 participants.

The data collected from the tapping activity includes the $\{x, y\}$ pixel coordinates of each tap alongside their corresponding timestamp - which is measured in seconds relative to the time of the first tap. Additionally, the tri-axial accelerometer waveform is recorded during the activity at a sampling frequency of 100Hz.

Figure 4.11 provides a visualisation of how the tapping activity was implemented by a HC and a PD participant. For each example, the pixel coordinates of each tap have been plotted relative to the location of the 'tap' buttons. The tap locations have also been plotted to scale for the iPhone 5, which both tests were implemented upon. The tap locations have been colour coded according to whether they were intended to tap the left button (red) or the right button (blue). It is evident that the HC participant not only achieves more taps during the 20 second period, but also that the taps are closer to the intended button. Conversely, the PD participant achieves less taps and their taps are more dispersed across the screen.

Voice Activity

During the voice activity, participants were asked to make a sustained \a\ ('Ahhh') phonation into the iPhone microphone at a steady volume for 10 seconds.

The data collected during the voice activity is a raw audio file of the voice recording at 44.1kHz as well as an audio recording of a five second count-down period prior to the participant starting the phonation. The voice activity was completed 65,022 times by 5,826 participants.

Figure 4.12 provides three example voice recordings for the purpose of demonstrating data quality and the ability to detect physiological measures as described in Chapter 2. Figure 4.12(a) demonstrates a healthy participant who holds and maintains the phonation at a steady amplitude throughout the test. Upon closer inspection of this recording (Figure 4.12(b)) it is evident that the fundamental

Figure 4.11: The tapping coordinates of (a) a HC participant and (b) a PD participant. Both participants used an iPhone 5 therefore this figure was generated via mapping the pixel coordinates to those of an iPhone 5 whose pixel dimensions are 1136×640. The total number of taps was calculated manually during post-processing and superimposed for visualisation purposes.

frequency and amplitude can easily be measured. However, Figure 4.12(c) demonstrates a recording from a PD participant whose amplitude is frequently changing. Recall that frequent changes in vocal amplitude is referred to as Jitter (§2.3.3) and is an abnormality distinct to neurodegenerative diseases including PD. Finally, Figure 4.12(d) provides an example of a noisy and improperly complete voice test as completed by a HC participant. The first seven seconds consists entirely of noise with the participant clearing their throat (the burst of amplitude just before six seconds). The participant initiates the test after seven seconds resulting in only three seconds of sustained phonation being recorded.

Memory Activity

The memory activity presented the participants with a grid of flowers, a number of which would light up in a random pattern which the participants would have to

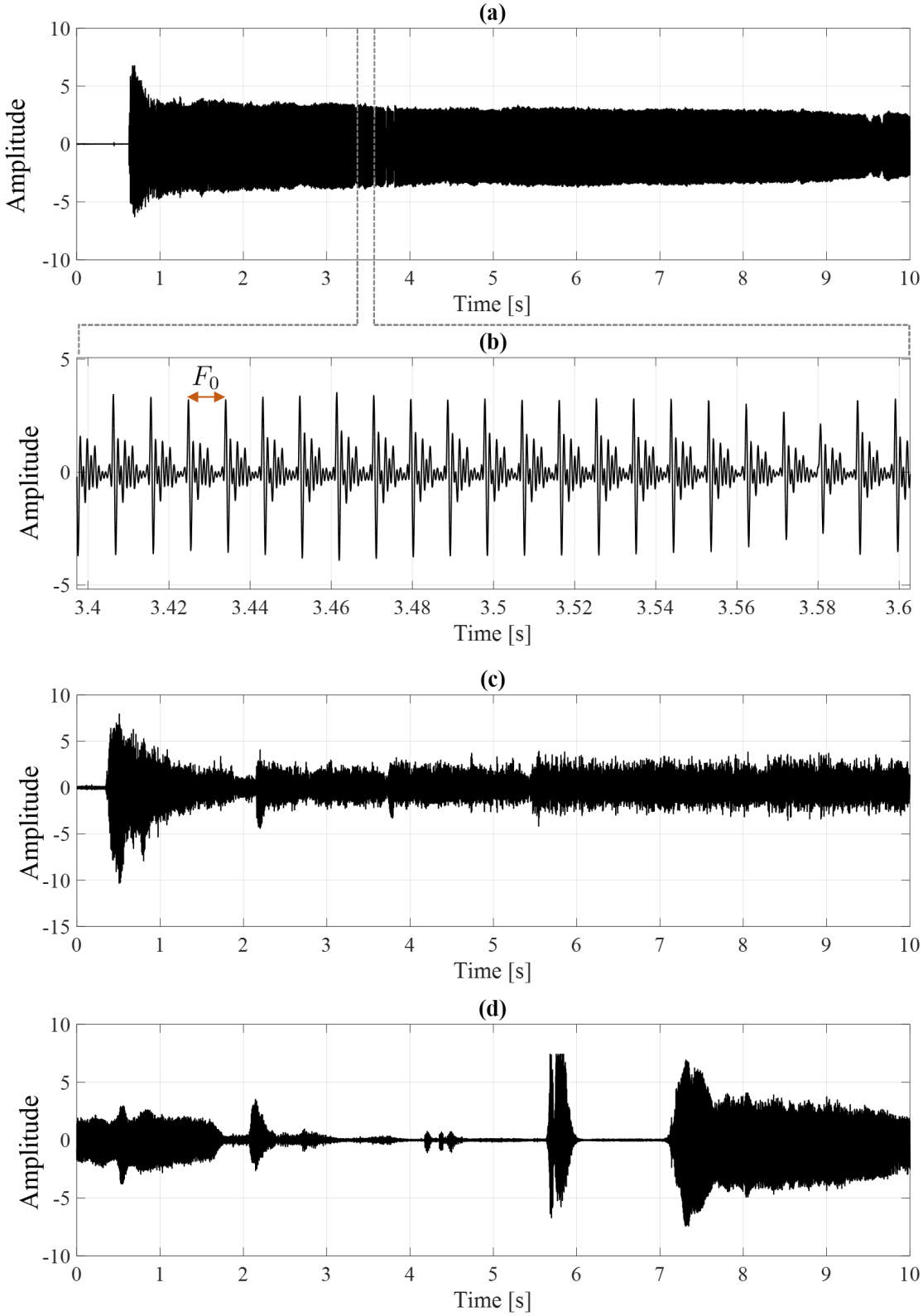


Figure 4.12: Example recording from the voice activity demonstrating (a) a healthy HC participant showing showing a phonation with a constant amplitude alongside an annotated segment shown in (b) with an example F_0 whereas (c) demonstrates a PD participant whose amplitude varies (Jitter) throughout the recording and (d) shows an example of a noisy and improperly complete test.

recall. There were five levels of difficulty during each implementation of the memory activity with the grid size and number of flowers to recall increasing at each level.

The data provided by the memory activity included the total score, the number of incorrect responses, and the number of completed levels. Significantly fewer participants completed the memory activity as it was introduced to the study a month after the other activities. The memory activity was completed 8,569 times by 968 participants.

4.4 Discussion

This chapter has described and provided examples of the data within both of the datasets used throughout this thesis. Prior to presenting major analyses on these datasets in the following four chapters, this discussion highlights their relative advantages and limitations.

The primary strength of the OxQUIP dataset is the reliability of all of its data. This includes the high quality raw signals in all three tests, the subsequent clinically derived features, and the clinically assigned MDS-UPDRS scores. Furthermore, the participant numbers are considerably larger than those used in previous studies whilst also containing diseased participants covering a wide severity range (Figure 4.1). The large sensor network also facilitates the analysis of lesser studied clinical features such as asymmetric gait.

The first limitation of this study is the infrequency of follow up examinations. As the OxQUIP study is hitch-hiking on participants' routine clinical care appointments, data is only collected once every three months. As discussed in Chapter 2, this is infrequent and incapable of detecting the higher frequency (daily or weekly) symptom fluctuation. An additional limitation is that all tests are focused on motion analysis and do not investigate other symptoms such as tremor, dexterity, voice, or memory. Therefore the data being collected is focused on one, albeit a very common, area motor symptoms may present.

The first and foremost advantage of the mPower dataset is the unprecedented number of participants and activity instances. In total, 9,520 participants completed

196,996 instances of data across all activities, severity surveys, and demographic surveys. Further, many participants completed multiple tests on a longitudinal basis. As tests can be completed at the participants leisure, many participants contribute instances on a highly frequent, often daily, basis. The study of longitudinal test contribution and participant retention will be discussed in detail during Chapter 6. Furthermore, the mPower cohort is extremely diverse in age and severity; allowing the symptoms and performance of young healthy controls (a usually under-represented demographic) to be compared against older healthy controls and diseased subjects.

The second advantage is that the activities aim to assess multiple aspects of the disease including body motion, voice, dexterity, and non-motor impairments. This will enable a far more detailed and comprehensive analysis of the overall disease severity whilst also presenting the first opportunity to compare how the symptoms between multiple activities relate to one another.

As mPower is the first open-source, remotely collected, longitudinal, and 'big' dataset, it would be naïve to assume it is without its limitations. Firstly, as a purely remotely collected dataset, the demographic data has not undergone validation from a clinician. As such, it is necessary to assume all demographic data is accurate and honest. Furthermore, many of the demographic survey questions have a binary response. This could be problematic when a participant is answering whether they have received a professional diagnosis as the binary response does not account for parkinsonian disorders. The same limitation is true for the medication usage as the type and dosage of medication cannot be specified.

Table 4.4: Comparison of the OxQUIP and mPower datasets.

	OxQUIP	mPower
Number of Participants	162	9,520
Environment	Clinical	Remote
Measurement Device	IMU	Smart-phone
Measurement Quality	Very high	Very variable
Activity Type(s)	Body Motion	Body Motion, voice, dexterity, non-motor
Activity frequency	3 Months	Daily
Severity Score	Full MDS-UPDRS	Subset of MDS-UPDRS Parts I and II
Severity Scoring	Movement disorder specialist	Self-reported

Secondly, the examples of the walking and voice data collected during the mPower study demonstrated a high level of variability in quality (Figures 4.10 and 4.12). In the case of the walking test, sources of noise may include improper test completion and the completion of tests in inconsistent environments. Neither of these sources of noise are present in the corresponding OxQUIP walking test data. In the case of the voice test, sources of noise may include improper test completion and the presence of background noise. It is here hypothesised that the tapping test is the most commonly completed as it is the simplest and least intrusive test to perform. It is also less susceptible to test environment and consequently will suffer from less sources of noise. Although improper test completion is still a potential source of noise to the tapping activity, the intuitive application interface will likely decrease the number of incorrectly implemented instances.

Thirdly, there is also the potential of noise entering the mPower MDS-UPDRS survey. Although the efficacy of self-reporting Sections I and II of the MDS-UPDRS have received some attention, and often show agreement between the scores given by patients and movement disorder specialists, these studies have focused on considerably smaller participant numbers often from a single medical facility. As the mPower dataset contains a far larger number of participants from medical centers across the United States, the level of intra-rater variability remains unknown. A further source of noise in the MDS-UPDRS survey is that it is being completed by people who do not have PD and therefore have no prior experience or benchmark with the scoring system.

Table 4.4 provides a final and succinct comparison of the datasets presented by this chapter.

5

Disease Quantification During Routine Clinical Care

Related Publications:

1) J. Prince, J. Fitzgerald, C. Antoniades, M. De Vos, "Differentiation of Parkinsonian Disorders Using Wearable Sensors and Machine Learning", Movement Disorders (Manuscript in preparation)

5.1 Introduction

Chapter 2 provided an exhaustive background detailing how digital sensors have previously been used in a clinical environment for objective disease assessment. The majority of these studies suffered from small cohort sizes at usually mild or severe stages of disease progression. Further, many of these studies implemented proof-of-concept experimental protocols which utilised few sensors for a purely cross sectional assessment. The OxQUIP study, as described in Chapter 4, overcomes the vast majority of these limitations via using a large network of sensors as part of routine clinical care.

The first objective of this chapter is to replicate the cross sectional confirmatory analyses of previous studies utilising three clinically administered motor tests. This is intended to reveal the underlying characteristics of the novel dataset in a bid to identify features that are distinct of Parkinson's disease (PD) and that also correlate with clinically assigned MDS-UPDRS scores.

Secondly, this chapter performs a cross sectional disease classification and severity prediction using data collected from participants at their first visit. Far fewer studies have gone beyond classical confirmatory analyses (§3.1) to perform

cross sectional classification and regression tasks. Data is available from three motor tests therefore a quantitative comparison between the inter-test and intra-test correlations is presented and whether classification and regression are improved via fusing the features from all tests.

Thirdly, as the OxQUIP study collects data from each participant once every three months, it facilitates the implementation of several previously impossible longitudinal analyses. These analyses focus on determining whether disease deterioration or longitudinal symptom variation can be detected at the standard frequency of clinical assessment. Confirmatory statistical tests are employed to inspect for novel longitudinal behaviours of features over the course of nine months.

Finally, longitudinal classification and regression tasks are performed. This is the first work to present a classification task that incorporates longitudinal feature variation into the design matrix. Additionally, this is the first work to present a regression task that tracks disease severity using features from common clinical motor tests.

This chapter provides a comprehensive overview of how wearable sensors can be used as part of routine clinical care for objective disease assessment. The OxQUIP study is purely clinical and the subsequent dataset is very high quality and consists entirely of clinical features (as opposed to technical features - see §2.3.1). Despite its highlighted shortcomings, the MDS-UPDRS is assigned by a movement disorder specialist and therefore acts as the gold standard of disease severity throughout this chapter. As such, the findings of this chapter provide the thesis with a benchmark regarding the potential and challenges associated with disease assessment using wearable sensors.

The primary contributions of this chapter to the field of clinical Parkinson's disease quantification are:

- The replication of the cross sectional confirmatory analyses of previous studies on the large, diverse, multi-sensor, and multi-test OxQUIP dataset.

- An assessment as to the ability of features from three common clinical tests to perform cross sectional disease classification and severity prediction.
- An investigation into the longitudinal properties of features; determining whether disease groups show a different amount of longitudinal change and fluctuation.
- The first implementation of a longitudinal disease classification task that investigates whether the classification procedure benefits from the inclusion of longitudinal feature behaviour. This is performed using data collected as part of routine longitudinal clinical care.
- The implementation of a longitudinal regression task to assess the ability of features to accurately track disease severity.

5.2 Background

The identification of objective biomarkers that can be used for quantitative PD assessment are widely sought in a variety of in-clinic motor tests [80]. As postural instability and bradykinesia are indicative of the presence of PD, a substantial amount of literature exists wherein distinct PD body motion characteristics have been identified that are not found in a healthy population [89, 295]. However, the technologies considered to be the 'gold standard' in deriving these biomarkers are expensive and often lack the ability to effectively assess gait variability [296–298]. As such, the adoption of low cost Inertial Measurement Units (IMUs) has risen as an alternative means of deriving digital biomarkers in a clinical environment. Indeed, IMUs have shown the potential to accurately derive many of the biomarkers that 'gold standard' technologies have identified as showing differentiating ability between PD and HC participants [84, 89, 96, 299, 300].

However, as IMUs are still an emerging technology for the in-clinic assessment of PD, their clinical relevance and uptake have been hindered by several important factors. Firstly, previous studies have utilised small cohort sizes, usually with

less than 30 participants [20, 80]. Secondly, experimental protocols have varied wherein the number and location of IMUs, test environment, and test procedure have differed between studies; restricting the generalisation of findings [94, 301]. As symptoms often present at different body locations (including lower limbs, upper limbs, and trunk), the use of a small number of sensors is unlikely to capture the relevant biomarkers for many symptoms.

A small number of datasets have emerged attempting to address these limitations; aided by several commercial IMU systems recently entering the market [55, 131, 302]. Studies utilising these datasets have focused on performing cross-sectional body motion analyses, using a large array of IMUs on larger cohorts at a wide range of disease progressions. Furthermore, the IMU usage within these datasets often implement multiple body motion tests including gait, Timed-Up-and-Go (TUG), and the sway test [93, 115, 131]. The primary focus of these studies have been validating the previously identified clinical biomarkers that show statistical differences between disease groups.

Due to the unpredictable and non-linear progression of PD, the ability to objectively quantify disease symptoms and progression on a longitudinal basis would improve the understanding of the long-term symptom fluctuations. Combined with the known inter- and intra-rater subjectivity of the MDS-UPDRS scoring system, it is difficult to determine if the measured longitudinal variation of severity is caused by worsening symptoms or simply by the variability of the scoring system. Few studies have attempted longitudinal monitoring with IMUs, and those that have have suffered from the aforementioned limitations of small subject numbers [116, 303]. Furthermore, the longitudinal data collected by these studies are highly variable and infrequent; with data collection often occurring at baseline and then at follow up, which can range from 12 to 24 months later [116, 123, 304]. These infrequent measurements will be insensitive to the natural variation of symptoms that occur on a more frequent basis [163, 305]. Thus, the use of longitudinal tests utilising IMUs during routine clinical care would create a more robust platform to assess disease progression as natural symptom variation can be accounted for.

In this chapter, both a cross sectional and a longitudinal analysis are performed on a large cohort of participants at many stages of disease progression. The cross sectional analysis aims to firstly replicate the analyses of previous studies, as described in §2.3.1, on the novel and rich OxQUIP dataset. This includes traditional statistical tests aimed at identifying features that are distinct of PD from gait, TUG, and sway tests as collected by a large network of IMUs. The analyses of previous studies are extended through the implementation of cross sectional disease classification and regression tasks. The results of previous studies are further extended via a longitudinal analysis. Initially, the longitudinal behaviour of features are assessed for a difference between the disease groups at a frequency of three months. The effect of including longitudinal test data on disease classification is then compared against the traditional cross sectional classification procedure. Finally, longitudinal disease severity prediction is performed on a participant-to-participant basis in a bid to enable objective tracking of disease severity.

5.3 Methods

Dataset and Clinical Feature Description

All participants used in this chapter originate from the OxQUIP study. As this chapter is focused on identifying diseased participants, no differentiation is made between typical and atypical Parkinsonism[1]. All participants who receive treatment (levodopa and Deep Brain Stimulation) are in the 'ON' state for all test instances. The feature sets extracted from each test are those described in §4.2 and are listed in Appendix A.1.

[1]The associated publication of this chapter concerns the differentiation of typical from atypical parkinsonism. The cross sectional analysis performed in this chapter follows an identical analysis procedure to the associated publication but for the purpose of differentiating all parkinsonism disorders (typical and atypical) from healthy controls.

Table 5.1: Demographics of the OxQUIP participant subsets.

	# Participants	# Instances	Age	MDS-UPDRS	Years Since Diagnosis
	Cross Sectional Baseline Subset				
HC	44	44	66.4 ± 7.6	5.4 ± 6.28	N/A
PD	101	101	67.5 ± 8.2	30.1 ± 15.7	5.6 ± 4.7
	Longitudinal Classification Subset				
HC	21	84	67.1 ± 8.1	5.6 ± 6.9	N/A
PD	52	208	67.3 ± 8.3	26.0 ± 13.0	5.8 ± 5.1
	Longitudinal Regression Subset				
HC	44	147	66.4 ± 7.6	4.7 ± 4.1	N/A
PD	101	338	67.5 ± 8.2	29.6 ± 16.5	5.6 ± 4.7

5.3.1 Cross Sectional Baseline Disease Quantification

Cross Sectional Baseline Subset

The participants who completed all three motor tests (walking, TUG, and sway) at baseline visit are included in all baseline analyses and are herein referred to as the *Cross Sectional Baseline Subset*. The demographics of these participants are given in Table 5.1.

Cross Sectional Feature Hypothesis and Correlation Tests

This section performs a confirmatory statistical analysis on the clinical feature set arising from each test individually. This is the most common form of analysis reported in previous literature and therefore enables the findings of this chapter to be validated against previous work [123, 131]. Statistical differences are sought through the use of hypothesis tests whereas correlation analyses are performed to reveal any relationships between the identified biomarkers and the clinically assigned MDS-UPDRS scores.

For each test, independent t-tests are used to inspect for statistical significance between disease groups for all features using a significance level of 0.05. The mean and standard deviation of all features found to possess p-values below the significance level are reported alongside the corresponding p-value. All features identified as showing significant differences are herein referred to as the *significant feature set*. The Spearman's correlation coefficient between each feature in the significant

feature set and the total MDS-UPDRS is reported alongside the corresponding correlation p-value

Due to the likelihood that many of the features within each significant feature set are correlated (such as Gait Speed Left and Gait Speed Right), LASSO feature selection is implemented such as to reduce the significant feature set into a dense subset of features. The resulting features and LASSO weightings are reported.

Cross Sectional Disease Classification

The previous section replicated the analysis most widely used in existing literature. In this section, the exploratory and confirmatory statistical tests commonly performed by previous studies are extended via determining the ability of the features to also perform disease classification.

For each test, repeated 10-fold cross validation is performed using the full feature set from each test separately. Within each repeat of cross validation, the cross sectional baseline subset undergoes minority class balancing prior to being assigned to a fold. Within each fold, the training and validation sets undergo zero-mean unit-variance normalisation with respect to the mean and standard deviation of the training set. Feature selection (LASSO) is performed on each training set, with the selected features subsequently being extracted from the validation set[2].

Two classifiers are implemented: Logistic Regression (LR) and Random Forest (RF). In the case of RF classification, the training and validation sets do not undergo normalisation nor feature selection due to the inherent ability of RF to function successfully when presented with a high number of features (§3.2.4). These classifiers are chosen due to their popularity in the clinical and engineering fields respectively and enable a validity check via a comparison of their results.

Finally, a cross sectional classification task utilising the features from all three tests simultaneously is implemented - presenting this thesis' first study of a multi-source dataset. The feature sets from each of the three separate tests

[2] Although feature selection was performed during the confirmatory analyses in the previous section, the results of the confirmatory analyses are not utilised during the classification procedure so as to prevent data leakage between the training and validation sets.

are concatenated and used during this classification process. This classification task intends to reveal whether the features arising from each test present unique information and whether this benefits the classification performance.

The accuracy, sensitivity, specificity, and F_1-score are reported for each classifier alongside their respective standard errors.

Cross Sectional Disease Severity Regression

The inspection of the OxQUIP MDS-UPDRS scores of §4.2 demonstrated that the PD population possess a wide range of severities that partially overlap with those of the HC participants. As such, the use of a regression analysis may be more appropriate at disease quantification than binary classification which is insensitive to disease severity. In this section, the statistical tests of previous studies are extended via determining the ability of the features to perform disease severity regression using the Cross Sectional Baseline Subset.

As in the classification analysis, repeated 10-fold cross validation is firstly performed using the total feature sets from each test separately. Similarly, within each fold the training and validation sets again undergo zero-mean unit-variance normalisation with respect to the training set. Feature selection (LASSO) is performed on each training set using the binary outcomes, with the selected features being extracted from the validation set. This regression procedure is then implemented upon the multi-source dataset containing the feature sets of all three tests simultaneously.

Two regression algorithms are implemented: Linear Regression (LiR) and Random Forest. The mean and standard deviation of the root mean squared error (RMSE) and the mean absolute error (MAE) for all regression tasks are reported.

5.3.2 Longitudinal Disease Monitoring

Throughout this longitudinal disease monitoring analysis the features from all tests are used together. The focus of this section is to reveal novel longitudinal char-

acteristics of features and whether these characteristics influence the performance of classification and regression tasks.

Longitudinal Participant Subset

Two participant subsets are used during the longitudinal analysis:

a The *Longitudinal Classification Subset* are the participants who completed all tests during their first four visits. This number of visits results in performing longitudinal monitoring over a nine month period. The inclusion of participants with exactly four visits is based on a trade off between participant numbers and length of longitudinal analysis. As the OxQUIP study is still active and not all participants have completed the study, only 32 participants have made five or more visits at present, which is deemed an insufficient sample size to perform a longitudinal analysis upon.

b The *Longitudinal Regression Subset* contains every instance where all tests were completed. As such, each participant will contribute n_v instances of data where n_v is the number of visits where they have completed all tests. This results in the Longitudinal Regression Subset being the largest of the subsets.

The demographics of both longitudinal subsets are given in Table 5.1.

Longitudinal Feature Variation

This analysis investigates the longitudinal variation of features within the longitudinal classification subset. Two analyses are performed intended to determine if (i) the disease groups show a different degree of change in feature value over a nine month period and (ii) the disease groups show different amounts of feature fluctuation over a nine month period.

Firstly, paired t-tests are performed between the feature values at baseline and the feature values at the nine month follow up. This is performed separately for each disease group. This inspects whether feature values vary significantly at a participant-to-participant level between the two visits. This is a typical testing

technique employed by the small number of previous studies using IMUs on a longitudinal basis. The results of the of the limited number of previous studies can be extended to exploit the more regular data collection protocol of the OxQUIP study. Any feature identified as showing a statistical difference between baseline and nine month follow up in the PD group, but not in the HC group, undergo additional hypothesis testing. Paired t-tests are subsequently performed between the feature values at baseline and the feature values at the three and six month follow ups respectively. This enables the progression of features, relative to baseline, to be better understood at a three month frequency.

Secondly, it is possible to investigate the fluctuation of feature values over the nine month period. For each participant, the standard deviation of each feature value is calculated using the four instances contributed throughout the study. Independent t-tests are performed between the standard deviations of each feature between the disease groups. This therefore enables a comparison between the amount of longitudinal fluctuation shown by each disease group.

Longitudinal Disease Classification

When performing the cross-sectional classification task, any symptom variation identified by the previous analysis would not be accounted for which may prove detrimental to the classification accuracy. Two forms of longitudinal classification are performed in this section with the intention of determining whether the inclusion of longitudinal test data benefits the classification procedure.

Firstly, during *static classification* a classification model is trained and validated using the data contributed at each individual visit by the Longitudinal Classification Subset. This is equivalent to performing a cross-sectional classification task at each visit individually. This assumes the data contributed by each participant at each visit is independent and therefore this approach does not account for any longitudinal variation of feature values.

Conversely, during *dynamic classification*, the design matrix formed at visit v is dynamically updated to incorporate longitudinal feature variation. This is

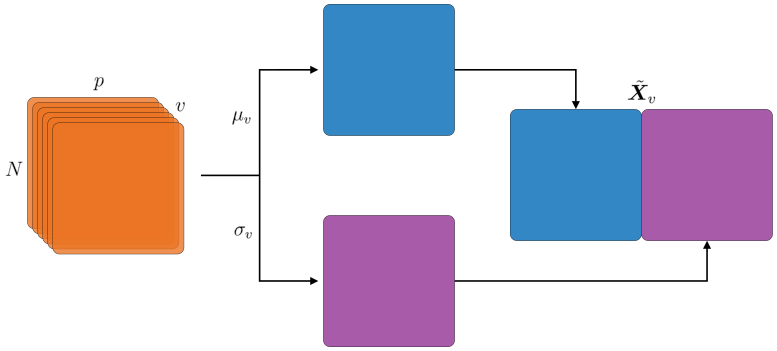


Figure 5.1: Visualisation of the design matrix formulation during dynamic learning classification.

achieved via utilising the current feature set from visit v in addition to all preceding feature sets. The design matrices from each separate visit are arranged into a $N \times p \times v$ tensor of N participants, p features, and v visits. The mean and standard deviation feature values across the visits are calculated and concatenated into a new design matrix, $\tilde{\boldsymbol{X}}_v$, which subsequently contains $2p$ features. As such, $\tilde{\boldsymbol{X}}_v$ is incorporating the longitudinal variation of each feature shown by each participant. This process is shown schematically in Figure 5.1.

In both classification approaches, logistic regression and random forest classifiers are trained and validated using repeated 10-fold cross validation. During dynamic classification, $\tilde{\boldsymbol{X}}_v$ is divided into training and validation sets with LASSO feature selection bring implemented on each training set.

Longitudinal Disease Severity Regression

In the Longitudinal Regression Subset, participants contribute multiple and a variable number of instances as data collection is still in progress. Therefore, during the training and validation of models, cross-validation *by participant* it utilised wherein all of a participant's instances are assigned to whichever set (training or validation) the participant is assigned. This is to ensure the validation set does not include instances from a participant who also contributed instances to the training set, as this has shown to highly bias results [237].

Repeated 10-fold cross validation is used which utilises the same feature normalisation and selection processes as in the baseline regression of §5.3.1. As all of

a participant's instances are either in the training or validation set, it is possible to visualise the estimated MDS-UPDRS of a single participant on a longitudinal basis. The RMSE and MAE are reported.

5.4 Results

Baseline Disease Quantification

Baseline Hypothesis Tests and MDS-UPDRS Correlations

Of the 108 features measured during the walking test, 38 were found to show significant difference in value between the disease groups. Of the 22 features measured during the sway test, only three were found to show significant difference between the disease groups. Of the 33 features measured during the sway test, only three were found to show significant difference between the disease groups. Tables 5.2, 5.3, and 5.4 provide the results of the independent t-tests and the Spearman correlation coefficients of the walking, sway, and TUG significant feature sets respectively.

Figure 5.2 is a correlation matrix demonstrating how the features within each of the tests are correlated. Furthermore, Figure 5.2 also demonstrates that feature correlations exist between tests (inter-source correlations). The presence of highly correlated features within each source (intra-source correlations) validates the use of a sparsity inducing feature selection technique. Table 5.5 provides the features selected as the most relevant by LASSO feature selection when implemented on each test individually. The features within each test are ordered by the magnitude of their weighting.

Cross Sectional Baseline Classification & Regression

The first four columns of Table 5.6 provide the cross sectional baseline classification metrics. These are given for each of the separate tests and for the multi-test classification.

Table 5.2: Hypothesis and correlation results for the walking significant feature set. TDS: Terminal Double Support, SLS: Single Limb Support, ROM: Range Of Motion, GCT: Gait Cycle Time (given as a % of total cycle time). These results correspond to the methods outlined in §5.3.1. A description of all features is given in Appendix A.1 and visualised in Figure A.1

Walk Feature	HC ($\mu \pm \sigma$)	PD ($\mu \pm \sigma$)	p-value	Correlation: R	Correlation: p-value
Gait Speed L (m/s)	1.24 ± 0.16	1.08 ± 0.27	0.001	-0.551	<0.001
Gait Speed R (m/s)	1.27 ± 0.16	1.11 ± 0.27	0.001	-0.542	<0.001
Foot Strike Angle L (degrees)	25.13 ± 5.21	20.31 ± 6.83	<0.001	-0.449	<0.001
Foot Strike Angle R (degrees)	26.52 ± 5.65	21.52 ± 6.77	<0.001	-0.518	<0.001
Toe Off Angle R (degrees)	38.57 ± 4.61	36.09 ± 6.09	0.019	-0.387	<0.001
Stride Length L (m)	1.29 ± 0.15	1.15 ± 0.22	0.001	-0.529	<0.001
Stride Length R (m)	1.32 ± 0.15	1.18 ± 0.22	0.001	-0.516	<0.001
TDS R (%GCT)	9.30 ± 1.41	10.25 ± 2.95	0.048	0.447	<0.001
Coronal ROM (degrees)	5.99 ± 1.77	4.78 ± 2.28	0.003	-0.149	0.083
Arm Swing Velocity L (degrees/s)	238.22 ± 59.67	181.62 ± 96.35	0.001	-0.271	0.001
Arm Swing Velocity R (degrees/s)	218.78 ± 57.82	184.38 ± 103.30	0.045	-0.229	0.007
Arm Range of Motion L (degrees)	59.64 ± 16.85	36.55 ± 22.65	<0.001	-0.366	<0.001
Arm Range of Motion R (degrees)	53.78 ± 17.64	37.54 ± 24.90	<0.001	-0.324	<0.001
Turn Angle (degrees)	175.96 ± 15.26	168.46 ± 19.77	0.030	-0.361	<0.001
Turn Duration (s)	2.09 ± 0.26	2.44 ± 0.58	<0.001	0.532	<0.001
Turn Velocity (degrees/s)	195.07 ± 36.69	153.39 ± 43.66	<0.001	-0.644	<0.001
Steps in Turn (n)	3.77 ± 0.55	4.16 ± 0.92	0.013	0.317	<0.001
Cadence L (steps/min) STD	2.18 ± 1.05	2.79 ± 1.45	0.014	0.542	<0.001
Cadence R (steps/min) STD	2.21 ± 1.04	2.85 ± 1.66	0.021	0.535	<0.001
Double Support L (%) STD	1.07 ± 0.28	1.42 ± 0.90	0.015	0.567	<0.001
Double Support R (%) STD	1.06 ± 0.28	1.42 ± 0.89	0.010	0.573	<0.001
Cycle Duration L (s) STD	0.02 ± 0.01	0.03 ± 0.03	0.029	0.530	<0.001
Cycle Duration R (s) STD	0.02 ± 0.01	0.03 ± 0.03	0.044	0.520	<0.001
Foot Strike Angle R (degrees) STD	1.74 ± 0.67	2.14 ± 0.79	0.004	0.344	<0.001
Toe Off Angle L (degrees) STD	1.07 ± 0.37	1.25 ± 0.55	0.047	0.348	<0.001
SLS L (%) STD	0.67 ± 0.16	0.94 ± 0.61	0.004	0.560	<0.001
SLS R (%) STD	0.72 ± 0.17	0.98 ± 0.66	0.010	0.561	<0.001
Stance L (%) STD	0.70 ± 0.18	0.97 ± 0.58	0.003	0.614	<0.001
Stance R (%) STD	0.64 ± 0.17	0.94 ± 0.61	0.002	0.588	<0.001
Step Duration L (s) STD	0.01 ± 0.004	0.01 ± 0.02	0.017	0.528	<0.001
Step Duration R (s) STD	0.01 ± 0.006	0.02 ± 0.02	0.050	0.506	<0.001
Stride Length L (m) STD	0.04 ± 0.01	0.04 ± 0.01	0.033	0.335	<0.001
Swing L (%) STD	0.70 ± 0.18	0.97 ± 0.58	0.003	0.614	<0.001
Swing R (%) STD	0.64 ± 0.17	0.94 ± 0.60	0.002	0.588	<0.001
TDS L (%) STD	0.68 ± 0.19	0.93 ± 0.56	0.005	0.577	<0.001
TDS R (%) STD	0.72 ± 0.19	0.97 ± 0.71	0.025	0.541	<0.001
Arm ROM L (degrees) STD	9.71 ± 6.80	7.06 ± 4.59	0.008	-0.122	0.158
Turn Velocity (degrees/s) STD	26.61 ± 13.77	19.42 ± 12.00	0.002	-0.352	<0.001

Table 5.3: Hypothesis and correlation results for the sway significant feature set. These results correspond to the methods outlined in §5.3.1.

Sway Feature	HC ($\mu \pm \sigma$)	PD ($\mu \pm \sigma$)	p-value	Correlation: R	Correlation: p-value
Mean Velocity (m/s)	0.14 ± 0.07	0.22 ± 0.17	<0.001	0.26	<0.001
Mean Velocity (Coronal) (m/s)	0.03 ± 0.03	0.05 ± 0.04	0.04	0.27	<0.001
Mean Velocity (Sagittal) (m/s)	0.13 ± 0.07	0.21 ± 0.17	<0.001	0.25	<0.001

Table 5.4: Hypothesis and correlation results for the TUG significant feature set. These results correspond to the methods outlined in §5.3.1.

Sway Feature	HC ($\mu \pm \sigma$)	PD ($\mu \pm \sigma$)	p-value	Correlation: R	Correlation: p-value
Duration (s)	9.28 ± 1.55	12.36 ± 6.84	<0.001	0.59	<0.001
Sit to Stand Duration (s)	0.98 ± 0.13	1.13 ± 0.18	<0.001	0.37	<0.001
Stand to Sit Duration (s)	0.84 ± 0.17	0.93 ± 0.20	0.01	0.39	<0.001
Turns Angle (degrees)	178.67 ± 5.47	169.40 ± 18.84	<0.001	-0.51	<0.001
Turn Duration (s)	1.97 ± 0.27	2.35 ± 0.57	<0.001	0.51	<0.001
Turn Velocity (degrees/s)	212.68 ± 38.89	166.90 ± 53.91	<0.001	-0.59	<0.001
Turns Angle (degrees) STD	5.67 ± 3.11	9.94 ± 10.50	0.01	0.42	<0.001
Turn Velocity (degrees/s) STD	26.17 ± 16.19	19.38 ± 13.17	0.01	-0.18	0.03

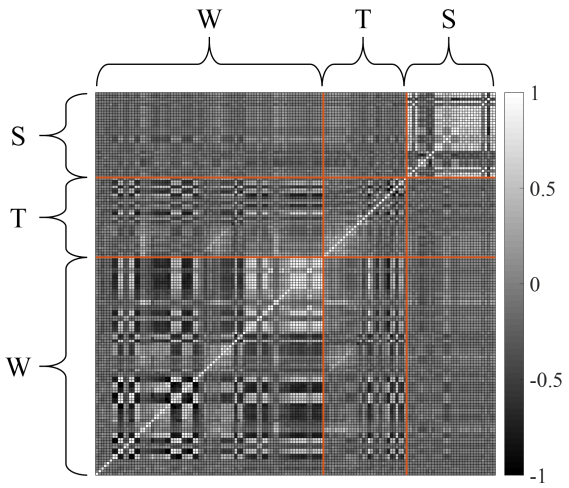


Figure 5.2: Inter- and intra-test feature correlations of the OxQUIP dataset. Units are Spearman's Rank Correlation Coefficient. W: Walking, T: TUG, and S: SWAY

The final two columns of Table 5.6 provide the baseline regression performances of each of the separate tests in addition to the regression performance of the multi-test regression.

Longitudinal Disease Monitoring

Longitudinal Feature Variations

Table 5.7 provides the features showing significance as identified by paired t-tests between the baseline and nine month follow up feature values. For each feature showing significance after nine months, the paired t-test results are also given between baseline and the three and six month follow ups.

Table 5.8 provides the results of the independent t-test results pertaining to feature fluctuation over the course of nine months. Figure 5.3 provides a visualisation

Table 5.5: The features and their corresponding weighting as selected from each test using LASSO. These results correspond to the methods outlined in §5.3.1. Key: % GCT - percentage of the gait cycle time.

Test	Feature	β
Walk	Gait Cycle Duration R (s) STD	-1.941
	Stride Length L (m) STD	-0.671
	Stride Length L (m)	0.354
	Gait Speed L (m/s)	0.181
	Single Limb Support L (%GCT) STD	0.160
	Steps in Turn (n)	0.028
	Coronal Range of Motion (degrees)	-0.013
	Foot Strike Angle R (degrees) STD	0.013
	Foot Strike Angle R (degrees)	-0.010
	Arm Range of Motion L (degrees)	-0.009
	Arm Range of Motion L (degrees) STD	-0.009
	Terminal Double Support R (%GCT)	-0.007
	Arm Range of Motion R (degrees)	-0.005
	Foot Strike Angle L (degrees)	-0.004
	Turn Velocity (degrees/s)	-0.003
	Arm Swing Velocity R (degrees/s)	0.002
	Turn Velocity (degrees/s) STD	-0.002
	Arm Swing Velocity L (degrees/s)	0.001
Sway	Mean Velocity (Sagittal) (m/s)	9.573
	Mean Velocity (m/s)	-9.362
	Mean Velocity (Coronal) (m/s)	4.606
TUG	Sit to Stand Duration (s)	0.501
	Turn Velocity (degrees/s)	-0.002
	Turns Angle (degrees)	-0.002

Table 5.6: OxQUIP Baseline Classification and Regression Results. These results correspond to the methods outlined in §5.3.1.

	Accuracy [%]	Sensitivity [%]	Specificity [%]	F_1 [%]	RMSE	MAE
	Logistic Regression & Linear Regression					
Walk	75.8 ± 1.5	74.7 ± 2.3	76.5 ± 2.2	74.4 ± 1.7	12.83 ± 0.85	9.91 ± 1.06
Sway	62.8 ± 1.5	52.8 ± 2.2	73.3 ± 2.2	57.0 ± 1.9	17.76 ± 1.02	14.45 ± 1.13
TUG	73.3 ± 1.5	69.0 ± 2.3	78.1 ± 2.0	70.7 ± 1.8	13.67 ± 0.86	11.17 ± 1.12
Walk + Sway + TUG	76.0 ± 1.3	73.5 ± 2.1	78.6 ± 2.0	74.4 ± 1.6	13.19 ± 0.45	10.73 ± 1.05
	Random Forest					
Walk	77.3 ± 1.4	72.7 ± 2.1	82.0 ± 2.0	75.1 ± 1.6	13.73 ± 0.75	10.87 ± 1.04
Sway	62.5 ± 1.5	51.9 ± 2.3	69.8 ± 2.2	58.9 ± 1.9	17.06 ± 0.92	13.87 ± 1.14
TUG	69.7 ± 1.4	70.5 ± 2.2	69.4 ± 2.2	68.7 ± 1.6	13.47 ± 0.68	10.74 ± 1.05
Walk + Sway + TUG	77.9 ± 1.3	75.6 ± 2.2	80.4 ± 1.9	76.3 ± 1.6	12.98 ± 0.73	10.28 ± 1.10

Table 5.7: Features that show a significant change over a three, six, and nine month period for PD participants, but not HC participants, as identified by paired t-tests. These results correspond to the methods outlined in §5.3.2.

	HC			PD		
Feature	3 Month	6 Month	9 Month	3 Month	6 Month	9 Month
First Step Range of Motion (degrees)	0.373	0.399	0.748	0.859	0.261	**0.030**
Duration (s)	0.107	0.373	0.374	0.471	0.252	**<0.001**
Toe Off Angle R (degrees)	0.310	0.548	0.405	0.316	0.124	**0.009**
Cadence L (steps/min) STD	0.390	0.612	0.752	**0.025**	**0.004**	**0.001**
Cadence R (steps/min) STD	0.391	0.756	0.831	**0.009**	**0.011**	**<0.001**
Double Support L (%GCT) STD	0.574	0.540	0.280	0.141	**0.027**	**0.007**
Double Support R (%GCT) STD	0.567	0.530	0.551	0.212	0.104	**0.007**
Gait Cycle Duration L (s) STD	0.262	0.186	0.804	0.175	0.070	**0.045**
Gait Cycle Duration R (s) STD	0.204	0.330	0.772	0.066	0.090	**0.021**
Gait Speed L (m/s) STD	0.846	0.691	0.881	**0.027**	0.056	**0.003**
Gait Speed R (m/s) STD	0.635	0.728	0.772	0.064	0.072	**0.009**
Toe Off Angle R (degrees) STD	0.708	0.702	0.513	**0.008**	0.109	**0.038**
Single Limb Support L (%GCT) STD	0.294	0.489	0.351	0.094	0.192	**0.005**
Single Limb Support R (%GCT) STD	0.644	0.407	0.720	0.732	0.612	**0.024**
Stance L (%GCT) STD	0.892	0.606	0.883	0.587	0.130	**0.029**
Stance R (%GCT) STD	0.627	0.445	0.351	**0.019**	0.091	**0.004**
Step Duration L (s) STD	0.428	1.000	1.000	**0.003**	**0.038**	**0.003**
Step Duration R (s) STD	0.666	0.330	0.330	0.595	0.399	**0.020**
Stride Length L (m) STD	0.460	0.117	0.297	0.057	0.203	**0.023**
Stride Length R (m) STD	0.899	0.297	0.287	0.166	0.336	**0.015**
Swing L (%GCT) STD	0.892	0.606	0.895	0.587	0.130	**0.029**
Swing R (%GCT) STD	0.618	0.445	0.351	**0.019**	0.091	**0.004**
Terminal Double Support L (%GCT) STD	0.829	0.722	0.609	**0.012**	**0.040**	**0.006**
Terminal Double Support R (%GCT) STD	0.301	0.965	0.353	0.926	0.924	**0.022**

of the fluctuation of a single feature for each disease group separately.

The MDS-UPDRS of the PD participants at the three, six, and nine month follow ups are 24.24 ± 13.3, 26.6 ± 14.6, and 28.0 ± 16.9 respectively. The MDS-UPDRS of the HC participants at the three, six, and nine month follow ups are 5.55 ± 7.00, 5.71 ± 5.41, and 4.21 ± 3.92 respectively. No significant changes were found between the baseline and nine month follow up MDS-UPDRS scores for the HC participants [p = 0.984] or PD participants [p = 0.719].

Longitudinal Classification

The accuracies of the static and dynamic longitudinal classification strategies over a nine month period are shown in Figure 5.4. Recall that static classification is equivalent to performing a cross sectional classification task at each visit separately using the longitudinal classification subset.

Table 5.8: The features showing a significantly different degree of fluctuation over a nine month period between disease groups. These results correspond to the methods outlined in §5.3.2.

Feature Name	HC	PD	p-value
Toe Out Angle L (degrees)	5.29 ± 3.13	3.97 ± 1.90	0.031
Elevation at Midswing L (cm) STD	0.05 ± 0.03	0.08 ± 0.06	0.033
Foot Strike Angle L (degrees) STD	0.27 ± 0.16	0.47 ± 0.40	0.036
Turn Velocity (degrees/s) STD	10.5 ± 6.27	7.19 ± 5.01	0.017
Duration (s)	0.67 ± 0.34	0.93 ± 0.54	0.042
Turns - Duration (s)	0.14 ± 0.07	0.22 ± 0.16	0.039
95% Ellipse Axis 2 Radius (m/s^2)	0.04 ± 0.03	0.09 ± 0.09	0.033
Centroidal Frequency (Sagittal) (Hz)	0.10 ± 0.05	0.16 ± 0.11	0.020
Path Length (Sagittal) (m/s^2)	0.82 ± 0.53	1.61 ± 1.79	0.049
RMS Sway (m/s^2)	0.02 ± 0.01	0.04 ± 0.04	0.034
RMS Sway (Sagittal) (m/s^2)	0.02 ± 0.01	0.03 ± 0.03	0.028
95% Ellipse Axis 2 Radius (degrees)	0.24 ± 0.15	0.50 ± 0.54	0.032
RMS Sway (degrees)	0.10 ± 0.06	0.21 ± 0.24	0.034
RMS Sway (Sagittal) (degrees)	0.10 ± 0.06	0.20 ± 0.20	0.028

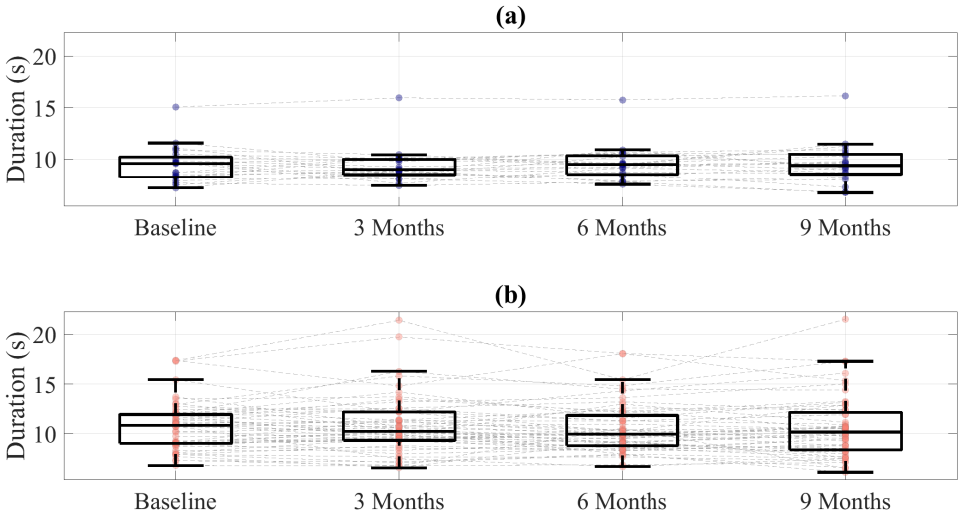


Figure 5.3: The fluctuation of the TUG test duration over a nine month period for (a) healthy controls and (b) Parkinson's participants. This feature was selected as it was identified as showing a significant pair-wise change over the course of nine months whilst also showing a significant difference in the degree of fluctuation between disease groups. These results correspond to the methods outlined in §5.3.2.

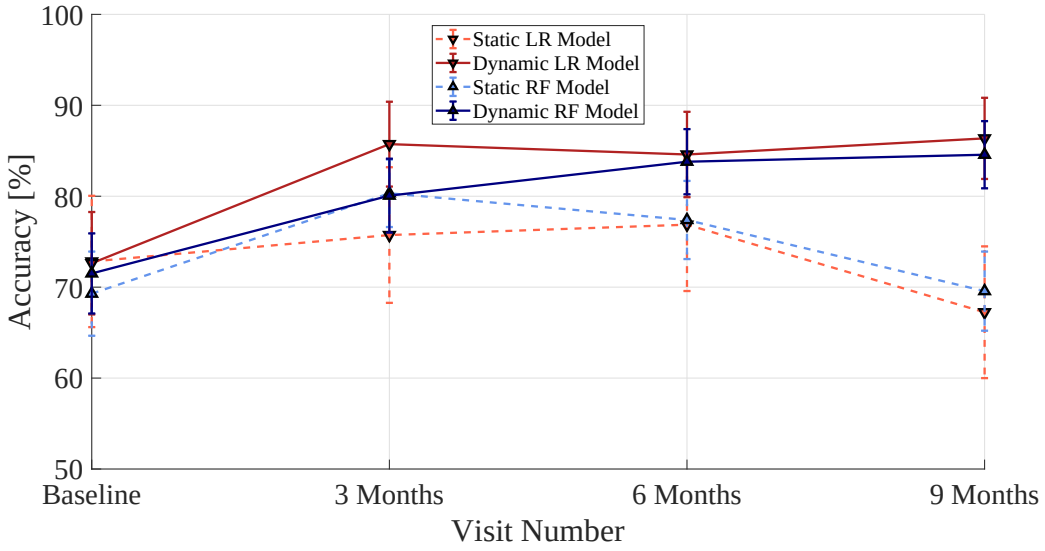


Figure 5.4: Longitudinal classification results of the static and dynamic approaches. These results correspond to the methods outlined in §5.3.2.

At baseline, the static and dynamic accuracies show no statistical differences for both the LR [$p = 0.652$] and RF [$p = 0.461$] models. Conversely, at 9 months, the classification accuracies between static and dynamic classification show significant differences when using LR [$p < 0.001$] and RF [$p < 0.001$]. There is no significant change in MDS-UPDRS for this subset over this period [$p = 0.719$] thus demonstrating that the improvement in accuracy is not a consequence of worsening disease severity.

Longitudinal MDS-UPDRS Regression

The results of the longitudinal regression for both regression algorithms (LR and RF) are given in Table 5.9. Figure 5.5 demonstrates the results of the longitudinal regression analysis for three participants who contributed data over a 12 month follow up period.

Table 5.9: The errors of the two regression algorithms when implemented on longitudinal data. These results correspond to the methods outlined in §5.3.2.

	RMSE	MAE
Linear Regression	12.13 ± 0.76	8.37 ± 2.1
Random Forest	12.11 ± 0.77	7.13 ± 2.3

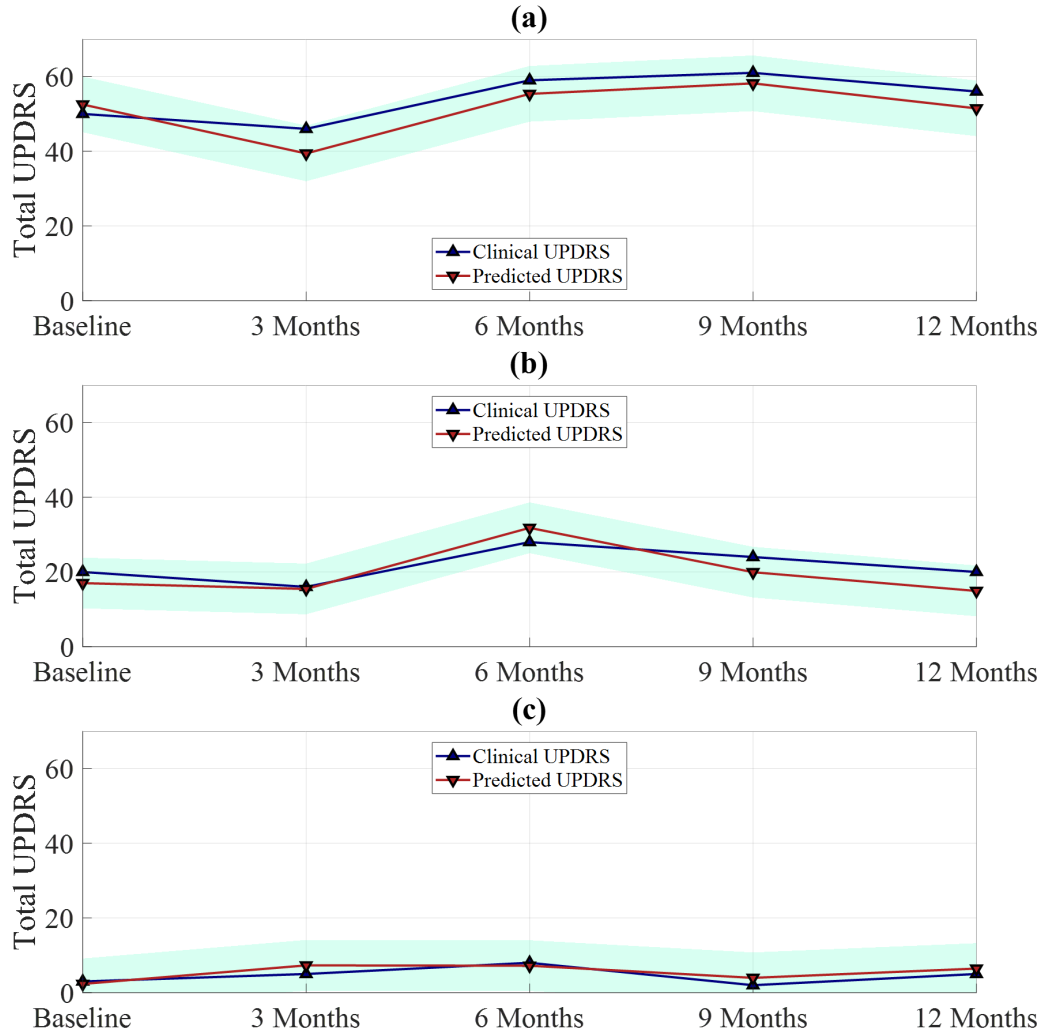


Figure 5.5: Longitudinal Regression demonstrations over a 12 month period for a (a) severe PD participant (b) mild PD participant and (c) a healthy participant. The green area represents the 5 - 95% confidence interval as calculated from the error of the test participants. These results correspond to the methods outlined in §5.3.2.

5.5 Discussion

The assessment of Parkinson's disease relies entirely on a clinically performed scoring system that is known to fall victim to multiple sources of subjectivity. Digital sensors have been suggested as a means of providing objective disease assessment but their clinical uptake have been hindered by the experimental inconsistencies of previous studies. In this chapter, the ability of digital sensors to provide objective disease assessment as part of routine clinical care has been investigated. Initially, a cross sectional confirmatory analysis used by previous work was replicated such

as to identify features that are distinct of PD and that show correlations with MDS-UPDRS. These findings were extended to assess the ability of these features to perform cross sectional disease classification and regression. Unlike previous studies, the OxQUIP study performed follow up assessments every three months facilitating a range of longitudinal analyses. The first longitudinal analysis in this chapter revealed novel longitudinal characteristics of features over the course of nine months; highlighting that PD participants often show a larger degree of variation and a larger amount of change over this period in many features. The inclusion of these longitudinal characteristics during classification proved highly beneficial when compared to the traditional cross sectional classification approach. Finally, a regression task was performed demonstrating the ability of these features to accurate track disease severity.

During the baseline cross sectional confirmatory analysis, many features were identified across the three tests that showed significant differences between PD and HC participants. Many of these features have been suggested by previous studies that utilised a smaller number of subjects such as PD participants having a decreased walking velocity, step length, and turning velocity. It is evident that many of these biomarkers are highly correlated with one another, potentially rendering their collection redundant. This is demonstrated in Figure 5.2 wherein it is visible that many inter-source and intra-source correlations exist. Most notably, the walking and TUG tests yield many highly correlated features; with each test also demonstrating a degree of inter-test correlations. Alternatively, the sway test produces many inter-test correlated features yet little intra-test correlations. As such, the application of LASSO feature selection demonstrated that a far smaller sub-set of features possess unique information, shown in Table 5.5. Further, many of these features also showed strong and significant correlations with the clinically assigned MDS-UPDRS scores.

An important observation that has been unattainable in previous studies is that many of the features of the confirmatory analyses originate from IMUs at different body locations. This is most notable in Table 5.2 where features pertaining to arm, leg, and torso motion are present. Furthermore, these features also suggest

that there is a distinction between the left and right hand side of the body for many of these features; thus highlighting the asymmetrical gait characteristics described in the literature [125, 306, 307]. This finding emphasises the importance of utilising a large sensor network as opposed to the potentially insensitive single sensor networks used by many previous studies [125].

When performing cross sectional classification using each test separately, the walking test yields the highest cross sectional classification accuracy of 77.3%. During chapter 2 the gait test generally reported higher classification accuracies when compared to the sway test (62.8%), consistent with the findings in this work. The TUG test performs similarly well (73.3%) to the walking test which is attributed to the high level of correlation between the walking and TUG feature sets demonstrating their near equivalence. However, this is the first work to compare the classification performance of these clinical tests using a consistent cohort and sensor network. Interestingly, the use of a RF classifier produces lower classification accuracies than the LR. Throughout this work, the same RF parameters were utilised for all classification tasks to increase interpretation and generalisation. Accordingly, each base learner of a random forest randomly uses $\sqrt{p}$ features from each feature set. As the LASSO implementation during the confirmatory analyses demonstrated, the sway and TUG tasks not only consist of fewer features than the walking test, but also produce less unique features (Table 5.5). As such, as no feature selection occurs prior to the design matrix being input to the RF, less nodes in the sway and TUG models are likely to contain distinctive features resulting in a degree of over-fitting occurring on the training data.

As would have been expected by the high level of correlation between many of the features and the MDS-UPDRS found during the correlation tests (Table 5.2, 5.3, and 5.4), the cross sectional baseline regression analysis showed a strong ability for the feature sets to predict MDS-UPDRS. Counterintuitively, many studies have performed correlation analyses between MDS-UPDRS and motor features yet not continued to implement a regression analysis [20, 105, 106]. As each MDS-UPDRS question is ranked from 0 - 4, a MAE of 10.28 (using all tests and a RF regression)

is deemed highly successful. As the maximum total MDS-UPDRS score is 260 (65 items), the MAE found here corresponds to an average error rate of 3.85% [71].

The cross sectional confirmatory analyses validated many findings of previous studies on a novel dataset whilst also proving the feature set capable of successful disease classification and severity regression. The longitudinal analyses aimed to extend these findings by revealing novel longitudinal characteristics of features and symptoms. Specifically, how features varied on a longitudinal basis was investigated in two ways.

Firstly, Table 5.7 presents the features showing a significant level of change between baseline and nine month follow up in the PD population. Importantly, these features do not show a significant level of change in the HC population. Further, a minority of these features show a significant level of change at the three and six month follow up. This result is suggestive of the analysis detecting the gradual worsening of symptoms over time. Upon completion of the OxQUIP study, this result can be validated via the inclusion of the remaining longitudinal instances. If a feature were to successfully detect symptom deterioration it would show and maintain a significant difference from the baseline value.

Secondly, to gain an insight into the symptom variance between visits the fluctuation of features over the course of four visits was examined on a participant-to-participant basis. Table 5.8 provides previously unseen characteristics of symptom fluctuation in a clinical environment. It was found that many features showed a wide degree of variation even though environmental conditions, experimental protocols, and medication states were kept constant. Indeed, 14 features were identified showing a significantly different degree of fluctuation in PD participants' than in HC participants'. This process is visualised in Figure 5.4 wherein the duration of the TUG test for each participant is shown to vary, often by a large amount in PD participants, between consecutive visits. However, it should still be noted that measuring of features at three month intervals is still insensitive to daily symptom fluctuations which may act as a confounding factor. Ideally, the use of a

more frequent data collection strategy could be employed enabling the detection of the subjects that consistently show deterioration or improvement.

During the longitudinal classification analysis, a comparison was made between traditional cross sectional classification tasks (static classification) and a classification task that incorporated longitudinal feature behaviour (dynamic classification); demonstrated in Figure 5.4. Dynamic learning proved to significantly improve the performance of classification whilst using both a LR and a RF classifier. Conversely, static classification showed no such improvement. Indeed, the variation of the static classification is caused by natural feature variation in the cohort. The improvement demonstrated by dynamic learning is attributed to two findings of the longitudinal feature variation analysis. Firstly, feature fluctuation was shown to differ significantly in many features between the disease groups. During dynamic classification, this fluctuation is incorporated into the design matrix via finding the standard deviation of each feature at each visit. Secondly, these fluctuations may also be thought of as noise, thus, via also including the mean feature value the fluctuation noise is filtered out to better represent the 'true' value. This has previously been suggested via demonstrating the intraclass correlation of step counts (measured using IMUs) increased significantly over the course of seven days in a cohort of multiple sclerosis participants [308].

There are several clinical and analytical limitations to this chapter. Firstly. a limitation that is present in previous studies and remains in this study is the inherent infrequency of clinical PD assessment. A participant undergoes assessment roughly every three months. It is widely reported that symptom severity varies at a far more frequent rate, especially during the early stages of disease progression [58, 114, 309]. However, this limitation is challenging to overcome as it is impractical to increase the frequency of clinical tests (§2.2).

Furthermore, this analysis has purposely not differentiated between disease groups. This is because the focus of this chapter is on differentiating HC from a PD cohort at a wide array of disease progressions. As previous studies have usually utilised PD participants with moderate to severe symptoms who are usually receiving

treatment, there exists limited evidence of biomarkers distinctive of early PD [37, 84, 132, 310–312]. Future work on this dataset will focus on identifying longitudinal characteristics that differentiate early stages of the disease (pre treatment) from later stages. At present, only 17 early stage participants have been enrolled with even fewer (N = 10) having contributed four or more test instances. The differentiation of typical and atypical Parkinsonian diseases is also an emerging research area for which the analysis in this chapter has been applied to using the PSP participants enrolled in the OxQUIP study. The differentiation of typical from atypical is particularly important as, although the symptoms can appear very similar, the treatments required differ.

Similarly, disease treatment, such as levodopa, is taken daily and induces daily variations in symptoms. The effects of treatment on biomarkers is a highly active area of research in the field of PD objectification but is not addressed in the present work. Although the OxQUIP dataset contains participants in both the ON and OFF states, this treatment response is beyond the scope of the current study but is an intuitive extension of these results. The effects of treatment on the biomarkers identified as distinctive of PD here should be investigated; with a particular emphasis of typical and atypical differentiation.

An analytical limitation of the current analysis is the lack of accounting for longitudinal variation in both feature values and their corresponding responses (MDS-UPDRS) during the longitudinal regression. Many more sophisticated methodologies are applicable to this scenario and whose use is motivated by these results. Such methodologies include various novelty detection strategies, Hidden Markov Models, and Gaussian Processes [210, 313–315]. Accounting for longitudinal variation in response and feature values using such methods may further improve the regression performance, as was observed during the dynamic classification task.

Finally, it is worthy to note that the OxQUIP study is still active and therefore the number of participants will continue to increase as will the number of longitudinal observations. The analysis procedures utilised throughout this chapter will be applied to the complete dataset upon completion.

5.6 Conclusion

This chapter has assessed the ability of wearable sensors to perform objective disease quantification as part of routine clinical care. Cross sectional statistical hypothesis tests revealed many features as showing significant differences between disease groups whilst also showing the ability to perform successful disease classification and severity prediction. The longitudinal analyses identified a previously unseen set of features that show different degrees of variation between the disease groups. A longitudinal classification which incorporated symptom variation was implemented and shown to significantly outperform the traditional cross sectional classification approach. Finally, a regression task was implemented demonstrating the ability to accurate track disease severity on a longitudinal basis. In the next chapter, longitudinal characteristics from additional motor and non-motor tests are explored regarding their ability to be monitored in a remote environment. As the data used in the following chapter was collected using smart-phones, tests are contributed at a much higher frequency than in a clinical setting as presented in this chapter.

6

Remote Monitoring of Symptom Progression Using Smart-phones

Related Publications:

1) J. Prince, S. Arora, M. De Vos, "Big data in Parkinson's disease: using smart-phones to remotely detect longitudinal disease phenotypes", Physiological Measurement, vol 39, no 4, p 0044005, April 2018.

6.1 Introduction

The previous chapter demonstrated how various forms of disease quantification can be achieved via using wearable sensors during routine clinical care. However, due to the OxQUIP study being constrained to a clinical environment, it still suffers from the limitations found in previous studies, namely the infrequency and inconvenience of tests.

This chapter steps out of the controlled clinical environment and aims to address the challenge of remotely performing longitudinal disease monitoring. It was suggested in chapter 5 that a larger degree of longitudinal feature fluctuation and change could be detected in PD participants when using tests that were three months apart. Many participants in the mPower study contribute tests on a daily basis, facilitating the inspection of feature fluctuation at a much higher frequency. Via extracting clinically validated metrics from a motor and non-motor test, their longitudinal behaviours are quantified such as to investigate whether learning periods can be detected in a remote environment and whether they differ between disease groups.

Initially, this chapter describes how a popular test used to clinically assess dexterity has been adapted by several studies to be quantified using digital sensors. The mPower dataset presents the opportunity to further these studies through the use of longitudinal monitoring on a larger cohort whilst also incorporating an additional non-motor (memory) test into the analysis. The ability of these tests to reveal differences in longitudinal behaviour and whether learning occurs at different rates between disease groups is assessed. Further, underlying characteristics pertaining to the self-reported MDS-UPDRS, participant demographics, and participant retention rates of the mPower dataset are investigated so as to better understand how data collected in a remote environment differs from that collected in a clinical environment.

The primary contributions of this chapter to the field of Parkinson's disease monitoring are:

- The first investigation into whether longitudinal motor and non-motor learning impairments can be detected in the large and remotely collected mPower dataset.
- An assessment as to whether disease groups demonstrate a different level of longitudinal symptom fluctuation and if a different number of test repetitions are needed for disease groups to reach a steady test performance.
- An exploration into how changes in longitudinal test performance can be quantified and subsequently if they better correlate with the MDS-UPDRS than baseline test performances.
- Determining whether the longitudinal behaviour of motor and non-motor symptoms are related.
- An inspection into the potential limitations of the self-reported MDS-UPDRS system via determining the floor-ceiling effects of the individual scoring system's elements.

6.2 Background

As discussed in Chapter 2, the current requirement for PD diagnosis and treatment to be performed in-clinic incurs a substantial cost for healthcare services. Additionally, with the projected number of people with PD to increase dramatically over the next decade, the ability to reliably perform assessments outside of a clinical environment is highly desirable [19]. These challenges, in addition to the many shortcomings of the MDS-UPDRS, have led to the active area of research identifying objective biomarkers to quantify the severity symptoms associated with PD though the use of digital sensors [20, 316].

In a clinical environment, the Alternating Finger Tapping (AFT) activity has demonstrated that subjects with PD show an impaired motor-performance when compared to healthy subjects [157, 161, 317]. Such impairments include hastening, faltering, or freezing during the AFT test. Subsequently, these impairments have been detected through digital sensors resulting in features such as tapping speed and rhythm being identified as showing strong capabilities of predicting symptom severity [138, 139]. Additionally, quantitative approaches to detecting non-motor impairment in PD subjects have also shown promise through memory examinations such as the Serial Reaction Time (SRT) Test. Findings commonly suggest that subjects with PD show impaired sequence and implicit learning when compared to healthy subjects [197, 199, 318].

Reiterating the primary downfall of previous studies utilising wearable sensors stated in Chapter 2; they suffer from low subject numbers and lack of longitudinal data as they were confined to a hospital environment, which seriously limits the scalability of their studies as data collected in a home environment can be more confounded by noise. Table 6.1 provides a summary of the most recent AFT studies demonstrating their subject and instance limitations. Presently, the degree of scalability of previous findings is unknown as validation on long time-scales, on a large subject group, and in a non-clinical lab environment has yet to be undertaken. Furthermore, both the motor and non-motor learning impairments discovered in subjects with PD have only been suggested over a very short period of

Citation	# Of Subjects (PD/HC)	# of Measurements	Measurement Device	Location
(Arroyo-Gallego et al., 2017) [156]	21/23	51	smart-phone	Clinic
(Kassavetis et al., 2016) [228]	14/0	14[1]	smart-phone	Clinic
(Picillo et al., 2016) [140]	123/0	492[2]	Video-camera	Clinic
(Arora et al., 2015) [99]	10/10	18[3]	smart-phone	Remote
(Printy et al., 2014) [138]	18/0	54	smart-phone & Glove	Clinic
(Memedi et al., 2013) [161]	95/10	12,011[4]	Handheld Computer	Clinic & Remote
(De Frias et al., 2007) [151]	50/48	196	Response Console	Clinic
(Tavares et al., 2005) [139]	33/0	66	MIDI-Keyboard	Clinic
(Lee et al., 2016) [230]	57/87	432	smart-phone	Clinic
(Zhan et al., 2016) [79]	121/105	1,600[5]	smart-phone	Remote
(Kraus et al., 2005) [319]	411/0	NA	Contact Pencil	Clinic
(Bot et al., 2016) mPower [78]	8,003	78,883	smart-phone	Remote
(This study) mPower	312/236	48,892	smart-phone	Remote

Table 6.1: Summary of the most recent studies quantifying a tapping activity in Parkinson's disease. The data collected in the mPower study, shown in the bottom row, is used in this study.

time (usually within 24 hours), commonly over the course of several test repetitions or a single clinical visit [153, 157].

The advancement of micro-electronic systems has enabled wearable sensors to become more commonplace in everyday life, most notably though their embedment in smart-phones. Recent studies into PD have created smart-phone applications enabling participants to perform tests multiple times a day, over the course of many months, all in a non-clinical environment [78, 79]. The accelerometer in smart-phones allows assessment of gait, tremor, and balance, whilst the touch-screen is used for the AFT activity and for memory games. Subsequently, these studies have collected databases which contain tests over multiple modalities on a longitudinal basis from greatly increased subject numbers. smart-phones also enable the user to perform regular self-assessments regarding their disease severity,

[1]Not stated, assumed to be one measurement per subject.

[2]One measurement for each hand in each medication state (ON/OFF).

[3]Average of 2.7 tests per day for an average of 34.4 days ($\sim$ 93 tests) assumed to be split evenly over five test types.

[4]Coming from multiple studies. Measurement frequency is highly inconsistent with between $\sim$28 to 2 measurements per subject and no longitudinal analysis was performed.

[5]8,000 instances in the study are assumed to be split evenly across five modalities (voice, gait, balance, reaction, & dexterity.)

symptom prevalence, and medication adherence. However, many challenges have been identified when collecting remote data compared to clinical data [237, 316]. The variability of test environment, smart-phone placement, and smart-phone orientation have a significant influence on the test data. It is challenging to determine whether differences between tests are due to environmental factors or caused by a longitudinal change in impairment. Findings from clinical datasets suggest subjects with PD show impairments in motor and non-motor learning in the AFT and memory tests on a short term basis [157, 197].

6.3 Methods

6.3.1 Data Description

This chapter focuses purely on the data acquired in the tapping and memory activities from the mPower study [78].

Although many features from the AFT have shown to be discriminative between PD and HC subjects, the focus of this work is on the 'Total number of taps' completed during each test instance as it is a highly interpretable feature that has been repeatedly validated in a clinical environment, often showing the strongest relation to disease severity.

As the memory activity was included in the mPower study in a later version than the other activities, it subsequently has significantly fewer test instances. Similarly to in the tapping activity, the focus of this work is on a single and highly interpretable feature; the 'Total memory score'.

6.3.2 Short Term Behaviour

It has been shown in clinical environments that PD and HC participants show variable performances over a small number of repetitions at the AFT test [153]. The first aim is to determine if PD participants show an impaired performance at the AFT. Secondly, it is investigated whether the PD participants' performance changes at a different rate to HCs over a short number of repetitions and whether these changes

could be induced by a response to medication; enabling the differentiation of the performance change due to learning from the performance change due to medication.

In the tapping activity, to be included in the short term behaviour (STB) analysis, participants were required to have contributed 5 or more test instances within the first 24 hours of their first test instance (n = 600). Due to the lower number of participants and instances in the memory activity, the inclusion criteria was altered so that participants who contributed 5 or more test within the first 72 hours of their first test were included (n = 112).

By inspecting the relative test performances between the PD and HC participant groups it is possible to determine whether PD participants show impairment at baseline. Next, a comparison of the performance of each participant group after the fifth repetition of an activity with their baseline performance is implemented using the student t-test. The choice of limiting this analysis to the first five tests is based on previous works demonstrating that both the AFT and SRT tests show learning within five repetitions [199, 320]. In order to determine if medication is having an effect on test performance, the above procedure is repeated on all participants who are in the same medication state at baseline and at fifth visit, thus isolating the change in performance due to practise via removing any influence of medication.

6.3.3 Longitudinal Behaviour

Previous studies focus on the short term behaviour of the AFT test as they did not have access to longitudinal tests (Table 6.1). As the mPower database contains longitudinal data, it is possible to extend the analysis to determine whether longitudinal variation occurs within the performance of HC and PD participants. Adopting a similar approach to the study of the short term behaviour, it is investigated whether PD participants are more likely to show transient behaviour and whether the rate at which PD participants learn new tasks is different to HCs. As each test is labelled with a medication time-stamp, an attempt is made to determine if PD participants who are not taking medication are more likely

to show transient behaviour than those who are taking medication. This high-frequency analysis may provide useful insight into the effect of medication on test performance, and help clinicians make informed decisions to optimize drug dosage and times for each individual PD participant.

For the tapping test, any participant who contributed 20 or more tapping test instances is included in the long term behaviour (LTB) analysis (548 participants). Alternatively, due to there being less memory test instances, any participant who contributed 10 or more memory test instances is included in the long term behaviour analysis (121 participants).

Progression Ratio

To quantify the longitudinal performance of a participant the Progression Ratio (PR) metric is defined. The PR continuously compares a participant's average starting performance to their performance at any other given test instance. When this process is repeated for all instances, multiple PR values are obtained which form the Progression Ratio Waveform (PRW).

If a participant completes a total of N instances of a test, and an averaging window size of n (where $N >> n$) is selected, the PRW will have $J = N - (n-1)$ test points. The j^{th} point in the PRW is calculated via:

$$\mathrm{PRW}_j = \frac{\frac{1}{n}\sum_{j}^{j+(n-1)} f_j}{\mu_{BV}} \tag{6.1}$$

where:

$$\mu_{BV} = \frac{1}{n}\sum_{j=1}^{n} f_j \tag{6.2}$$

with f_j being the feature value at test number j. The PRW is taking the average feature performance over the first n tests, and then finding the ratio between this average starting performance and the average performance of all other windows (of length n). In this study, $n = 5$. The averaging window size of five is used as this is consistent with and builds on the previous short term analysis which focused

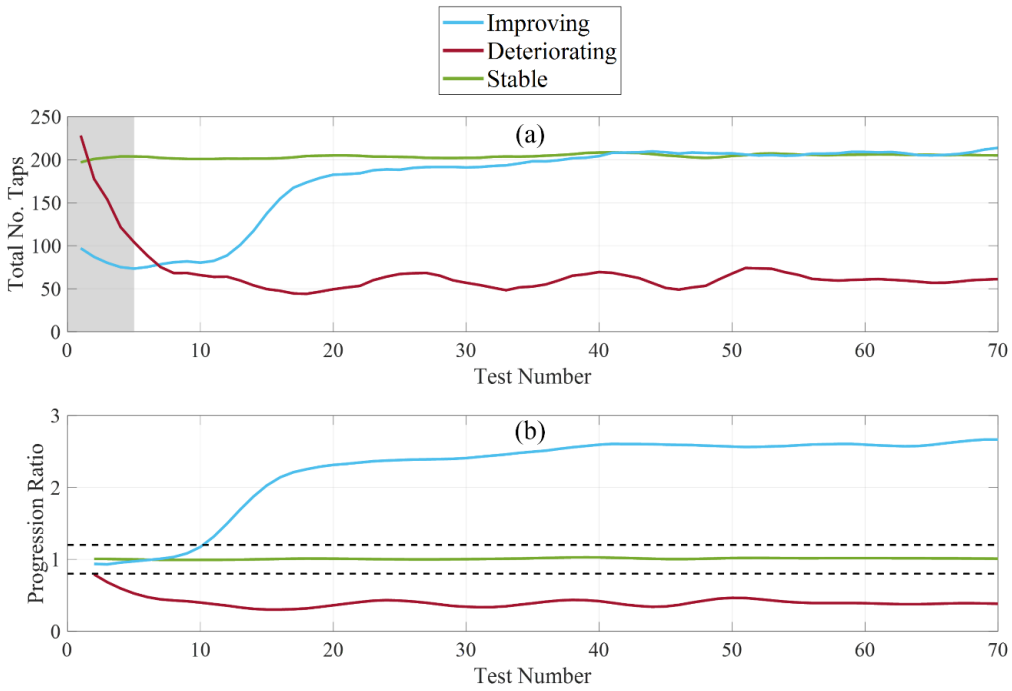


Figure 6.1: Visualisation of the PWR formation for three participants. Plot (a) shows the number of taps by each participant in their first 70 test instances. The instances that fall within the grey box are averaged and form σ_{BV} in Equation 6.2. Plot (b) shows the corresponding PRW for each participant with the Improving and Deteriorating thresholds (1.2 and 0.8 respectively) shown by the dotted lines.

on the first five test performances. This enables the separation of the learning behaviours seen in the short and long term analyses.

According to Equation 6.1, if a feature value increases over repetition, the PRW will tend to be greater than one, whereas if the feature value decreases over repetition, the PRW will be less than one. A feature showing little or no change over time will have a PRW at a constant value of one. Participants are subsequently classified into Learning-Phenotypes (L-PTs) based on μ_{prw}, the mean of their PRW, according to:

$$\text{L-PT} = \begin{cases} \mu_{prw} > 1.2 & \text{Improving} \\ \mu_{prw} < 0.8 & \text{Deteriorating} \\ 0.8 \leq \mu_{prw} \leq 1.2 & \text{Stable} \end{cases} \tag{6.3}$$

The μ_{prw} therefore enables the determination whether a participant's performance is varying longitudinally. In Equation 6.3, the thresholds for classification are based on observing an average performance change of 20% from baseline. Previous investigations into short term learning in the AFT and reaction time tests report

performance changes of between 16 - 24% [153, 321]. As the data in this study is collected remotely, test performances can vary as a consequence of test environment, thus, using a value of 20% ensures only subjects who show a consistently different longitudinal performance with their baseline performance are classified as being either improving or deteriorating. Figure 6.1(a) demonstrates the longitudinal behaviour of three participants (one from each longitudinal phenotype) and Figure 6.1(b) shows their corresponding PRWs.

Steady State Index

In addition to detecting the presence of transient and varying longitudinal performance in a new task, an additional aim is to determine if PD participants take longer to learn this task. An additional metric, the Steady State Index (SSI), is defined which estimates the test number at which a participant reaches a steady state of performance. The SSI continuously compares a participant's average performance over their final n tests, which is assumed to be their Steady State Performance (SSP), to their performance at any other test instance. In the same manner as the progression ratio waveform was calculated, the Steady State Waveform (SSW) is calculated as:

$$\mathrm{SSW}_j = \frac{\frac{1}{n}\sum_{j}^{j+(n-1)} f_j}{SSP} \tag{6.4}$$

where:

$$SSP = \frac{1}{n}\sum_{j=N-n}^{N} f_j \tag{6.5}$$

Whereas for the progression ratio waveform wherein the ratio between current performance and the mean starting performance was found, when determining the SSW the ratio between the current performance and the mean finishing performance is found. In order to determine the Steady State Index, the indices of tests whose performances are within 20% of the SSP are found and denoted in a binary vector:

$$\mathbf{I} = (SSW > 1.2)\,\&\,(SSW < 0.8) \tag{6.6}$$

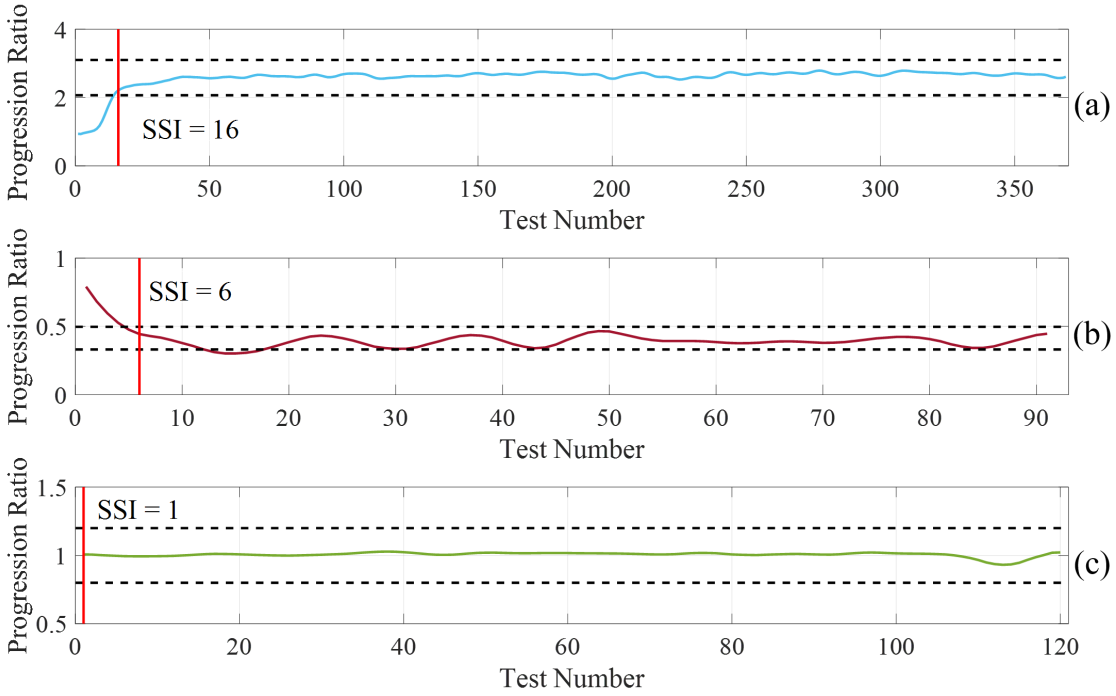


Figure 6.2: Visualisation of the SSI determination for same three participants as presented in Figure 6.1. The dotted lines represent the final value ±20% thresholds for each participant.

From the binary **I** vector, the index of the first two consecutive TRUE instances is set as the SSI.

Figure 6.2 provides a visualisation of the calculation of the SSI for the same three participants as presented in Figure 6.1.

6.3.4 Relation to Disease Severity

Features from the AFT activity and memory function tests have been shown to correlate well with clinically assigned severity [139, 194, 322]). However, the above studies have mostly consisted of one-off clinical visits on a small cohort of subjects with mild/severe severities. The above findings are yet to be tested on a large cohort of subjects, using remotely collected data, encompassing a wide range of MDS-UPDRS severity scores. This work expands on previous approaches via incorporating the longitudinal progression of participant performance. To investigate the potential links between longitudinal performances and disease severity the correlation between

disease severity and participants' steady state performances are found and compared with the correlation between disease severity and baseline performances.

The PD participants who were included in the Longitudinal Behaviour Subset and who also had at completed least one MDS-UPDRS assessment were included in this analysis. This subset is herein referred to as the Performance-Severity Subset (PSS).

The reliability of the mPower MDS-UPDRS survey is further explored via determining if the 'Floor-Ceiling' effect is present in the mPower subset of questions. The floor-ceiling effect occurs in a scoring system when the extremity scores are insensitive to small changes in the parameter being measured, resulting in a high proportion of scores taking the highest or lowest possible values. Firstly, the PSS is divided into two groups based on their total MDS-UPDRS score severity. The correlations between the test performances and severity scores of severe participants are compared against the same correlation in not-severe participants. A participant is labelled as Severe if their total MDS-UPDRS score is above the mean plus one standard deviation of the population's total MDS-UPDRS score (MDS-UPDRS threshold = 20.0). Secondly, participants are divided into two groups based on the test performances. The correlations between the test performances and severity scores of participants with severely impaired performances are compared against the same correlation in participants whose performances are not severely impaired. A participant is labelled as having a severe performance if their baseline performance is below the mean minus one standard deviation of the population's baseline performance (tapping threshold = 99.1 and memory threshold = 191.8). Finally, the relationship between disease severity and the number of years since diagnosis is investigated. This is firstly performed on all PD participants who contributed a MDS-UPDRS survey (N = 534). Secondly, the PSS cohort is split into two groups based on how recently the participants received their professional diagnosis. The correlation between the years since diagnosis and the participants' MDS-UPDRS scores and test performances are found. A participant is labelled as having a recent diagnosis if they received their diagnosis within the last seven years. Conversely, a participant is labelled as having a longstanding diagnosis if they have had a diagnosis

for seven years or longer. The seven year threshold was chosen as it is the mean plus one standard deviation of the years of diagnosis for PD participants in both the tapping and memory tests. Finding these correlations, in a large population with a wide range of severity scores, will enable conclusions to be drawn regarding whether previously clinically found features are suitable for severity prediction on subjects with low/mild severities in a remote environment. Finally, for each individual question in the mPower MDS-UPDRS survey, as well as the total MDS-UPDRS score, the floor-ceiling effects are evaluated via calculating the percentage of responses that are the highest and lowest possible values.

6.3.5 Relationship between Tapping and Memory Tasks

This analysis aims to find if a relationship exists between the longitudinal behaviour in the tapping and memory activities. All participants who were in the Longitudinal Behaviour Subsets for both the tapping and memory activities (n = 107) are selected and it is investigated whether their longitudinal behaviour is consistent between the activities.

6.3.6 Disease Characteristics between Sexes

It has been suggested that oestrogens may have a neuroprotective effect against PD; in animals the substantia nigra has shown to be more susceptible to neuron degradation in low oestrogen environments [323]. Several studies have demonstrated that PD incidence can be as much as two times higher in males than in females [65]. Subsequently, due to their lower oestrogen levels, it is proposed that males are more likely to develop PD at a younger age than females.

Utilizing the entire mPower demographics dataset, it is investigated whether the proportion of male and female participants with PD is in agreement with previously reported incidence between sexes. For the PD participants who completed the demographics survey and at least one tapping test (N = 1,060) or at least one memory test (N = 297) the effects of sex on smart-phone test performance (as quantified using the number of taps and memory score) is studied. The study is

	Tapping			Memory		
	N	Age	MDS-UPDRS	N	Age	MDS-UPDRS
	Short Term Behvaiour					
YHC	406	30.2 ± 8.8	N/A	10	32.8 ± 12.1	N/A
HC	57	60.3 ± 7.7	N/A	11	64.2 ± 10.4	N/A
PD	137	63.4 ± 7.3	N/A	91	65.5 ± 6.2	N/A
	Longitudinal Behvaiour					
YHC	150	32.5 ± 9.5	N/A	10	34.3 ± 9.1	N/A
HC	86	61.9 ± 7.7	N/A	14	65.1 ± 7.0	N/A
PD	312	63.8 ± 6.8	N/A	97	65.6 ± 6.3	N/A
	Performance Severity Subset					
Severe MDS-UPDRS	28	64.3 ± 6.2	27.6 ± 4.1	16	65.5 ± 7.3	25.8 ± 3.2
Not Severe MDS-UPDRS	233	63.9 ± 6.8	9.6 ± 5.9	67	63.8 ± 6.5	9.6 ± 5.6
Severe Performance	59	63.6 ± 6.8	13.5 ± 8.1	9	64.1 ± 6.4	13.4 ± 7.7
Not Severe Performance	204	63.9 ± 6.7	10.5 ± 7.4	74	63.6 ± 6.5	10.5 ± 6.9
Longitudinal Diagnosis	48	62.6 ± 7.5	17.0 ± 8.8	18	64.9 ± 7.5	16.5 ± 7.32
Recent Diagnosis	215	64.1 ± 6.5	11.5 ± 6.8	65	65.8 ± 6.0	11.3 ± 6.3

Table 6.2: The characteristics of participants in each of the three main analysis subsets. The Short-Term Behaviour (STB) and Longitudinal Behvaiour (LTB) subset inclusion criteria do not require participants to have contributed a MDS-UPDRS survey, they therefore contain many participants who did not contribute any MDS-UPDRS data and hence this has been omitted.

concluded by looking at the disease characteristics between sexes via implementing Binomial proportion tests between the male and female participants within each of the L-PTs. This is performed for the tapping and memory tests, and is aimed to determine whether participant sex is a confounding factor in L-PT classification.

6.3.7 Data Subsets

For the first two analysis sections outlined above, the eligible participants are further divided into three subsets; Young Healthy Control, Age Matched Healthy Controls, and Parkinson's Disease participants. These three groups are denoted by YHC, HC, and PD respectively. The YHC subset contains any participant under the age of 50 years old who self-reports as not having PD. The HC subset contains any participants who are 50 years old or older and who report as not having PD. The PD group contains any participant who self-reports as having a professional diagnosis of PD. The characteristics of all subsets can se seen in Table 6.2.

6.4 Results

Tapping

Short-Term Behaviour

At baseline performance, PD participants show significant impairment compared to both HC [p <0.001] and YHC [p <0.001] participants. Similarly, after five repetitions PD participants remain significantly impaired compared to HC [p = 0.02] and YHC [p < 0.001] participants.

Performing paired t-tests between the baseline performance and performance at the fifth repetition for the YHC, HC, and PD groups yields p-values of 0.60, 0.27, and 0.03 respectively, demonstrating that the only participant group showing a significant level of change at the 0.05 level are the PD participants.

Further, for all participants it is found that the change in performance over the first five test instances correlates with baseline performance [R = -0.48]. This intuitively states that participants with a lower baseline performance tend to improve by a larger degree than participants at a higher baseline performance.

Across each of the first five test instances, the PD participants showed a higher level of performance fluctuation (58 ± 1.25 taps) than both the HC (48 ± 1.8 taps) and YHC (47.8 ± 1.3 taps) subsets.

Of the 137 PD participants, 11 reported that do not take any medication for their Parkinson's disease symptoms, and the remaining 126 participants reported taking medication. No difference is seen in baseline performance between participants who take medication and those who do not [p=0.58].

Of the 126 participants who take medication, 22 reported being in the OFF state ('before medication') and 33 reported being in the ON state ('after medication') at baseline. No difference is seen in performance between the participants in the ON and OFF states [p = 0.59]. Seventy participants were identified as being in the same medication state at baseline and at fifth repetition. These participants also showed a significant change between baseline and fifth visit [p = 0.04] thus demonstrating that the significant changes in performance are not a response to medication.

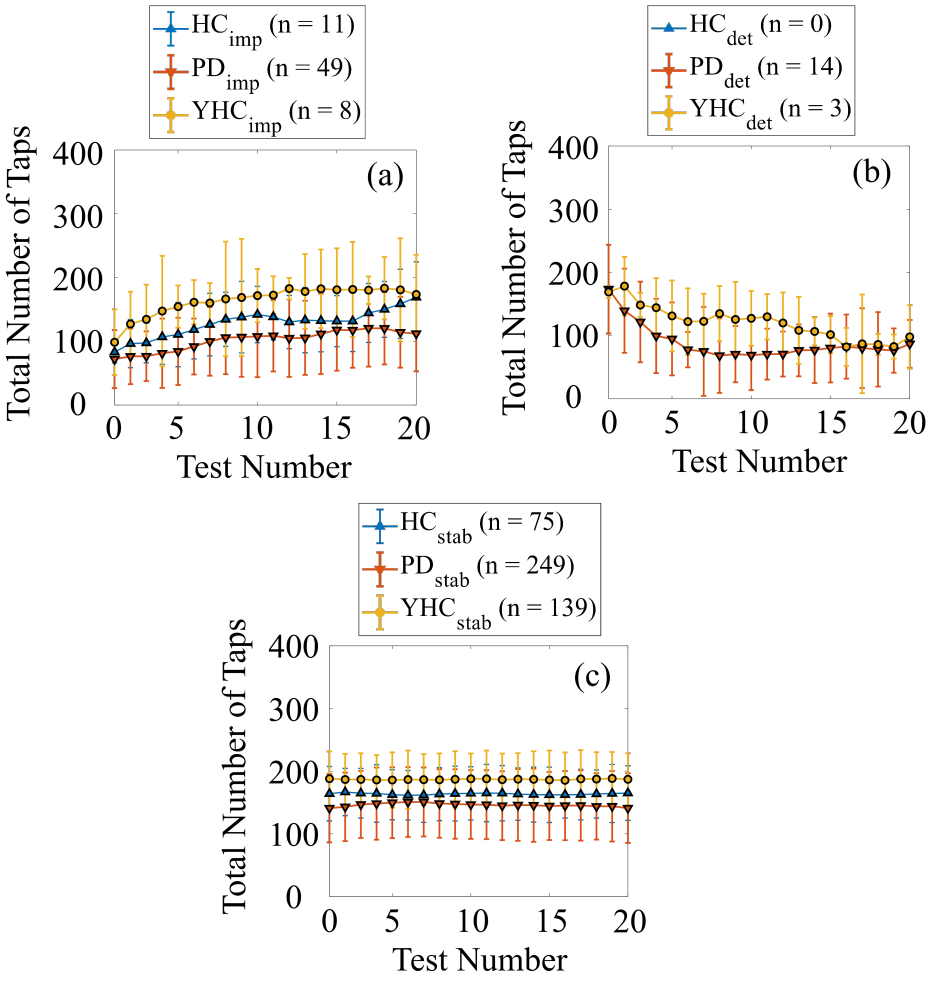


Figure 6.3: The longitudinal behaviour over the first 20 test instances in the tapping activity for participants identified as (a) improving (b) deteriorating and (c) stable. These results correspond to the methods outlined in §6.3.3 when applied to the tapping activity.

Longitudinal Behaviour

In the Longitudinal Behaviour Subset data, significant impairment between the PD and HC participants [$p<0.001$] and the PD and YHC participants [$p<0.001$] in baseline performance is found.

Each participant's Learning-Phenotype is determined based on Equations 6.1 and 6.3. The longitudinal performance of each participant group, and each of the three L-PTs, can be seen in Figure 6.3.

Over 20 test instances, the improving YHC, HC, and PD participants increased their performance by an average of 75, 86, and 40 taps respectively [Figure 6.3(a)].

	% of Sub-Group	Baseline Value ($\mu \pm \sigma$)	Final Value ($\mu \pm \sigma$)	Progress Ratio ($\mu \pm \sigma$)	Steady State Index ($\mu \pm \sigma$)
			YHC (N = 150)		
Improving	5.3	96.0 ± 51.6	182.8 ± 94.9	1.33 ± 0.18	5.3 ± 4.1
Deteriorating	2	172.3 ± 9.5	122.7 ± 68.1	0.78 ± 0.02	7.3 ± 1.5
Stable	92.7	183.9 ± 43.5	180.0 ± 43.4	1.00 ± 0.07	2.0 ± 4.5
			HC (N = 86)		
Improving	12.8	80.6 ± 16.8	115.8 ± 51.5	1.30 ± 0.09	4.3 ± 3.1
Deteriorating	0	N/A	N/A	N/A	N/A
Stable	87.2	157.0 ± 43.4	165.0 ± 35.5	1.01 ± 0.08	1.1 ± 1.0
			PD (N = 312)		
Improving	15.7	81.7 ± 45.5	136.4 ± 57.4	1.46 ± 0.45	16.6 ± 26.6
Deteriorating	4.5	146.6 ± 70.2	76.6 ± 30.1	0.71 ± 0.08	11.3 ± 10.7
Stable	79.8	136.8 ± 54.7	140.0 ± 54.9	1.01 ± 0.09	2.3 ± 5.4

Table 6.3: Baseline and longitudinal characteristics of each of the three Learning Phenotypes in the three participant groups in the tapping activity. These results correspond to the methods outlined in §6.3.3 when applied to the tapping activity.

The improving PD and HC participants showed a significant difference between their baseline performance and 20^{th}test performance, both with $p<0.001$, whereas the improving YHC participants approached significance with $p=0.051$.

No significant changes in performance were seen for any of the stable YHC, HC, or PD participants. The stable PD participants maintained a significant impairment in performance compared to the HC and YHC participants over the first 20 test instances [Figure 6.3(c)].

The proportion of each sub-group showing transient (not stable) behaviour can be seen in Table 6.3. The Binomial proportional test is used to determine if PD participants are more likely to show transient behaviour than HCs. It is found that a significantly larger proportion of PD participants show transient behaviour compared to the YHC participants [$Z = 3.67$, $p < 0.001$] whereas the proportion of PD and HC participants showing transient behaviour approaches significance [$Z = 1.56$, $p = 0.059$]. To account for the effects of age, a Binomial proportional test was also performed between the YHC and HC participants yielding insignificant differences in the number of transient participants between the groups [$p = 0.134$].

In addition to determining if PD participants are more likely to show transient behaviour, the Steady State Index (SSI) is used to measure if PD participants take

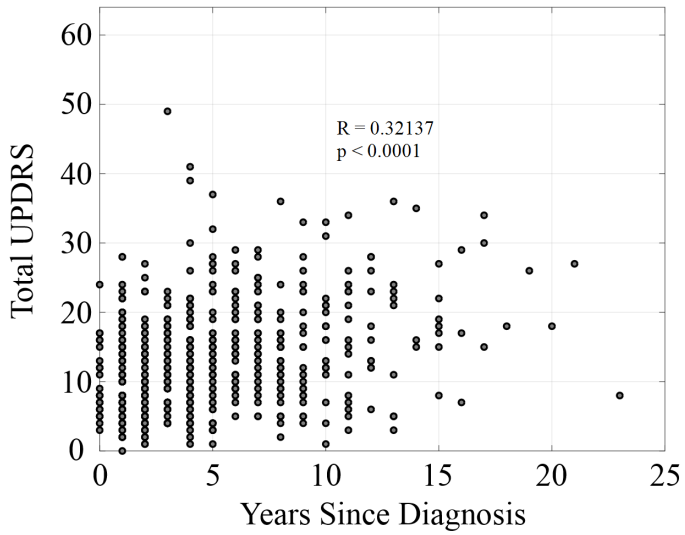


Figure 6.4: The correlation of the total MDS-UPDRS with years since diagnosis. These results correspond to the methods outlined in §6.3.4.

longer to learn a task. The PD participants show significantly larger SSI values than the HCs [p = 0.013] and YHCs [p = 0.015]. Further dividing the PD group into its L-PTs (final column of Table 6.3), it is seen that this difference is caused by the much larger SSIs of the improving and deteriorating participants. Across the entire Longitudinal Behaviour subset, the time (in days) of each test relative to the first test is found to correlate very highly with test index [Pearson R = 0.95], validating that SSI is a suitable measure for time taken to reach a steady state.

Of the 312 PD participants in the Longitudinal Behaviour Subset, 24 self-reported as not-medicated, and the remaining 288 as medicated. Six of the non-medicated participants showed transient behaviour and the remaining 18 were stable. No difference is found between the proportion of medicated participants that showed transient behaviour compared to non-medicated participants [Z = 0.28, p = 0.77].

Severity Correlation

Prior to splitting the Performance-Severity Subset (PSS) based on severities, from the first row of Table 6.4 it is seen that in all cases the steady state performance has a stronger correlation with severity than baseline performance.

		Total MDS-UPDRS		Section II MDS-UPDRS		MDS-UPDRS 2.4	
		BV	SSP	BV	SSP	BV	SSP
	All Severities	-0.15	-0.22	-0.16	-0.21	-0.17	-0.24
MDS-UPDRS	Severe	**-0.36**	**-0.54**	**-0.32**	**-0.41**	-0.27	**-0.47**
	Not Severe	-0.17	-0.19	-0.18	-0.19	-0.16	-0.20
Tapping	Severe	**-0.37**	**-0.49**	**-0.36**	**-0.40**	**-0.34**	**-0.43**
	Not Severe	-0.04	-0.11	-0.06	-0.09	-0.09	-0.17

Table 6.4: Spearman correlation coefficients between Baseline Values (BV) and Steady State Performances (SSP) with severity scores for the tapping activity with significant correlations (at the 0.05 level) shown in bold. These results correspond to the methods outlined in §6.3.4 when applied to the tapping activity.

When splitting the PSS based on MDS-UPDRS severity, it is again observed that in all cases the steady state performance has a stronger correlation with severity when compared to the baseline performance. Similarly, when splitting the PSS based on tapping performance, it is found that the steady state performance has a stronger correlation with severity than baseline performance.

In both splitting approaches, it is found that the performance of severe participants systematically have stronger correlations with severity scores when compared to non-severe participants. Further, it is found that splitting the participants in such a way produces statistically significant correlations of increased and moderate strengths.

In addition to correlating tapping performance with the total MDS-UPDRS and Section II MDS-UPDRS (motor experiences of daily life) scores, it is found that MDS-UPDRS 2.4 ('Over the past week have you usually had troubles handling your food and easting utensils') also correlated significantly with performance in both splitting techniques.

Figure 6.4 demonstrates the relationship between years since diagnosis and total MDS-UPDRS for all PD participants. The correlations between years since diagnosis and severity scores for the tapping PSS are given in Table 6.5. Further, weak correlations were found to exist between years since diagnosis and age

		MDS-UPDRS			Performance	
		Total	Section II	MDS-UPDRS 2.4	Baseline	Steady State
	All Diagnoses	**0.34**	**0.36**	0.23	0.02	0.05
Diagnosis	Longstanding	**0.35**	**0.41**	**0.38**	0.02	0.06
	Recent	0.25	0.23	0.11	0.02	0.04

Table 6.5: Spearman correlation coefficients between participants' years since diagnosis and their severity scores and tapping performances with significant correlations (at the 0.05 level) shown in bold. These results correspond to the methods outlined in §6.3.4 when applied to the tapping activity.

in the Longstanding Diagnosis group (R = 0.01) and in the recent diagnosis group (R = 0.03).

The MDS-UPDRS floor effects were large within each of the 16 individual questions. On average, 47.1% of responses to each question were the lowest possible score (0 – 'Normal'). Conversely, the ceiling effects within each question were small, with an average 'Severe' response rate of 1.0% with six of the questions not containing any 'Severe' subjects. The MDS-UPDRS 2.4 question consisted of 54.2% 'Normal' responses whereas no instances of the maximum score of four ('Severe') were reported. These effects are not reflected in the total MDS-UPDRS score (summation of each question) with no subjects reporting the lowest or highest scores possible (zero and 64 respectively).

Memory

Short-Term Behaviour

At baseline performance, the 91 PD participants showed no significant memory impairment (as quantified using the total memory score) compared to the 11 HC and the 10 YHC participants. However, after five repetitions PD participants differed from the YHC group significantly [p = 0.04].

Performing paired t-tests between the baseline performance and performance at the fifth repetition for the YHC, HC, and PD groups yields p-values of 0.29, 0.15, and 0.04 respectively. This finding is consistent with the equivalent result in the tapping activity and demonstrates that the only participant group showing a significant level of change at the 0.05 level are the PD participants.

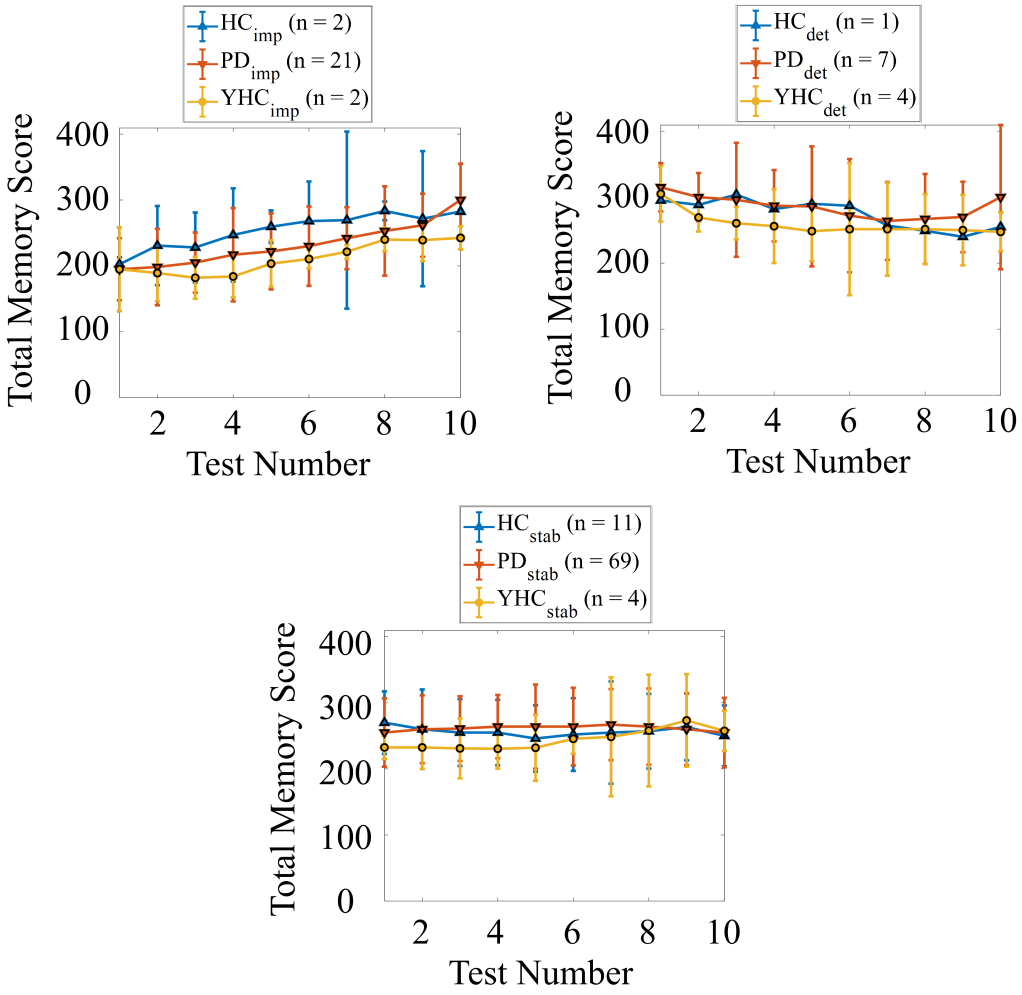


Figure 6.5: The longitudinal behaviour over the first 10 test instances in the memory activity for participants identified as (a) improving (b) deteriorating and (c) stable. These results correspond to the methods outlined in §6.3.3 when applied to the memory activity.

For all participants it is again found that the change in performance over the first five test instances correlates with baseline performance [R = -0.49], showing that participants with a lower baseline memory performance are more likely to improve than participants with a higher baseline performance.

Longitudinal Behaviour

In the Longitudinal Behaviour Subset data, no significant memory impairment is found between PD, HC, or YHC participants in baseline performance. The small sample size available is noted when drawing conclusions from these tests.

Over 10 test instances, the improving YHC, HC, and PD participants increased

	% of Sub-Group	Baseline Value ($\mu \pm \sigma$)	Final Value ($\mu \pm \sigma$)	Progress Ratio ($\mu \pm \sigma$)	Steady State Index ($\mu \pm \sigma$)
			YHC (N = 10)		
Improving	20.0	195.0 ± 63.6	225.0 ± 63.6	1.26 ± 0.12	5.5 ± 3.5
Deteriorating	40.0	287.5 ± 42.1	223.7 ± 19.7	0.86 ± 0.03	5.3 ± 2.5
Stable	40.0	233.8 ± 17.5	235.0 ± 55.8	1.01 ± 0.06	1.0 ± 0
			HC (N = 14)		
Improving	14.3	202.5 ± 10.6	300.0 ± 56.6	1.16 ± 0.03	1.5 ± 0.7
Deteriorating	7.1	295.0 ± 0	240.0 ± 0	0.87 ± 0	7.0 ± 0
Stable	78.6	273.6 ± 47.2	254.6 ± 62.1	0.98 ± 0.04	1.0 ± 0
			PD (N = 97)		
Improving	21.7	200.9 ± 47.1	255.0 ± 65.5	1.24 ± 0.19	5.8 ± 6.2
Deteriorating	7.2	323.6 ± 36.6	231.0 ± 44.2	0.80 ± 0.10	11.6 ± 13.5
Stable	71.7	257.3 ± 51.4	255.9 ± 52.7	0.99 ± 0.05	1.8 ± 4.8

Table 6.6: Baseline and longitudinal characteristics in the memory activity of each of the three Learning Phenotypes in the three participant groups. These results correspond to the methods outlined in §6.3.3 when applied to the memory activity.

their memory score by an average of 47.5, 80.0, and 80.4 points respectively [Figure 6.5(a)] although none of the improving groups showed a significant change between baseline and 10^{th}instance performance.

The proportion of each sub-group showing transient behaviour can be seen in Table 6.6. Due to the insufficient number of participants in each of the YHC and HC Binomial proportion tests are inappropriate.

No difference is found between any of the group SSI values. Table 6.6 demonstrates that the deteriorating PD participants have the longest average SSI, although this L-PT is also restricted in participant numbers (N = 7).

Severity Correlation

The first row of Table 6.7 shows that the steady state performance tends to have a weaker correlation with severity than baseline performance. This result is also seen when splitting the participants based on MDS-UPDRS severity and based on performance impairment.

However, in both splitting approaches the performance of the severe/impaired participants have stronger correlations with severity scores than when compared to non-severe/non-impaired participants. Consistent with the equivalent results in the

		Total MDS-UPDRS		Section I MDS-UPDRS		MDS-UPDRS 1.1	
		BV	SSP	BV	SSP	BV	SSP
	All Severities	-0.14	-0.04	-0.16	-0.08	-0.01	-0.03
MDS-UPDRS	Severe	**-0.43**	-0.10	-0.18	-0.08	**-0.33**	-0.20
	Not Severe	-0.13	-0.01	-0.17	-0.14	-0.01	-0.06
Memory	Severe	**-0.33**	-0.08	**-0.51**	-0.16	**-0.36**	-0.01
	Not Severe	-0.03	-0.01	-0.06	-0.19	-0.13	-0.09

Table 6.7: Spearman correlation coefficients between Baseline Values (BV) and Steady State Performances (SSP) with severity scores for the memory activity. These results correspond to the methods outlined in §6.3.4 when applied to the memory activity.

tapping analysis, it is again found that the severe/impaired participants produce several statistically significant correlations with moderate strength.

In addition to correlating memory performance with the total MDS-UPDRS and Section I MDS-UPDRS (non-motor experiences of daily life) scores, the MDS-UPDRS 1.1 ('Over the past week have you had problems remembering things, following conversations, paying attention, thinking clearly, or finding your way around the house or in town?') is found to also correlate significantly with performance in both splitting techniques.

The correlations between years since diagnosis and severity scores for the memory PSS are given in Table 6.8. Again, weak correlations are found to exist between years since diagnosis and age in the Longstanding Diagnosis group (R = 0.07) and in the recent diagnosis group (R = 0.04).

The MDS-UPDRS floor effects were again large within each of the 16 individual questions. On average, 47.8% of responses to each question were the lowest possible score. Consistent with the tapping activity, the ceiling effects within each question were small, with an average 'Severe' response rate of 0.7%. In 13 of the questions, no subjects reported a 'Severe' score. The MDS-UPDRS 1.1 question consisted of 54.8% 'Normal' responses whereas no subjects reported the maximum score of four ('Severe'). These effects are not seen in the total MDS-UPDRS score (summation of each question) with no subjects reporting the lowest or highest scores possible.

		MDS-UPDRS			Performance	
		Total	Section II	MDS-UPDRS 2.4	Baseline	Steady State
	All Diagnoses	**0.35**	0.19	0.04	-0.11	-0.10
Diagnosis	Longstanding	0.10	0.06	0.04	-0.19	0.02
	Recent	0.22	0.19	0.02	-0.17	-0.05

Table 6.8: Spearman correlation coefficients between participants' years since diagnosis and their severity scores and memory performances with significant correlations (at the 0.05 level) shown in bold. These results correspond to the methods outlined in §6.3.4 when applied to the memory activity.

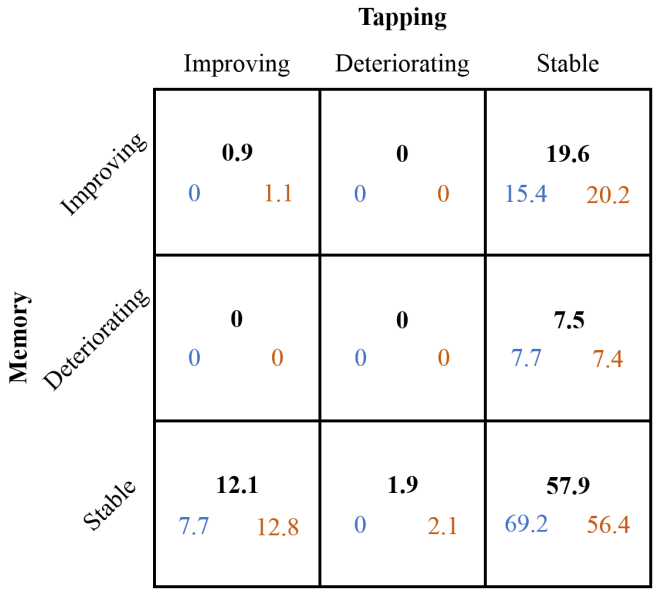


Figure 6.6: Confusion plot of inter-activity longitudinal behaviour. Numbers quoted are the percentage of the population that fall into each segment. The central black figures are the percentages of all 107 participants in the analysis. The lower left numbers in blue are the percentage of the 13 HC participants. The lower right numbers in red are the percentage of the 94 PD participants. These results correspond to the methods outlined in §6.3.5.

Relationship between Motor and non-Motor Tasks

Of the 107 participants who were in both the tapping and memory longitudinal analyses there were 13 HCs and 94 PDs. Figure 6.6 shows how these participants performed longitudinally in both the tapping and memory activities.

Disease Characteristics between Sexes

A total of 1,087 participants have a professional diagnosis of PD and 5,581 participants are healthy. Within the PD participants, a higher proportion of males than females is found, with males accounting for 65.8% of cases. Interestingly, this sex imbalance is present in the healthy participants also, with males accounting for 80.8% of the healthy cases. However, of the 5,314 male participants only 13.5% are PD cases. Of the 1,461 female participants, 25.3% are PD cases.

Although an imbalance exists between sexes (more males than females in cohorts of PD and health controls), the participants are well age matched with respect to the ages at which males report disease onset (56.6 ± 9.6 years) showing no significant difference to that of females (56.2 ± 9.0 years) [p = 0.53]. This finding is overall consistent with the age at which professional diagnosis occurs; males (58.4 ± 9.0 years) and females (57.9 ± 8.4 years) [p = 0.44].

In the tapping activity, no difference in starting performance between males (135.4 ± 61.3 taps) and females (133.9 ± 58.0 taps) [p = 0.73] is found. Similarly, in the memory activity, the starting performance of males (254.5 ± 57.1) shows no significant difference to that of females (256.9 ± 63.2) [p = 0.74].

In the tapping activity, the results of the Binomial proportion tests showed no statistical difference between the proportion of male (N = 179) to female (N = 133) participants in the improving [p = 0.61], deteriorating [p = 0.91], and stable [p = 0.42] phenotypes. Similarly, in the memory activity, the proportion of male (N = 58) and female (N = 39) participants in the improving [p = 0.75], deteriorating [p = 0.88], and stable [p = 0.38] phenotypes showed no significant differences.

6.5 Discussion

In a hospital environment the collection of high frequency longitudinal data from a large subject group would be costly and strategically difficult. Subsequently, research objectives are now being turned to exploiting remotely acquired datasets, such as in the mPower study, to validate clinical findings in large cohorts whilst

gaining new insights into the longitudinal characteristics of PD. In this chapter, it has been demonstrated that clinically validated features can be used to remotely identify task learning using smart-phones. From a large cohort of PD subjects and healthy controls, this chapter has revealed a small fraction of subjects show transient longitudinal behaviour which has previously gone undetected in smaller datasets. Furthermore, subjects with PD are more likely to show transient longitudinal behaviour whilst also tending to take longer to reach a steady performance in the alternating finger tapping test and the steady state performances are found to correlate better with disease severity than baseline performance. When investigating the floor-ceiling effect in the mPower MDS-UPDRS survey, significant correlations between performance and severity are found within the most severe participants, but not in the less severe subjects, whilst many of the individual MDS-UPDRS questions show a large floor effect that is not found in the total MDS-UPDRS score.

On the large mPower cohort, this chapter has validated the clinical finding that PD participants show an impaired motor performance and the ability to improve the AFT activity over a period of five repetitions. Participants who are in the same medication state at baseline and at their fifth repetition showed a significant performance change. This result demonstrates that the change in performance is not a consequence of medication response, as has previously been suggested, but rather due to practise and learning of the test [324]. Intuitively, it is found, across all subject groups and ages, that subjects with a lower baseline performance are more likely to improve than those with a higher baseline performance.

In the tapping activity, only a small fraction of all subjects showed transient behaviour and that the PD group tended to improve by a lesser degree over the first 20 instances than both the HC groups. This finding is unlikely to have been detected in previous studies due to the low subject numbers, lack of longitudinal data, and the small proportion of subjects showing this behaviour. The majority of subjects in all groups showed stable longitudinal behaviour, although the PD group was found to have a significantly larger proportion of subjects showing transient behaviour compared to both the healthy control groups. Consequently,

the PD participants took a significantly larger number of test instances before they reached their steady state performance. These longitudinal variations should be incorporated into the medication response models which have formed the majority of the work on the mPower dataset so far [237, 238].

Although no impairment is seen in the PD participants at baseline in the memory activity, the PD participants are the only group to show a significant change between baseline and the fifth repetition of the activity, consistent with the tapping activity. It is again observed that subjects with a lower starting performance tend to improve by a larger degree than subjects with a higher starting performance. The longitudinal memory behaviour analysis was more restricted in participant numbers than the tapping activity meaning proportional tests were inappropriate and were not performed. Unlike the majority of the non-motor learning in PD literature, the mPower study does not utilise the Serial Reaction Time Test, which assesses both long and short term memory through multiple tests showing a repeated sequence of stimuli. Subsequently, the mPower memory activity does not provide a means of assessing declarative learning, which has shown significant impairment in PD subjects [199].

In the tapping activity, the steady state performance is found to be a better predictor of severity than baseline performance. This suggests that starting performance at a new motor task is less representative of disease severity than the performance after the test has been learnt. However, this phenomenon is not seen in the memory activity. The number of subjects showing learning behaviour in the memory activity is smaller than in the tapping activity, thus the smaller amount of longitudinal change occurring in the performance combined with the generally weaker baseline correlations is reflected by the intuitive decrease in correlation strength with SSP.

When correlating years since diagnosis and total MDS-UPDRS for all PD participants, as shown in Figure 6.4, a significant correlation of moderate strength ($R = 0.32$, $p < 0.0001$) is observed; intuitively demonstrating that disease severity worsens with time. This finding is also seen when limiting the sample to the

participants in each activity in the PSS. However, although the Section II score (motor) of MDS-UPDRS shows a similarly strong correlation with years since diagnosis, weak correlations are found between the Section I score (non-motor) and years since diagnosis. This result may be interpreted in two ways. Firstly, this could imply that motor deterioration is more pronounced than non-motor deterioration, which has also been suggested in a cohort of 707 subjects over a two year period [325]. Secondly, accounting for the fact the MDS-UPDRS is self-assessed, it may be that participants are more aware of motor deterioration than non-motor. Using longitudinal smart-phone data and clinically assessed MDS-UPDRS collected from a larger cohort, future studies could make more informed inferences regarding the difference in rates of motor and non-motor degradation over time. Conversely, weak correlations were found between years since diagnosis and the baseline and steady state performance measures in both activities.

In the investigation into floor-ceiling effects, it was found that in both activities the performance of severe subjects consistently had a stronger correlation with severity scores than the non-severe subjects. Additionally, although very large floor effects are seen within each of the individual questions, neither floor nor ceiling effects are found in the total MDS-UPDRS score. These findings are a consequence of many subjects reporting low/mild severity for several questions and no severity for the majority of questions, thus, for many subjects, the total MDS-UPDRS severity is being diluted by the many questions being reported as zero. Accordingly, only subjects experiencing mild/severe symptoms in multiple aspects of the disease are grouped as Severe as they report many instances of mild/severe severities in the MDS-UPDRS survey. Hence, in the Non-Severe group, many subjects may report a severe impairment localised to one aspect of the disease, but no impairment in the remaining aspects, and therefore go undetected. This is reflected in the large divide in correlation strengths between feature performance and disease severities between the severe and non-severe groups.

There is an imbalance between sexes in the mPower dataset with roughly 80% of all participants being male. The imbalance between sexes in the mPower PD

population is found to be consistent with previous studies, demonstrating a higher incidence rate in males than in females. Although an imbalance exists between sexes, it is still appropriate to compare the performance and diagnosis metrics of the sexes as both samples contain a large number of participants even after their age matching. Fortunately, the Longitudinal Behaviour Subset is less severely affected by the sex imbalance. Binomial proportion tests demonstrate that the proportion of male and female participants in each L-PT is statistically similar, thus ruling out participant sex as a confounding factor in L-PT classification.

The 'big' data collected through the use of smart-phones is enabling multiple aspects of Parkinson's disease to be studied concurrently. Determining the complex relationship between motor and non-motor symptoms has thus far been restricted by clinical studies investigating a single activity such as gait, tapping, memory, or cognition. The multiple activity types being completed on smart-phone collection systems on a regular basis present the opportunity to overcome this restriction. When comparing the longitudinal behaviour of a small number of participants across the tapping and memory activities, a non-diagonal relationship between their learning-phenotypes is observed.

A recurring limitation with remotely collected big datasets is the lack of longitudinal subject compliance [78, 156]. In the tapping activity, 548 (6.8%) of the initial 8,003 subjects contributed at least 20 instances. Furthermore, compliance levels greatly differ between HC and PD participants. Of the 5,357 HC participants at baseline, only 236 (4.4%) contributed at least 20 instances. Conversely, 312 of the 1,060 PD participants (29.4%) at baseline contributed at least 20 instances. In the memory activity, it was necessary to lower the minimum number of instances required to be eligible for analysis to 10, resulting in 121 (12.5%) of the initial 968 subjects being eligible. The reduced number of subjects participating in the memory activity is a knock-on effect of this limitation as it was introduced over a month after the initial application release. Many of the participants who were active at the start of the study had stopped using the mPower application by the time the memory activity was released whilst the number of new participants

enrolling in the study had plateaued. However, although the compliance level is relatively low, the longitudinal data from these datasets is still significantly larger than in previous studies. The use of passive monitoring (continuous data collection when the smart-phone application is not actively being used) ensures a high level of compliance and is currently being implemented by several studies [79, 223]. However, this type of data is nearly entirely unlabelled making activity recognition and feature extraction more challenging.

This discussion is finalised by considering the future applications of the longitudinal analysis framework whilst speculating on techniques that may improve remote data collection. In this study, the focus was on a single feature from each activity, however, exploring the longitudinal behaviour of additional features may facilitate better prediction of other aspects of the subset of MDS-UPDRS survey. For example, in the tapping activity, monitoring features pertaining to fatigue or finger displacement could act as proxy measures for resting tremor, for which there is a specific question in the mPower MDS-UPDRS survey.

An additional progression of this study would be to utilize the additional gait and voice activities within the mPower dataset. The associated challenges in these data are the influence of differing environmental conditions, such as inclined walking surfaces and background noise, making it challenging to extract features that are robust at both the inter- and intra-participant level. During the description of the mPower dataset in Chapter 4, it was demonstrated in Figure 4.10 that the mPower gait signals are highly variable in quality, making it challenging to robustly extract a clinically interpretable feature on a longitudinal basis. Furthermore, any features extracted from the gait activity would only be beneficial to longitudinal analyses if they are interpretable and relatable to at least one of the limited number of clinical severity measures. Finally, it is common in clinical based studies to also recruit the PD participant's spouse; usually yielding an approximately age and sex balanced control participant. If volunteers of remote studies were to be encouraged to ask their spouse to also enrol, the resulting datasets may not only

be larger, but may also not contain the imbalances this research has highlighted are present in the current dataset.

6.6 Conclusion

Using the remotely collected mPower dataset, this chapter investigated the short-term and longitudinal behaviour of PD and HC participants in a motor test and a separate non-motor test. Significant impairments in the motor activity were found in the PD population on both a short and long term basis, whereas no such impairment existed in the non-motor test. In both tests, a previously unseen small fraction of participants were identified as showing transient longitudinal behaviour, with this transient behaviour being more common in PD participants than in HCs. By performing the analyses of this chapter, a deeper understanding has been gained into the characteristics of the dataset. Specifically, the levels of noise, demographic imbalances, inconsistent test completion, and poor participant retention rates were highlighted as factors not present in the OxQUIP dataset. In the next chapter, disease classification is performed using the mPower dataset. However, due to the characteristics of the mPower dataset identified in this chapter, more sophisticated dataset management and classification procedures are required in order to exploit the full richness of the mPower dataset.

7

Remote Disease Classification in the Presence of Missing Data

Related Publications:

1) J. Prince and M. De Vos, "A Deep Learning Framework for the Remote Detection of Parkinson's Disease Using Smart-phone Sensor Data" in Engineering in Medicine and Biology Society (EMBC), July 2018, Hawaii. IEEE

2) J. Prince, F. Andreotti, M. De Vos, "Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-wise Missing Data", IEEE Transactions on Biomedical Engineering, August 2018.

7.1 Introduction

Chapters 5 and 6 demonstrated the ability of digital sensors to perform in-clinic disease classification and remote disease monitoring using smart-phones respectively. The next step in the quantification of PD, as addressed in this Chapter, is the transition of the clinically performed classification tasks of Chapter 5 to the noisy and real world mPower dataset. Whereas the previous chapter focused on the longitudinal monitoring of a single biomarker from each of the tapping and memory activities, this chapter concentrates on how multiple feature sets from all four activity types can be used for disease classification.

As discussed in Chapter 4, the OxQUIP dataset is of exceptionally high quality in the sense that the data obtained is free of noise and contains very few missing instances. Conversely, when data is collected 'in the wild' it is often heavily corrupted by noise and missingness; as is the case in the mPower dataset. As such, more sophisticated dataset management and pre-processing techniques are required in order to fully delve into and explore the richness of the mPower dataset.

This chapter firstly presents a literature review focused on the use of smartphones for the classification of PD that extends from that presented in Chapter 2. This highlights the challenges experienced and the limitations of previous studies attempting to perform remote disease classification. Secondly, this chapter presents multi-source ensemble learning, a novel methodology which combines a dataset deconstruction technique with an ensemble learning strategy which subsequently enables participants with incomplete data to be included in the development of classification algorithms. During the development of this methodology, a series of convolutional neural networks are implemented which act as a further means of compensating for the noise in the dataset. Finally, capitalising on the large number of participants in the dataset, a series of feature selection algorithms are implemented so as to determine the efficacy of the presented model in comparison with two alternative methods that are widely used in the literature. This process includes the use of variable sample size bootstrap sampling to reveal the affect of sample size on the confidence of selected features.

The primary contributions of this chapter to the field of remote Parkinson's disease classification are:

- Remote disease classification is performed on the largest cohort of participants to date, including the use of data arising from multiple modalities/tests: tapping, gait, voice, and memory.
- Proposal of a novel method (Multi-Source Ensemble Learning) which facilitates the inclusion of all participants in the development of classification algorithms regardless of the presence of source-wise missing data.
- The use of a multi-branch convolutional neural network which is implemented across three separate modalities of raw data: tapping, gait, and voice
- The python scripts of all neural network models have been made freely available under a GNU GPL v3 license at `https://github.com/johnPrinceOx/Multi-Source_Ensemble_Learning_Neural_Networks`

- An exploration of feature selection strategies that focuses on the affect of sample size and the subsequent affect on classification accuracy.

7.2 Background

The largest limitation of smart-phone based datasets is the large quantity of missing data and poor volunteer retention rates. As participants are able to contribute data in multiple test types, it is commonly found that the vast majority of participants only complete a subset of tests [1, 232, 237]. Due to the heterogeneous nature of PD symptoms, the research community is particularly interested in determining the relationship between the different motor symptoms. When a dataset contains missing tests from many participants, it is commonly called a *source-wise* missing dataset, where the terms *source* and *test* are interchangeable. Source-wise missing datasets commonly occur and are especially prominent in datasets being collected in remote environments [79, 233].

Imputation techniques are commonly employed in the case of missing data [326]. However, in the case of source-wise missing medical data, imputation can be highly inappropriate. If a source is missing then the complete set of associated features is missing, therefore, a large number of features must be imputed from other sources which may contain no mutual information to the missing source, resulting in the imputations being poorly representative of their intended values [327]. An alternative approach in source-wise missing data is to discard all observations with incomplete data [233, 237]. The applicability of imputation techniques on multi-source missing data will be explored in the following chapter. Although discarding missing data guarantees a complete dataset, it is wasteful of potentially relevant data, especially if a large number of observations have incomplete data. In a recent study of remotely monitoring PD, 48.4% of the available participants were discarded from the study due to incomplete source data [232].

In this chapter, a method to compensate for source-wise missing data via combining a dataset deconstruction technique with ensemble learning is proposed. The method is shown to achieve a data retention rate of 100%, even if a majority

of observations have incomplete data. The method is applied to the large, remotely collected, and mostly incomplete mPower dataset for the purpose of PD classification. The classification ability of the method is compared to that of the current means of compensating for source-wise missing data and shows that the inclusion of additional data can improve classification and feature selection. The potential of learning deep convolutional neural network classifiers for this application is presented in tandem which is only possible because of the size of the dataset, which is an order of magnitude larger than the classical complete datasets.

7.3 Methods

7.3.1 Dataset Description

The data used in this chapter is taken entirely from the mPower study. Due to PD being most prominent in people over the age of 50 years old, only PD and HC participants 50 years or older are included in this analysis. Moreover, since the participant retention rate is inconsistent and low in the mPower study[1], this chapter focuses on the data provided 24 hours following the completion of their first source, resulting in 1,513 subjects. This is shown in Figure 7.1.

7.3.2 Dataset Deconstruction and Model Framework

Individual participants were able to contribute with any of the S sources (here $S = 4$: tapping, walking, voice, and memory) in any given permutation, thus the database comprises 2^S possible permutations of the available tests. Each source permutation may be represented as a binary vector, $\mathbf{I}[1...S]$, where $\mathbf{I}[i] = 1$ demonstrates that the i^{th} source has been contributed. In this research, a participant is assigned a binary source vector based on which sources were contributed. The binary source vector determines which *domain* a participant belongs to. A demonstration of the domain assignment process for a participant in the mPower dataset is given in Figure 7.1. Via assigning each participant to a domain, the initial source-wise missing dataset has subsequently been partitioned into 2^S smaller but complete domains; a process herein called *dataset deconstruction*.

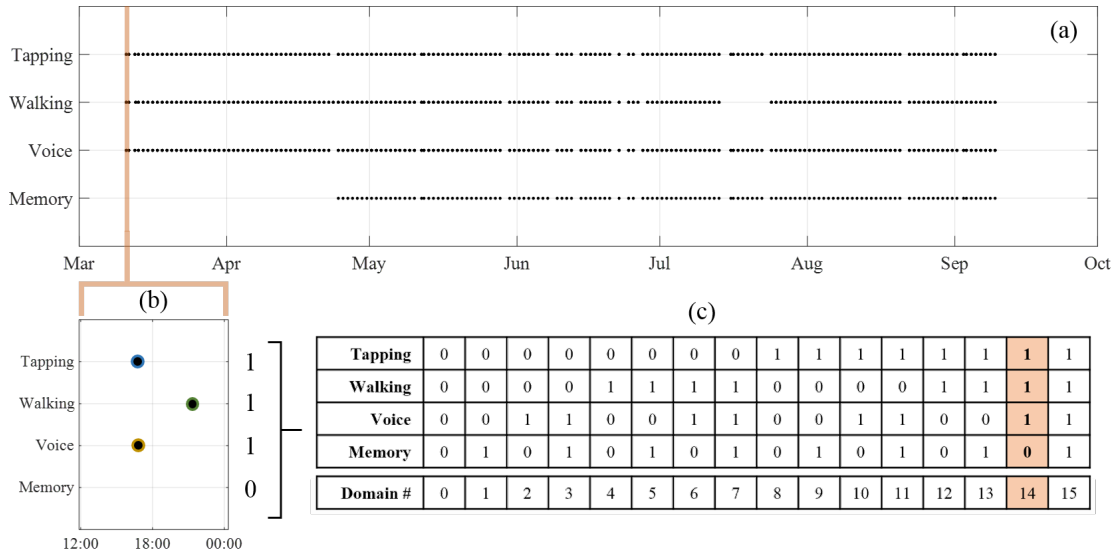


Figure 7.1: The domain assignment procedure. This participant contributed the largest number of test instances across all test types throughout the course of the study as shown in plot (a). In plot (b) the tests contributed within the first 24 hours are selected and their permutation is represented by a binary source vector. Subsequently, in plot (c), the participant assigned to the decimal integer domain which the binary source vector represents.

A degree of overlap exists with regards to which sources are present between certain domains. For example, in Figure 7.1 it is evident that domains 7, 13, and 15 contain the same sources (in addition to at least one other source) as present in domain 5. The ability for data representation to be shared across domains presents the opportunity to apply multi-task learning (M-TL) wherein multiple learning tasks are solved simultaneously [328]. In multi-task learning, machine learning algorithms are developed using the shared data representation between a target domain (such as domain 5) and appropriate transfer domains (such as domains 7, 13, and 15) simultaneously. Figure 7.2(a) provides a visualization of the target-transfer domain relationship of the current dataset. Indeed, M-TL has been applied using this dataset deconstruction technique but there are several limitations to its implementation [327]. The first limitation is that the number of participants in each domain is inconsistent which results in the number of participants used to train and test each of the M-TL models also being highly inconsistent; as is apparent in Figure 7.2(b). Thus, the results of a M-TL model when tested on each

domain could be being confounded by (i) the number of participants in the domain (ii) the characteristics of these participants, or (iii) the sources present each domain. These confounding factors make the M-TL results difficult to interpret.

To overcome these limitations, two special cases of the dataset deconstruction are noted. Firstly, the *single source domains* where only a single source is present; in the mPower dataset these are domains 1, 2, 4, and 8 which correspond to the memory, voice, walking, and tapping individual source domains respectively. For each single source domain, all participants who contributed the source are eligible to be used to develop the individual source model, regardless of which domain they have been assigned to. The example participant in Figure 7.1 would therefore be eligible to develop models for the tapping, walking, and voice single source domains. Subsequently, even participants with incomplete source data can be used in developing the individual source models thus ensuring a 100% participant retention rate. A participant with missing data who completed n sources ($0 < n \leq S$) is eligible to be used in developing n individual source models. Secondly, we note that the participants with complete source data (domain 15) are the only participants eligible to be used in developing *all* of the individual source models. Consequently, participants with complete source data will be assigned as a test group against which all individual source models will be tested. Furthermore, the individual source models can be fused in all permutations through the use of source ensembles. All models created through source ensembles can also be tested on the participants with complete source data.

The model framework can now be formally defined. Via deconstructing the source-wise missing dataset, it is possible to develop S individual source models using all participants, even those with missing data. The participants with complete source data are excluded from the training/validation of the individual source models and are reserved to act as a consistent test set against which all individual source models, and their permutations, may be tested. Having created a consistent test set for all models, the results of all models will be directly comparable to one another, thus removing the aforementioned confounding factors.

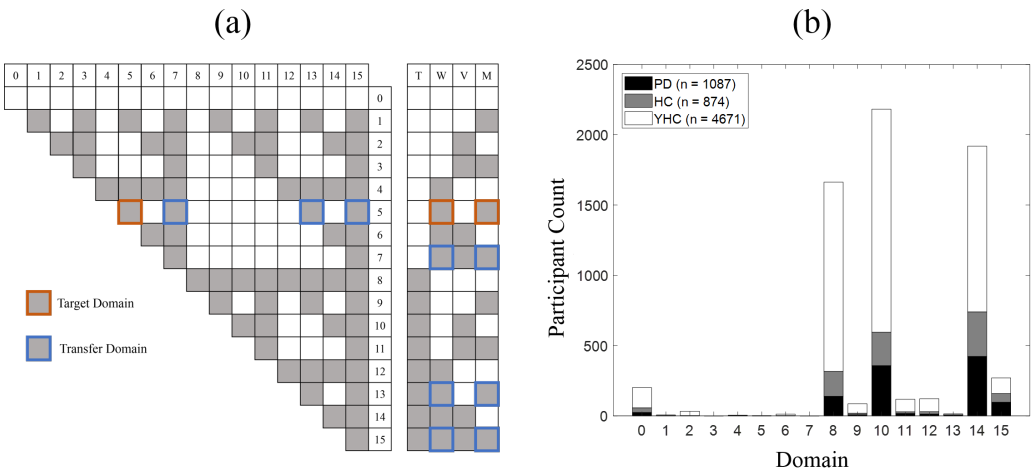


Figure 7.2: Domain Representations. (a) A visualisation of the target-transfer domain relationships for all domains and (b) a histogram showing the number of participants in each domain prior to the application of the target-transfer domain sharing. Note that YHC represent the health control participants under the age of 50 years old who are not included in this study.

Table 7.1: Demographic data of participants with incomplete data contributions who form the training and validation set for each source.

	Tapping	Walking	Voice	Memory
N	1323	624	1072	39
(PD/HC)	(799/524)	(377/247)	(679/393)	(25/14)
Age	62.4 ± 7.7	62.8 ± 7.6	62.9 ± 7.7	62.6 ± 8.7
% Male	68.6	71.2	67.8	61.5

In the mPower dataset, 1,380 participants were identified with incomplete source data who form the training and validation set. The instances and demographics of the training and validation participants are given in Table 7.1. There were 133 participants identified (87 PD/ 46 HC, age 62.9 ± 7.6, 71% Male) with complete source data who are assigned as the test set. Each participant in the test set contributed one instance in each of the sources. The effects of sex are assumed to be negligible on test performance as has been previously suggested on the mPower dataset [1, 238].

Table 7.2: The tapping feature set extracted from the touch-screen data. The features are separated into three types; Rhythm, Spatial Variability, and Fatigue.

Feature Name	Definition
	Rhythm
totalTaps	Total number of Taps
meanTimeTap	Mean time between successive taps
stdTimeTap	Standard deviation of time between successive taps
	Spatial Variability
xCoordsStd	Standard Deviation of x-pixel coordinates
xCoordsRange	Range of x-pixel coordinates
yCoordsStd	Standard Deviation of y-pixel coordinates
yCoordsRange	Range of y-pixel coordinates
pixelDistanceTotal	Summation of number of pixels between all taps
pixelDistanceMean	Mean number of pixels between successive taps
pixelDistanceStd	Standard deviation of number of pixels between successive taps
pixelDistanceIQR	Inter-quartile range of pixel distance between successive taps
	Fatigue
tapFatigueRatio	Ratio of number of taps in first 10 seconds and number of taps in final 10 seconds
tapFatigueHalf	The time at which half the total number of taps occurs

7.3.3 Individual Source Model Development

Feature Based Classifiers

Three classifiers were utilised that required an explicit feature set to be extracted in order to develop the individual source models. In this section, a detailed description of the feature set extracted from all sources is provided.

As discussed in chapter 4, the tapping activity provides raw data from two sensor types: the touch-screen and the tri-axial accelerometer. Features were extracted and concatenated from both sensors. The touch-screen features pertain to rhythm, spatial variability, and fatigue (Table 7.2) whilst a widely used set of signal based features are extracted from the accelerometer waveforms (Table 7.3) [79, 161]. The feature set is extracted from two second windows of the signal using a 0.25 second overlap between windows. The median and standard deviation of each feature across the windows is calculated and used to form the feature vector for the signal. This process is repeated for all signals.

From the walking activity the same set of signal based features were extracted from the tri-axial accelerometer and gyroscope. However, as the raw waveforms

Table 7.3: The signal based feature set extracted. The s_ax_ notation represents the sensor and axis of the signal from which the feature set is being extracted from respectively e.g. acc_x_mean represents the mean of the x-axis accelerometer signal.

Feature Name	Definition
s_ax_mean	Signal mean
s_ax_std	Signal standard deviation
s_ax_prc25	Signal 25th percentile
s_ax_prc75	Signal 75th percentile
s_ax_iqr	Signal inter-quartile range
s_ax_med	Signal median
s_ax_mod	Signal mode
s_ax_range	Signal range
s_ax_skew	Signal skewness
s_ax_kurt	Signal kurtosis
s_ax_mse	Signal mean squared energy
s_ax_entropy	Signal entropy
s_ax_dfc	Signal dominant frequency component
s_ax_fda	Signal dominant frequency amplitude
s_ax_loc_power	Locomation power density in range 0.5 - 3 Hz
s_ax_ar_coeff_1	Autoregression coefficient at time lag 1

are known to contain segments of noise or periods of no movement, features were only extracted from sections of the signal that were identified as gait as determined by a gait-segmentation algorithm as demonstrated in [215]. The specific gait segmentation algorithm used a magnitude based threshold on the accelerometer signals. The total acceleration magnitude, $|\boldsymbol{a}|$, was obtained via:

$$|\boldsymbol{a}| = \sqrt{\boldsymbol{x}^2 + \boldsymbol{y}^2 + \boldsymbol{z}^2} \tag{7.1}$$

where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, are the acceleration waveforms in the x, y, and z-axes respectively.

The acceleration magnitude is smoothed using a moving average filter with a window size of 0.3 seconds. The acceleration magnitude is normalised such that $|\boldsymbol{a}| \in [0, 1]$. The threshold was selected as the mean plus a single standard deviation of the acceleration magnitude. Features were then selected from the longest section of gait provided that the segment was five seconds or longer. If a segment of gait five seconds or longer did not exist, the participant was labelled as missing the

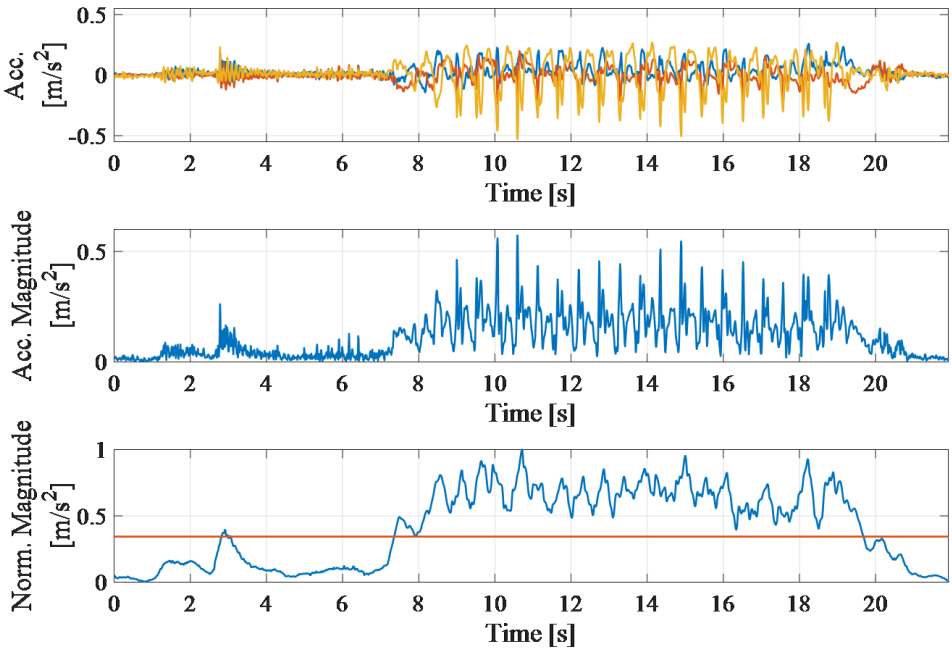


Figure 7.3: The gait segmentation procedure. (a) The raw tri-axial accelerometer signals (b) is the accelerometer magnitude as calculated through Equation 7.1 and (c) shows the acceleration magnitude having undergone the moving average filter. The red horizontal line is the threshold above which the signal is assigned as gait.

gait source. This process is demonstrated in Figure 7.3. The corresponding results of the gait segmentation algorithm are shown in Figure 7.4.

This gait segmentation algorithm was selected as it was specifically designed to be implemented on the raw waveforms collected from smart-phone embedded IMU's [215]. Furthermore, this algorithm has been tested and found the most effective (in terms of the smallest error rate when compared to nine other algorithms) using the waveforms collected at a variety of smart-phone sensor locations including the trouser front pocket.

For the voice activity, features were extracted from the raw audio recordings using the publicly available Matlab (Mathworks, USA) toolbox developed by [40]. These features utilize both temporal and frequency based metrics and have proven capable of detecting dysphonic PD participants in remote environments [39, 184]. The features extracted from each recording included measures of fundamental

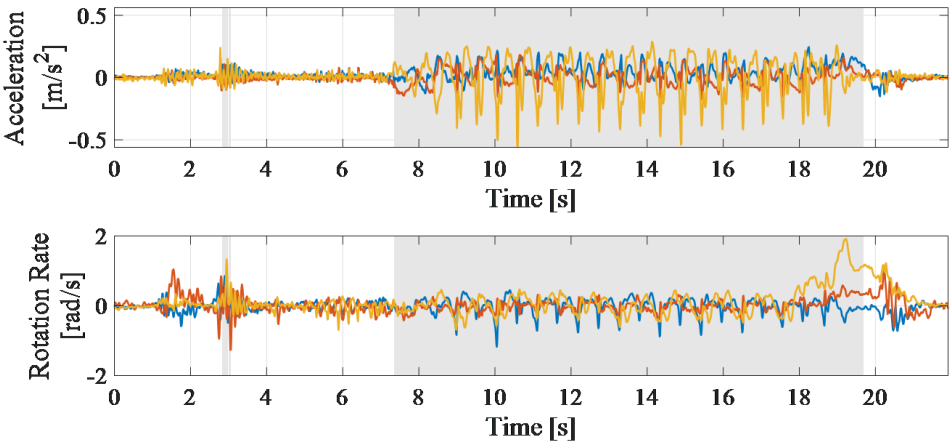


Figure 7.4: Having performed the gait segmentation algorithm, features are extracted from the acceleration and gyroscope signals identified as gait (shown in the grey box). Recall that the longest section of gait that is longer than 5 seconds is selected (∼7.5 - 19.9 seconds) and all other sections are discarded (∼2.9 - 3.1 seconds)

frequency and its corresponding amplitude and frequency variations (Jitter and Shimmer). Please refer to §2.3.3 for a review of speech based features and refer to [39, 40, 184] for an exhaustive list of all features extracted by this toolbox.

The memory activity provided only three features; total score, number of levels attempted, and number of incorrect responses.

The tapping, walking, voice, and memory activities provided 97, 180, 326, and 3 features respectively. Three feature based classifiers for each individual source were developed. Two of these classifiers are commonplace in the prediction of PD; logistic regression (LR) and random forests (RF) (§3.2).

The state-of-the-art Deep Neural Network (DNN) is also implemented on this feature set. All DNNs consisted of five layers with 200, 300, 50, 32, and 1 units respectively. All activation functions were rectified linear units (ReLU) due to their ability to improve training time whilst avoiding the vanishing gradient problem [276, 329].

During the training and validation of the LR models, least absolute shrinkage and selection operator (LASSO) feature selection is implemented on the training data.

Convolutional Neural Network Classifier

In addition to the three feature based base classifiers, a series of Convolutional Neural Networks (CNNs) were also implemented. CNNs are widely considered to be the state-of-the-art machine learning techniques and have received little attention in the field of PD classification to date due to the generally small size of PD datasets [2].

As mentioned in Section 3.3, CNNs do not require the definition of an explicit feature set but rather are capable of automatically learning features, or *filters*, directly from raw data. Furthermore, these filters are translationally invariant making CNNs particularly well suited to noisy raw data as in the mPower database. This is particularly evident when looking at the example of the raw walking activity data in Figure 7.4.

Of the four activities in the mPower database, the tapping, walking, and voice activities present data suitable to be used as inputs to a CNN (i.e. raw time-series) each requiring a small amount of pre-processing.

- In the tapping activity, sensor fusion is required so as to allow the unevenly sampled touch-screen data be used in conjunction with the evenly sampled accelerometer data [2]. Firstly, all tri-axial accelerometer signals were standardised to a length of 2,000 samples (20 seconds with an $fs = 100$ Hz). Zero padding was used to extend all signals whose length was less than 2,000. Secondly, as the touch-screen data is unevenly sampled, linear interpolation was used to create waveforms of equal length to the accelerometer waveforms. The touch-screen data for each recording was of length equal to the total number of taps (usually between 50 – 200 taps). Using the time-stamp of each tap, a new waveform with 2,000 samples was created for the x and y pixel coordinates. This is demonstrated in Figure 7.5. Consequently, the touch-screen pixel coordinate data is of equal length to the accelerometer data and is therefore suitable to be used as an input to a CNN.

- In the walking activity, the tri-axial accelerometer and tri-axial gyroscope signals ($fs = 100$ Hz) were used as the raw input. Unlike in the feature

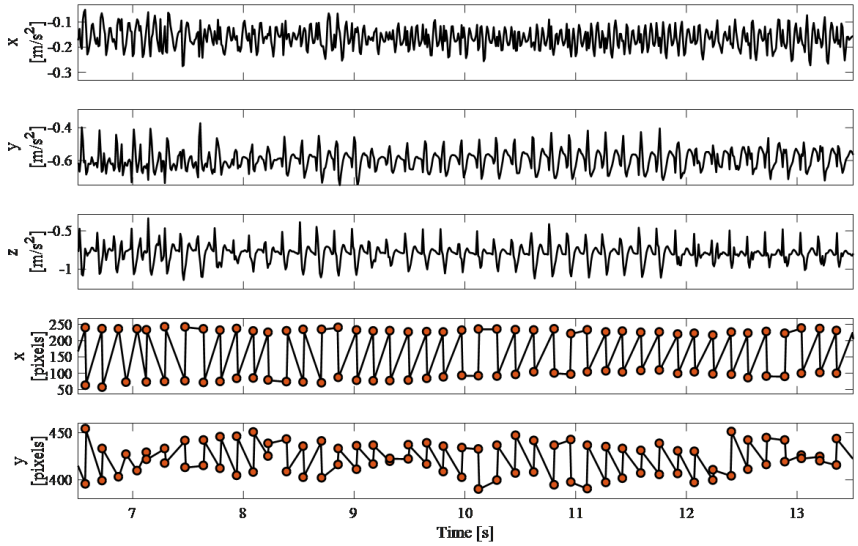


Figure 7.5: An example segment of the raw tapping data used as an input to the CNN. The upper three plots show the raw x, y, and z accelerometer signals respectively. The lower two plots show the linearly interpolated waveforms created from the x and y pixel coordinate data respectively.

based approach, very little pre-processing was required. All waveforms were standardised to a length of 3,500 samples; again using zero padding in the instances that were less than 35 seconds long.

- In the voice activity the raw voice recording signal was used ($fs = 42$ kHz). All waveforms were standardised to 420,000 samples; again using zero padding in the instances that were less than 10 seconds long.

The architecture of the multi-channel CNN used on all activity types is shown in Figure 7.6. While additional benefit might be obtained by using different network structures for the different sources, a generic general framework was opted for across all source data. Here, the concept of a variable first receptive field width is exploited across two convolutional branches [330]. In signal processing, it is favourable to inspect the frequency content of a larger section of the signal as opposed to a very small section. This theory also holds true when choosing the length of the first convolutional filter in a CNN. When using convolutional filters of a large width, the

frequency components of the data are better captured. Conversely, using filters of a small width better capture temporal aspects of the signal. Thus, the width of the first convolutional filter in each of the two branches is different so as to capture both the temporal and the frequency components of the data.

Alternative CNN architectures use small receptive fields but require many more layers and convolutional operations in order to capture the frequency components of the data [331, 332]. These networks are consequently very deep and require a large amount of computation power and training time to be optimized. Via the adoption of multiple branches of variable first receptive field widths, significantly fewer layers are required and often achieve similar model performances.

The multi-branch CNN architecture utilises ℓ_2-norm regularization, max-pooling, and batch normalisation layers as extra means of preventing overfitting, parameter reduction, and reduce training time respectively.

All CNN and DNN networks were developed using Keras with a Tensorflow (Google Inc., California) back-end using *Adam* optimization and model loss was calculated through binary cross-entropy [286].

Individual Source Model Evaluation

The classifiers developed for each individual source were trained and validated using the individual source domains (Table 7.1). During the training and validation procedure, each classifier underwent repeated stratified 10-fold cross validation. Prior to being separated into folds, the data is balanced so as to have equal number of PD to HC participants (using minority-class balancing).

In the feature based classifiers the training and validation sets are normalized to zero mean-unit variance using the means and standard deviations of the training feature set so as to avoid data leakage. The accuracy of each classifier is reported for the training and validation set. As the test set is imbalanced, when the individual source models are applied to the test set we also report the F_1-score (harmonic average) in addition to the classification accuracy.

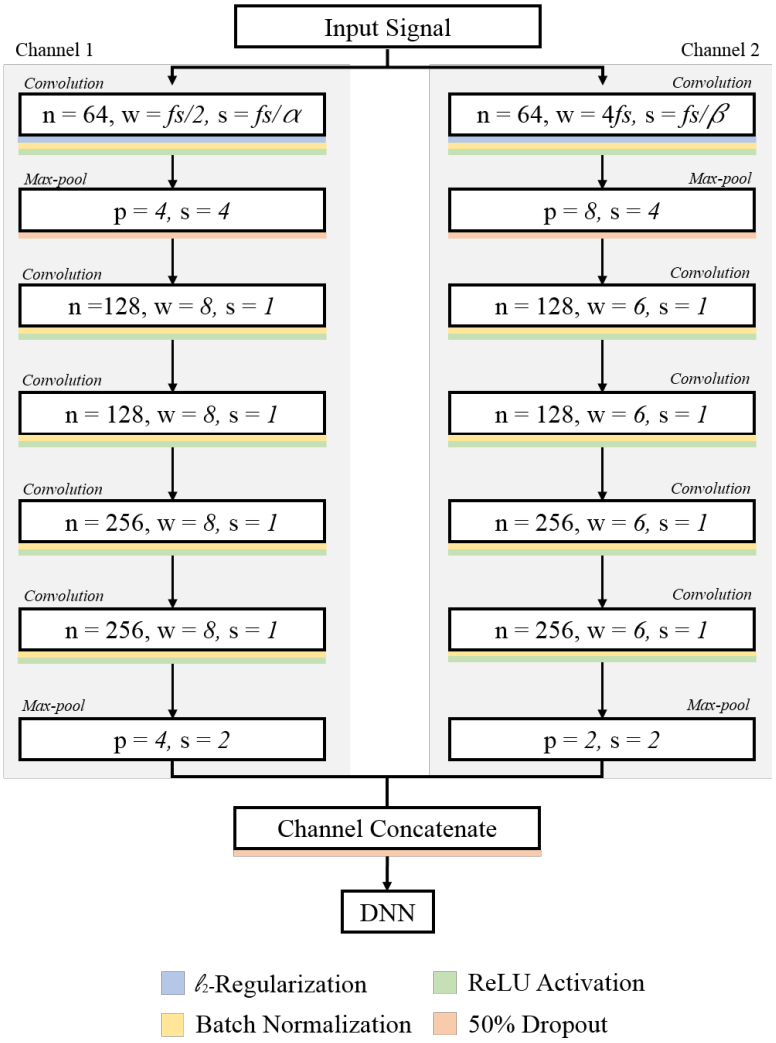


Figure 7.6: The multi-channel CNN architecture used on the voice, walking, and tapping source data. Here fs is the signal sampling frequency, n is the number of filters, w is the width of the filters, s is the stride length of convolutions, and p max-pool size. The size of the stride length in each branch in the first convolutional layer is controlled by α and β.

7.3.4 Ensemble Learning Approaches

Ensemble learning, often called classifier fusion, enables the predictions of multiple algorithms to be fused into a single prediction [288]. Often, the single ensemble prediction outperforms each of the 'base' algorithms due to the ensemble accounting for variability within the base algorithms' prediction ability [333].

Two forms of ensemble learning are presented (classifier and source) which enable two types of variability to be accounted for simultaneously.

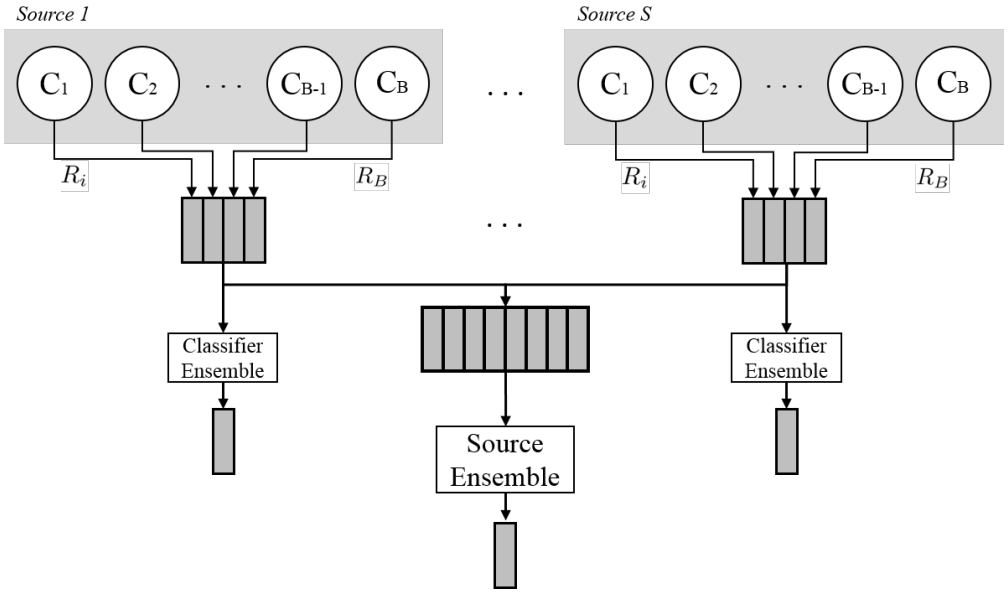


Figure 7.7: Schematic of the classifier and source ensemble procedures. There are B base classifiers within each source. R_i is the response from the i^{th} classifier. A classifier ensemble is implemented on the responses from all B classifiers from a single source. A source ensemble is performed on all responses from all classifiers from multiple sources.

Classifier Ensemble

Via performing a classifier ensemble within each individual source, whether the classification ability of an ensemble of multiple classifiers outperforms each individual base classifier is assessed.

Two popular classifier ensemble learning algorithms are implemented: majority voting and mean probability (§3.4)[334, 335]. In this implementation of classifier ensemble, we take the multiple classifiers *within each source* and ensemble their responses. Majority voting returns the binary response that occurs most frequently between the base classifiers. Alternatively, the mean probability ensemble is performed on the 'soft' response of each classifier and therefore returns a response based on the average probability of all classifiers for a single observation. A schematic of the classifier ensemble strategy is demonstrated in Figure 7.7.

The first classifier ensemble was formed using the three feature based classifiers (LR + RF + CNN). The second classifier ensemble was formed using both the feature based and the CNN classifiers (LR + RF + DNN + CNN). This enables

any difference between the CNN classifier to be directly compared against a purely feature based approach. The accuracy and F_1-score of all classifier ensembles is reported for each source individually.

Multi-Source Ensemble

In addition to classifier ensembles, an alternative usage of ensemble learning is presented in the form of source ensembles. Source ensembles serve two purposes in this research:

a The first purpose is to account for the source-wise missing data. As models have been developed for each source individually, source ensembles allow the fusion of these individual source models in all possible permutations; allowing all individual source models to be fused and implemented on the test participants.

b The second purpose is to account for the heterogeneity of symptom prevalence in PD. It is common for PD to manifest itself differently across the population. Thus, a participant may show mild to severe symptoms in one source but not in any others whereas another participant of equal PD severity may show mild to severe symptoms in a different source. Through source ensembles the classification predictions from multiple sources can be consolidated into a single prediction which accounts for symptoms in all sources; thus accounting for symptom heterogeneity.

Continuing the ensemble notation outlined in §3.4, when implementing a source ensemble on all S sources, and assuming that B base classifiers are trained for each source, the majority voting source ensemble learner is defined as:

$$\mathcal{F} = \begin{cases} 1 & \text{if } \sum_{i=1}^{B\times S} R_i \geq \frac{B\times S}{2} \\ 0 & \text{if } \sum_{i=1}^{B\times S} R_i < \frac{B\times S}{2} \end{cases} \tag{7.2}$$

and the mean probability source ensemble learner as:

$$\mathcal{F} = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^{B\times S} P(R_i=1)}{B\times S} \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

A schematic of the source ensemble strategy is demonstrated in Figure 7.7. It is important to note source ensembles are not performed on the response of the classifier ensemble. Rather, the response of each classifier from all sources are used. This approach ensures the maximum amount of classifier variability is present during the source ensemble.

Source ensembles are firstly performed using a single base classifier from each source; thus localising the results of source ensembles without the influence of classifier ensembles. This enables the comparison between the effects of classifier and source ensembles independently. Finally, source ensembles are performed using multiple base classifiers from all sources. As in the pure classifier ensemble, this is performed initially using only the feature based base classifiers across all sources, followed by using all feature based and CNN base classifiers across all sources.

7.3.5 Visualisation of the Multi-Source Ensemble Procedure

In order to further understand the benefit of performing the source ensemble, a novel means of visualizing the source ensemble process is provided [2].

Sammon Mapping

Using Sammon mapping, it is possible to reduce the high dimensional feature matrix into a latent space of a lower dimension. Sammon mapping [336] is a dimensionality reduction technique that aims to preserve the inherent geometric structure of data. As a point mapping algorithm, Sammon mapping reduces N feature vectors of p-dimensions into a new l-dimensional feature space where l is typically two or three dimensions and thus enables visualization. The mapping works via fitting all N observations in the latent space such that their interpoint distances are approximately equal to that of the original p-dimensional space. A trivial visualization is provided in Figure 7.8.

Mathematically this is achieved via determining and optimising (usually through gradient descent) a distance metric, E, between all feature vectors:

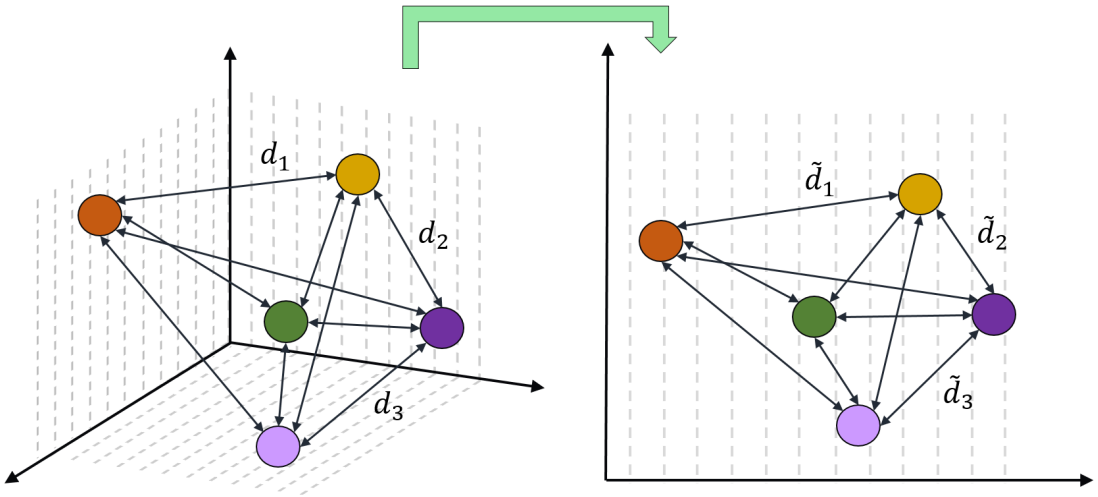


Figure 7.8: A trivial schematic of the Sammon mapping procedure. The original feature space ($p = 3$) is reduced into a 2-dimensional latent space. The Euclidean distances between each of the observations in the new latent space (e.g. $\tilde{d}_1$) have been optimised such as to preserve their distances in the original feature space (e.g. d_1).

$$E = \frac{1}{\sum_{i<j}[\tilde{d}_{ij}]} \sum_{i<j}^{N} \frac{[\tilde{d}_{ij} - d_{ij}]^2}{\tilde{d}_{ij}} \tag{7.4}$$

where d_{ij} is the distance between the i^{th} and j^{th} feature vectors in the original p-dimensional feature space and $\tilde{d}_{ij}$ is the distance between the i^{th} and j^{th} feature vectors in the new l-dimensional latent space.

As such, Sammon mapping is appropriate in the current application as it enables the reduction of the high dimensional feature space into a lower latent space allowing the visualization of participant location in the latent space. Furthermore, as it is possible visualise the location of all test participants in the latent space, it is therefore also possible to visualise which of these participants are correctly classified by each of the separate sources. This provides a visualisation of which participants are correctly classified by each of the separate sources, enabling the sources to be compared against each other on a participant-to-participant level.

Using the test participants, Sammon mapping is applied to the test participants' feature matrix such that it is reduced to a 2-dimensional latent space. Firstly, all test participants are plotted in the latent space according to their disease group

(PD vs HC). Secondly, the test participants correctly classified by each of the individual sources are plotted in the latent space. Finally, the test participants correctly classified by the multi-source ensemble are plotted and compared against each of the individual sources.

7.3.6 Model Comparison Approaches

As a means of model comparison, two alternative approaches are implemented for multi-source classification against which multi-source ensemble learning is compared. Both alternative approaches also present a platform for assessing the importance of features from all sources.

Complete Dataset Learning

This is the approach most commonly used in the case of incomplete multi-source datasets. Only the participants with complete source data are selected (i.e. the test set) and all participants with missing data are discarded. Classification models are developed using only the participants with complete source data. Also of interest is the study of which features across all sources are most pertinent in the classification procedure, as such, LASSO and Sparse-Group LASSO are employed for feature selection [337]. The latter of these approaches introduces regularization at the feature and source level. When implementing LASSO during Complete Dataset Learning, the design matrix $\mathbf{X}$ in Equation 3.12 is:

$$\mathbf{X}_c = [\mathbf{X}_c^T, \mathbf{X}_c^W, \mathbf{X}_c^V, \mathbf{X}_c^M] \tag{7.5}$$

where $\mathbf{X}_c^T, \mathbf{X}_c^W, \mathbf{X}_c^V$, and $\mathbf{X}_c^M$ are the tapping, walking, voice, and memory design matrices of the N_c participants who contributed all sources respectively and $\mathbf{X_c} \in \mathbb{R}^{p \times N_c}$ is the complete multi-source feature matrix containing p features.

The same complete design matrix ($\mathbf{X_c}$) is used during Sparse-Group LASSO, which is defined as:

$$\arg\min_{\beta} \frac{1}{2}||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2 \sum_{i=1}^{S} \sqrt{p_i}||\beta_i||_2 \tag{7.6}$$

where λ_1 and λ_2 determine the quantity of regularization at the feature and source levels respectively and p_i is the number of features in the i^{th} source.

The top 10 highest weighted features are reported from both the LASSO and SG-LASSO methods. As this section is focused on the influence of the feature selection process, only a LR model is implemented. Using the identified features from both methods, LR models are trained and validated and their classification accuracy reported.

Incomplete Dataset Learning

In this approach, a two-stage feature selection technique is implemented starting with the participants in each of the individual source domains. Firstly, within each of the individual source domains, the highest weighted features using LASSO are determined. Feature selection is occurring at the feature level within each individual source, thus enabling the use of the large number of participants with incomplete data. The features selected by LASSO from the i^{th} individual source are denoted as β^i. The non-zero weighted features are selected from each source are concatenated to give the feature weighting vector for all sources:

$$\beta_{ISL} = [\beta^T, \beta^W, \beta^V, \beta^M] \tag{7.7}$$

From β_{ISL}, the features with the highest weightings are reported. In the second step, all of the features in β_{ISL} are selected from $\mathbf{X}_c$ yielding:

$$\mathbf{X}_{ISL} = \mathbf{X}_c(\beta_{ISL}) \tag{7.8}$$

such that the features selected using the large number of participants with incomplete data have now been applied to the smaller number of participants with complete data. The second feature selection process is concluded via applying SG-LASSO on $\mathbf{X}_{ISL}$.

The highest weighted features from Complete Dataset Learning can be compared against those selected in the Incomplete Dataset Learning approach and allows conclusions to be drawn whether feature selection and model development is better performed at the feature or source level.

As in the Complete Dataset Learning approach, the top 10 highest weighted features from both the LASSO and SG-LASSO methods are reported alongside the classification accuracy of the subsequent LR models.

7.3.7 Effects of Sample Size on Feature Confidence

To further assess the influence of performing feature selection using participants with incomplete data, the effect sample size has on feature distributions is investigated. Here it is examined whether using a large sample size, as in the Incomplete Dataset Learning technique, provides a more robust platform to perform feature selection than in the traditional approaches, such as the Complete Dataset Learning technique.

This is achieved via utilising bootstrap sampling on the training and validation participants, using variable sample sizes. Bootstrap sampling is a non-parametric statistical technique that enables statistical measures to be estimated from a randomly sampled subset of the data [254]. To perform the bootstrap, $nSamp$ participants are randomly sampled with replacement from the N_i participants with incomplete source data. This subset of participants is known as the bootstrap sample. The feature set contributed by the bootstrap sample, f_{bs}, is selected and statistical measures from each feature is calculated. Statistical measures calculated from the bootstrap sample, $\mathbb{E}(f_{bs})$, are formally referred to bootstrap statistics. The process of selecting a bootstrap sample and calculating the corresponding bootstrap statistics is repeated B times. This process is summarized in Algorithm 1. Consequently, for each bootstrap statistic, B estimates of the true statistical value have been compiled.

It follows that if the bootstrap sample size is small, or the original population data contains a large degree of noise, the resulting bootstrap statistics will not be representative of the true value of the statistical measure. It is for this reason that performing bootstrapping using many sample sizes is beneficial in this study of feature selection. At each sample size it can be determined whether the estimated feature values from a bootstrap sample are representative of the true feature values of the entire population. This allows an assessment of sample size and whether it is appropriate to undergo feature selection.

Table 7.4: Classification Performances of all the Individual Source Model classifiers During Training and Validation. These results correspond to the methods outlined in §7.3.3.

	LR	RF	DNN	CNN
Memory	66.3 ± 28.3	53.9 ± 36.5	65.8 ± 25.8	n/a
Voice	62.9 ± 5.2	61.8 ± 5.6	69.7 ± 8.4	**72.5 ± 4.1**
Walking	59.5 ± 6.5	55.3 ± 6.2	68.2 ± 8.6	**72.6 ± 2.8**
Tapping	60.5 ± 4.8	61.8 ± 5.1	65.3 ± 3.1	**69.4 ± 3.5**

Presented is a visualization of how feature distributions vary with sample size via calculating the mean and standard deviation of the bootstrap statistics using $B = 10,000$ bootstrap samples at each sample size.

Algorithm 1 Procedure for bootstrap feature sampling using a variable sample size.

initalize: $nSamp$
while $(nSamp \leq N_i)$ **do**
  for $1 \rightarrow B$ **do**
    Bootstrap sample $nSamp$ participants
    Select the bootstrap sample's feature values, f_{bs}
    Calculate and save $\mathbb{E}(f_{bs})$
  end for
  increment: $nSamp$
end while

7.4 Results

Individual Source Model Performances

The classification accuracy for each classifier in each of the individual source models are presented in Table 7.4. For the balanced validation set, the mean and standard deviation results are shown for the repeated 10-fold stratified cross validation. The classification accuracy and corresponding F_1 scores of these models when applied to the test set are provided in Table 7.5.

Table 7.5: Classification Performances of all the Individual Source Model classifiers on the Test Participants. These results correspond to the methods outlined in §7.3.3.

	LR	RF	DNN	CNN
	Accuracy (%)			
Memory	64.7	69.2	57.9	n/a
Voice	69.2	69.2	69.9	**75.2**
Walking	63.2	68.4	70.7	**72.9**
Tapping	60.2	67.7	61.7	**70.7**
	F_1 (%)			
Memory	70.8	78.8	63.6	n/a
Voice	78.8	75.5	80.6	**83.2**
Walking	69.2	74.7	77.5	**77.5**
Tapping	65.4	72.6	72.7	**78.7**

Ensemble Model Performances

Classifier Ensemble

The results of performing the classifier ensemble on the test set within each individual source are presented in Table 7.6. Results are divided into two types (i) feature based classifiers only and (ii) feature and CNN based classifiers together. For each, the accuracy and F_1 for both ensemble algorithms are provided.

Multi-Source Ensemble

In the top portion of Table 7.7, the results of performing a source ensemble (across all four sources) on the test set for each classifier separately are provided. In the bottom portion of Table 7.7 the accuracy and F_1-score resulting from models which combine both classifier and source ensembles are provided.

Multi-Source Ensemble Visualisation

The top row of Figure 7.9 shows the visualisation of all of the test participants in the reduced 2-dimensional latent space; separated by their disease grouping. Rows 2, 3, 4, and 5 show the location of the correctly classified participants by the memory, voice, gait, and tapping activities respectively. The bottom row shows

Table 7.6: Classification performances of the classifier ensemble in each individual source. These results correspond to the methods outlined in §7.3.4.

	LR + RF + DNN		LR + RF + DNN + CNN	
	Majority Voting	Mean Probability	Majority Voting	Mean Probability
	Accuracy (%)			
Memory	69.2	63.2	n/a	n/a
Voice	68.4	71.4	72.9	78.2
Walking	69.9	69.2	77.4	74.4
Tapping	66.2	63.9	70.7	65.4
	F_1 (%)			
Memory	79.0	70.3	n/a	n/a
Voice	80.6	80.6	82.9	85.0
Walking	79.4	75.2	83.0	80.0
Tapping	76.9	71.1	78.9	71.9

Table 7.7: Classification performances of the source ensembles for each individual classifier and for the combined classifier and source ensemble. All units are %. These results correspond to the methods outlined in §7.3.4.

	Majority Voting		Mean Probability	
	Acc.	F_1	Acc.	F_1
LR	75.9	83.2	69.2	78.8
RF	74.4	82.7	75.2	82.0
DNN	69.2	80.0	70.7	76.9
CNN	71.4	82.1	79.7	85.9
LR + RF + DNN	78.9	85.3	76.7	82.9
LR + RF + DNN + CNN	82.0	87.1	82.0	87.1

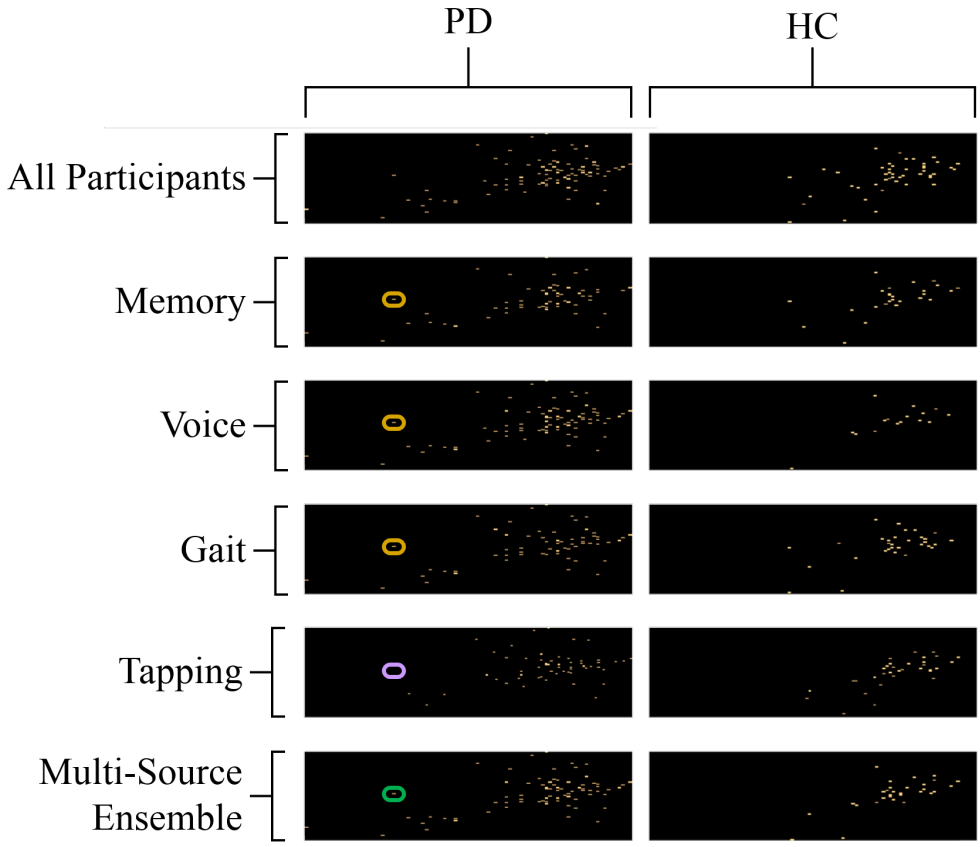


Figure 7.9: The visualisation of participants in the latent space according to disease grouping, individual source classification, and multi-source ensemble classification. These results correspond to the methods outlined in §7.3.5.

the location of the correctly classified participants by performing the multi-source ensemble learning technique.

To provide a single visual example of the multi-source ensemble technique, the yellow circles enclose a correctly classified participant by the memory, voice, and gait activities. However, the purple circle, in the same location in the latent space, demonstrates that the tapping activity did not correctly classify this same participant. However, as three of the four activities correctly classified this participant, it can be seen that this participant is correctly classified by the multi-source ensemble approach (in this example a mean probability ensemble was employed during the source-ensemble approach).

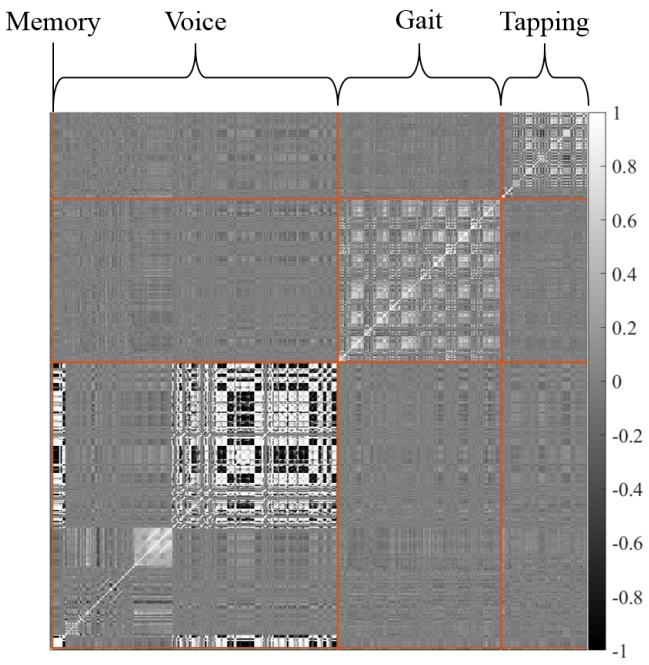


Figure 7.10: A 606 x 606 correlation matrix showing the inter- and intra-source feature Spearman's Rho correlations. These results correspond to the methods outlined in §7.3.6.

Model and Feature Selection Comparisons

Table 7.8 provides the classification accuracy of the Complete Dataset Learning and Incomplete Dataset Learning techniques. Recall that in Complete Dataset Learning, only the participants with complete source data have been used for selecting features to be utilised during model development. There are $N_c = 133$ participants with complete source data and a total of $p = 606$ features across the four sources thus $\mathbf{X_c} \in \mathbb{R}^{606\times133}$. Figure 7.10 demonstrates the inter- and intra-source feature correlations of $\mathbf{X}_c$. Conversely, recall that Incomplete Dataset Learning entails learning features from participants with missing source data and applying these features to the participants with complete source data. During Incomplete Dataset Learning, the dimensions of the design matrices undergoing LASSO in Equation 3.12 for the tapping, walking, voice, and memory sources are $\mathbf{X^T} \in \mathbb{R}^{97\times1323}$, $\mathbf{X^W} \in \mathbb{R}^{180\times624}$, $\mathbf{X^V} \in \mathbb{R}^{326\times1072}$, and $\mathbf{X^M} \in \mathbb{R}^{3\times39}$ respectively.

Table 7.9 provides the 10 features with the highest weights as determined via LASSO and SG-LASSO during the Complete Dataset Learning and the Incomplete

Table 7.8: The comparison of the classification accuracy (%) of the three approaches. These results correspond to the methods outlined in §7.3.6.

	LASSO	SG-LASSO
Complete Dataset Learning	71.4 ± 1.2	73.1 ± 3.3
Incomplete Dataset Learning	75.6 ± 1.3	73.2 ± 1.1
	LR + RF + DNN	LR + RF + DNN + CNN
Multi-Source Ensemble Learning	78.9	82.0

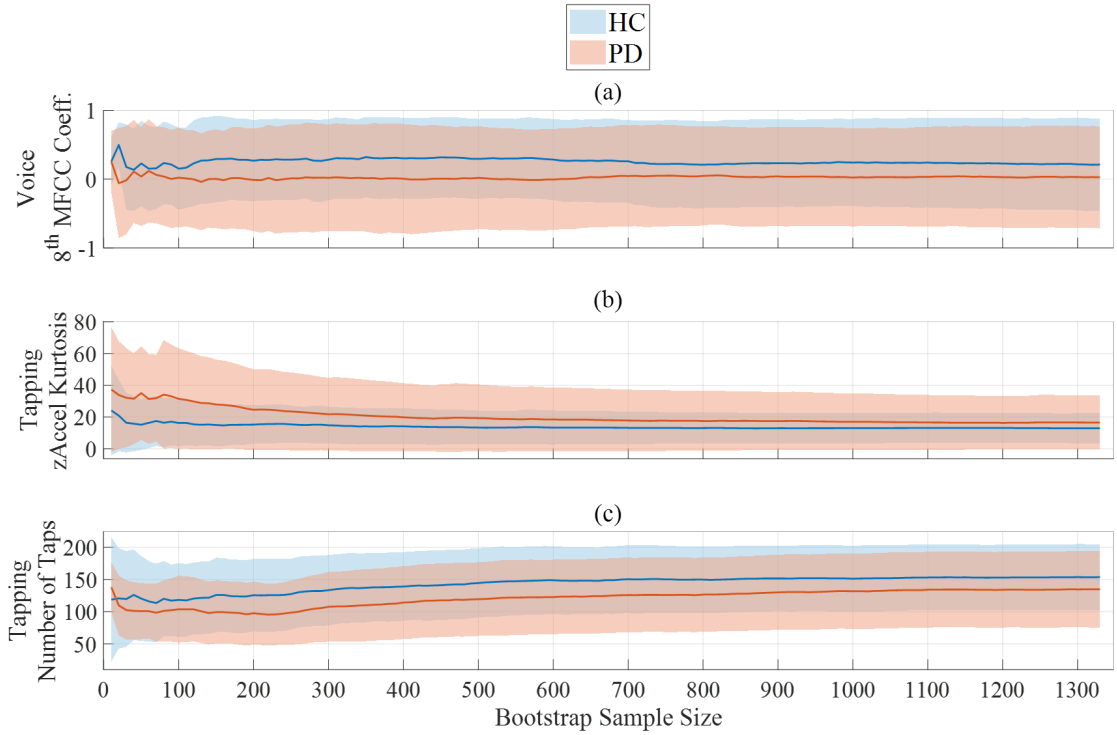


Figure 7.11: Demonstration of how three feature distributions vary with sample size. There are 10,000 bootstrap samples taken at each sample size. The solid line shows the mean of the mean of each bootstrap sample. The shaded area show the mean standard deviation of each bootstrap sample. These results correspond to the methods outlined in §7.3.7.

Dataset Learning techniques. Appendix A.2 provides an exhaustive description of each feature in Table 7.9.

Finally, Figure 7.11 shows how three features from the voice, walking, and tapping activities respectively vary with sample size.

Table 7.9: The top 10 features selected by the different approaches and different feature selection algorithms. These results correspond to the feature selection techniques outlined in §7.3.6.

Feature Rank	Complete Dataset Learning				Incomplete Dataset Learning			
	LASSO		SG-LASSO		LASSO		SG-LASSO	
	Source	Feature	Source	Feature	Source	Feature	Source	Feature
1	V	std-MFCC-10th coef	W	xGyro-DFA	V	std-8th delta	W	xGyro-Mean
2	W	yGyro-Skewness	V	Shimmer-F0prc75	V	std-MFCC-10th coef	W	xGyro-DFA
3	W	xGyro-Mean	W	zAccel-Median	T	zAccel-Kurtosis	T	zAccel-Median
4	T	xAccel-STD	V	mean-MFCC-8th coef	W	xGyro-Mean	V	mean-11th delta
5	V	std-8th delta	T	zAccel-Skewness	W	xGyro-DFA	W	yAccel-DFA-STD
6	V	GNE-NSR-TKEO	W	yAccel-Kurtosis-STD	V	HNR-mean	V	mean-MFCC-8th coef
7	W	xGyro-DFA	T	zAccel-AR-Lag-1	V	GNE-SNR-TKEO	W	zGyro-Median-STD
8	W	xAccel-Skewness	V	mean-delta delta 0th	V	Shimmer-F0-prc5	W	zAccel-Kurtosis
9	V	Shimmer-F0mean	T	IQR of Tap Distance	W	yGyro-DFA	V	mean-MFCC-5th coef
10	V	Shimmer-F0prc75	V	mean-12th delta-delta	V	Shimmer-F0-FM	W	zGyro-AR-Lag-1-STD

7.5 Discussion

Common techniques for analysing datasets with large quantities of missing data often result in a significantly smaller subset of the data being analysed. This chapter has presented a novel method for compensating for source-wise missing data through the combined use of dataset deconstruction and ensemble learning. This approach ensures a 100% participant retention rate without the need to perform imputation. Unlike previous work, multi-source ensemble learning identifies a consistent set of participants against which all models can be tested making the results highly interpretable. Due to the inclusion of a high number of participants and the robust fusion of multiple classification models, this method yields higher disease classification accuracies when used for remote detection of PD and to also be more appropriate at feature selection than traditional methods.

In total, 12 feature based classifiers and three CNN classifiers were developed. Four of the feature based classifiers utilised a state-of-the-art DNN. Interestingly, it is found that the DNN often gives consistent accuracies with the traditional feature based classifiers. This is indicative that the feature set is the limiting factor during classification as opposed to the classification technique. In the three sources where CNNs were developed, the CNN consistently outperformed the traditional feature based classifiers on both the training and test participants. The improvement seen by using CNNs is attributed to the translationally invariant nature of the filters identified from a large number of training instances. As the training sample size is large for all CNNs, it is highly likely that the resulting filters are robust at a population level and less likely to cause overfitting than in the traditional feature based approaches. This is the first study to the authors knowledge to implement a consistent CNN architecture on multiple types of smart-phone sensor data for the purpose of PD classification. These findings are consistent with the hypothesis that more sophisticated machine learning models used in tandem with a large cohort will improve remote PD classification [238].

The purpose of applying a classifier ensemble was to account for variability between classifiers thus providing additional means for accounting for the noise and

uncertainty that is inherent in the mPower dataset. The initial classifier ensemble made use of only the feature based classifiers. Although the ensemble often causes a small increase in performance within each separate source, the feature based classifier ensemble often fell short of the classification accuracy achieved by the respective source's CNN for all ensemble techniques. The second classifier ensemble made use of the feature based classifiers and the CNN classifier and in many cases reported accuracies higher than all base classifiers, including the respective source's CNN. Once more this is suggestive that the CNN is correctly classifying different subjects to the feature based classifiers on account of it using the convolutional based filters for classification as opposed to the hand-crafted feature set.

The benefit of the source ensemble approach is far more apparent than that of the classifier ensemble. When using a single classifier from multiple sources, the source ensemble outperforms the majority of single source classifiers. The cause of the improvement seen in the source ensemble approach is two-fold: (i) it accounts for the noise and variation within each of individual source models and (ii) it accounts for participants showing symptom heterogeneity wherein symptoms are only present in some sources.

The final ensemble approach demonstrated that the combined affect of source and classifier ensembles outperform all other ensemble approaches. It is again shown that ensembles at the source and classifier level show the highest classification accuracy when using both feature and CNN based classifiers. This finding is intuitive given that the classifier and source ensemble, when using the feature and CNN based classifiers, are accounting for the multiple types of noise and variability.

To test the efficacy of multi-source ensemble learning, two types of comparative studies were performed. Firstly, Complete Dataset Learning was performed which is the most common approach for datasets where source-wise missing data occurs. Here, models that are trained and validated on the participants with complete data showed an increase in classification accuracy when compared to single source models. It is important to note that of the 1,513 participants used in this study, only 133 (8.8%) had complete source data. As such, when using Complete Dataset

Learning, 91.2% of participants are being discarded. This is the standard approach used in the literature and is clearly a highly inefficient use of data [232, 237]. The second comparative study entailed an inspection of the influence of various feature selection techniques. During Incomplete Dataset Learning feature selection was performed within each source individually using the large number of participants with incomplete data and these features were applied to the participants with complete data for classification. It was shown that this feature selection approach yielded higher classification accuracy than the Complete Dataset Learning Approach.

The second comparative study entailed an inspection on the influence of feature selection. During Incomplete Dataset Learning we performed feature selection within each source individually using the large number of participants with incomplete data and applied these features to the participants with complete data for classification. This feature selection approach was found to yield higher classification accuracy than the Complete Dataset Learning Approach.

The inter-source feature relationship was inspected in several manners and all yield consistent findings. Firstly, Figure 7.10 demonstrates that features between sources show very little correlation (inter-source correlations) although large correlations exist within each sources (intra-source correlations). This lack of inter-source feature correlation explains why the use of SG-LASSO yields very similar results to that of traditional LASSO. Secondly, this finding was reinforced during the optimization of the SG-LASSO (Equation 7.6). It is evident that the SG-LASSO lasso procedure differs from traditional LASSO only through the addition of a regularisation term acting at the source-level. Indeed, during the optimization of the SG-LASSO equation, it was found that the value of λ_2 tended to be very close to zero as shown in Figure 7.12. This indicates that regularization occurs almost entirely at the feature level, and is virtually non-existent at the source level. As such, the SG-LASSO equation simply reduces down to that of traditional LASSO. Finally, this finding is further demonstrated by the similarity between features selected during the LASSO and SG-LASSO techniques in Table 7.9. It is also interesting to note that no features from the memory activity were in the top 10 highest weighted

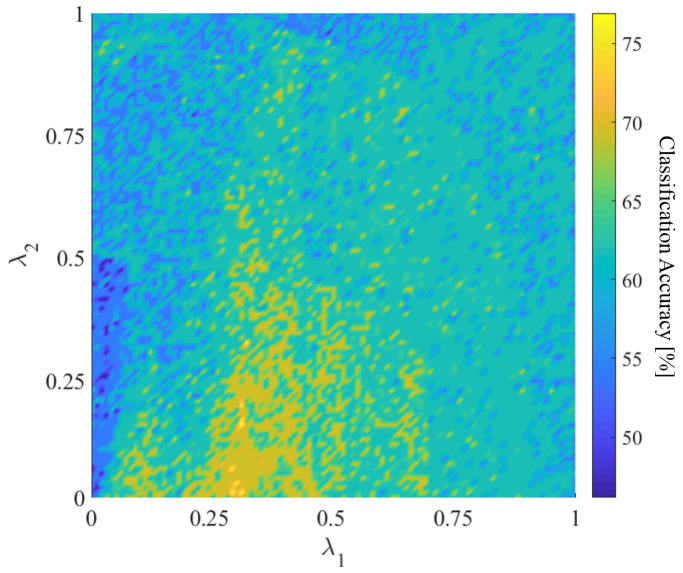


Figure 7.12: The classification accuracy achieved for each $\lambda_1 : \lambda_2$ ratio pair during sparse-group lasso optimisation. All units are in % This plot was obtained during the SG-LASSO optimisation outlined in §7.3.6.

features in any approach. Indeed, only 'Total Memory Score' was selected by all approaches but received very low weightings in the feature selection process.

Finally, the affect of sample size on feature distributions was explored, and consequently how sample size affects the confidence of feature selection. In Figure 7.11 it can be seen that feature values at small sample sizes are often transient, showing variable behaviour that is not representative of the population. However, with larger sample sizes the feature distributions reach a pseudo-steady state and show little variation. Figure 7.11(a) is an example of a voice feature that is selected by both the Incomplete Dataset Learning and the Complete Dataset Learning processes (Table 7.9). At sample sizes above 100 the feature distributions are stable and consistently differ between the disease groups. Figure 7.11(b) is an example of a tapping feature that is selected during Complete Dataset Learning, but not during Incomplete Dataset Learning. It is evident that at small sample sizes the feature values between the two groups appear to be large, but with the inclusion of more samples this difference is greatly reduced. Finally, in Figure 7.11(c) we show the behaviour of the common and interpretable 'Total Number

of Taps' feature from the tapping activity. Not only does this feature consistently show a difference between the disease groups, but it shows a gradual change of mean feature value. The variation of these feature distributions with sample size is attributed to high levels of noise in the feature set. Via the inclusion of more participants, the bootstrap estimates are less susceptible to noise and therefore provide better population estimates. As such, the features identified using the large sample size during Incomplete Dataset Learning are more robust and scalable than those identified by Complete Dataset Learning.

A limitation of the presented work is the assumption that all tests to have been completed correctly on the first attempt. Although mostly a correct assumption, it is expected that some tests used in this research were entirely noise and contain no relevant information. To alleviate the influence of this limitation, longitudinal test instances should be included. Included in the usage of longitudinal test instances should be the utilisation of multi-source ensemble learning for the creation of a continuous disease severity score. This would entail applying the method to longitudinal data and determining whether an objective multi-source composite score can be created that correlates with PD severity. Inclusion of the longitudinal data in the mPower dataset would increase the number of test instances by an order of magnitude; therefore creating an additional and substantial set of data to further test the potential of multi-source ensemble learning. An additional limitation in the present study is that the efficacy of multi-source ensemble learning on a dataset where inter-source relationships exist has not been evaluated. Future work should further assess multi-source ensemble learning (where feature selection only occurs at the feature level) via implementation on additional datasets wherein inter-source correlations are present. Similarly, multi-source ensemble learning has yet to be implemented on a dataset wherein a minority of observations possess missing data. In this scenario, fewer participants would be used during the development of individual source models than in their testing.

7.6 Conclusion

This chapter has explored the efficacy of using remotely collected data to perform disease classification. Due to the challenging data characteristics identified in the previous chapter, the application of traditional machine learning techniques would have resulted in over 90% of the participants being discarded from model development. As such, Multi-Source Ensemble Learning (M-SEL) has been proposed which enables all participants to be used in model development and validation whilst not requiring any imputation. Furthermore, during the implementation of M-SEL, a series of Convolutional Neural Networks (CNNs) were utilised on three different sources of raw time-series data for the purpose of automatic feature extraction and classification. Application of M-SEL achieves a classification accuracy of 82.0% demonstrating a large improvement compared to previous classification attempts on this dataset. Furthermore, multiple feature selection procedures were implemented and demonstrated the effect of sample size on feature distribution which further motivated the use of M-SEL in large and noisy datasets.

In the following chapter, M-SEL is further explored via its comparison to additional source-wise missing techniques on two vastly different datasets. This is intended to overcome the limitations highlighted in this chapter whilst determining the applicability of M-SEL when a minority of participants posses source-wise missing data.

8

Overcoming Source-wise Missing Data in Clinical and Remote Environments

Related Publications:

1) J. Prince, F. Andreotti, M. De Vos, "Effects of Source-wise Missing Data Strategies on Classifier Performance" in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, Brighton, (Invited Submission: 29thOctober 2018)

8.1 Introduction

The previous chapter motivated the formulation and implementation of multi-source ensemble learning (M-SEL) on the remotely collected mPower dataset. The primary benefit of this method is that it enables participants with source-wise missing data to be included in the development of machine learning models without the requirement to perform imputation. However, several potential limitations regarding the generalisation of M-SEL to other datatypes were highlighted in addition to how M-SEL performs under variable amounts of missing data when compared to alternative missing data strategies.

Traditional techniques for overcoming missing data, such as discarding and imputation, have been widely studied in relation to their effect on classifier performance. However, the appropriateness of these techniques are unknown when considering source-wise missing data. In this chapter, multiple techniques that compensate for missing data are applied to two multi-source datasets for the purpose of Parkinson's disease (PD) classification. This is performed via the simulation of variable amounts of source-wise missing data into complete datasets, and assessing the performance of each of the techniques under variable conditions.

Whereas the previous chapter focused on utilising the state-of-the-art classification techniques in order to maximise classification accuracy, this chapter is focused on providing a quantitative analysis of techniques for overcoming source-wise missing data using multiple datasets.

The primary contributions of this chapter are:

- A detailed quantitative assessment of missing data techniques on multiple source-wise missing datasets.
- The first implementation of a consistent classification protocol on two datasets for the purpose of PD classification.
- An implementation of a Deep Multi-Modal Autoencoder in the presence of missing data on multiple datasets with different degrees of noise and inter-source relationships.
- A quantitative assessment of the errors introduced by multiple techniques when applied to two source-wise missing datasets possessing different levels of noise.
- All python scripts for feature pre-processing and the training and validation of a Deep Multimodal Autoencoder model is freely available under a GNU GPL v3 license at `https://github.com/johnPrinceOx/Deep_Multi_Modal_Autoencoder_Neural_Networks`

8.2 Background

Imputation of missing values is common-place throughout engineering and statistics [338]. However, the success of imputation is dependent on a number of factors; most notably the quantity of missing data, the mechanism of the missingness, and the underlying relationship between the missing data and the present data [339]. Furthermore, studies comparing multiple imputation techniques have shown that these factors should govern which technique should be used in specific scenarios [340, 341].

The vast majoirty of studies investigating and utilising imputation techniques have been concerned with single-source datasets (§3.0.1) where it is common for there to be correlations within the feature set; a favourable trait for many popular imputation techniques [342, 343]. However, the applicability of these imputation techniques to multi-source datasets, where the features from separate sources may show little or no correlations, is poorly understood. Indeed, in the most recent review of multi-source learning techniques, the challenge of compensating for source-wise missing data is highlighted as an important aspect in need of more research [344, 345]. Furthermore, although [344] describes the concept of source-wise imputation as 'very much prone to error', there have been no quantitative studies validating the extent of this error, nor the impact of datatype upon the error.

The primary challenge in imputing missing source-data entails imputing the entire feature set of the missing source, as opposed to imputing a single feature as would be performed in a single-source dataset [346]. Accordingly, imputation of large quantities of features, from the potentially unrelated features of other sources, would negatively affect the performance of any machine learning model the dataset is subsequently used in [347].

The use of imputation learning in the previous chapter would have been highly inappropriate on account of 91.2% of the participants containing at least one missing source, whilst the underlying relationship between features being unknown. Alternatively, M-SEL was proposed as a methodology specifically for compensating for source-wise missing data and proved highly effective on the mPower dataset. However, as with traditional imputation techniques, M-SEL will only be applicable under certain circumstances and assumptions which are yet to be fully identified. Several limitations exist in the implementation of M-SEL of the previous chapter. Firstly, it is unknown how M-SEL compares to traditional imputation techniques under different amounts of missing data. Secondly, it is unknown how M-SEL performs on datasets of different structures i.e. when inter-source feature correlations exist. These are two important characteristics which this chapter aims to address.

In this chapter, a comparison is provided of three popular imputation techniques, a deep autoencoder, and M-SEL in regards to their ability to compensate for source-wise missing data when performing disease classification. To further assess each technique, each is implemented on two datasets that are juxtaposed by collection environment, test types, and data quality; enabling conclusions to be drawn regarding the influence of the underlying multi-source data structure and noise on each method. Finally, as the source-wise missing data is being simulated, the error induced by the imputation and deep autoencoding techniques can be quantified, allowing further insight into the applicability of each method.

8.3 Methods

8.3.1 Dataset Description

Data from both the mPower and OxQUIP datasets are used in this chapter. From the OxQUIP dataset, participants who contributed all sources at baseline visit were selected (N = 145). This is the same participant group used in §5.3.1 for the baseline clinical disease classification. The feature set of each of the three sources (S = 3) are the same as that described in §5.3.

From the mPower dataset, participants who contributed all sources at baseline (N = 133) are selected. This is the same participant group that formed the test set in §7.3.2 and that was used in the Complete Dataset Learning of §7.3.6. The feature set of each of the four sources (S = 4) are the same as that described in §7.3.3.

It should be noted that both of these participant groups have undergone extensive study previously in this thesis. As such, the previous results and conclusions may act as a 'benchmark' for which the results and conclusions from this analysis may act as a direct extension.

8.3.2 Simulation of Source-wise Missing Data

In each dataset, source-wise missing data is simulated on a participant-to-participant basis. Recall from the previous chapter that if a dataset contains S sources, and

participants are able to contribute sources in any permutation, a possible 2^S source-permutations exist. A participant is assigned a binary source vector representing which sources they have contributed with this vector determining to which domain the participant belongs (§7.3.2).

In this chapter, the reverse process is implemented through the simulation of source-wise missing data in otherwise complete multi-source datasets. As all participants in both datasets have contributed all sources, their binary source vectors are complete, $\mathbf{i} = \overrightarrow{1} = \{1\}^{1\times S}$. It is therefore possible to induce sparsity into the binary source vectors thus simulating the presence of source-wise missing data. A participant is assigned to a domain based on their randomly assigned binary source vector.

Furthermore, the extent to which the dataset is missing data can also be controlled using this approach. Firstly, the percentage of participants who will have incomplete data, m, can be fixed. Secondly, the distribution of the $m\%$ of participants within each of the domains can also be fixed. In §7.3.2, it is evident that the distribution of participants in each domain is highly variable. In this chapter, the probability of a participant being assigned to a domain (excluding the empty and complete domains) is equal such that:

$$P(p_n = d_i) = \frac{1}{2^S - 2} \tag{8.1}$$

where $P(p_n = d_i)$ is the probability that the n^{th} participant is assigned to the i^{th} domain. A schematic of the imputation process is provided in Figure 8.1.

A further benefit of utilising this domain assignment procedure is that it ensures the nature of the source-wise missing data is 'Missing Completely At Random' (MCAR) [348]. Missing Completely At Random data naturally occurs very rarely and is known to produce an unbiased platform for analyses [349, 350]. Examples of domain distributions for a three source dataset are given in Figure 8.2.

As such, the performance of each missing data strategy can be assessed when presented with different amounts of missing data. The techniques can therefore be compared and their relative successes at different degrees of missing data determined.

Figure 8.1: Simulation of missing data in a complete multi-source dataset containing N participants and a total of p features across all sources.

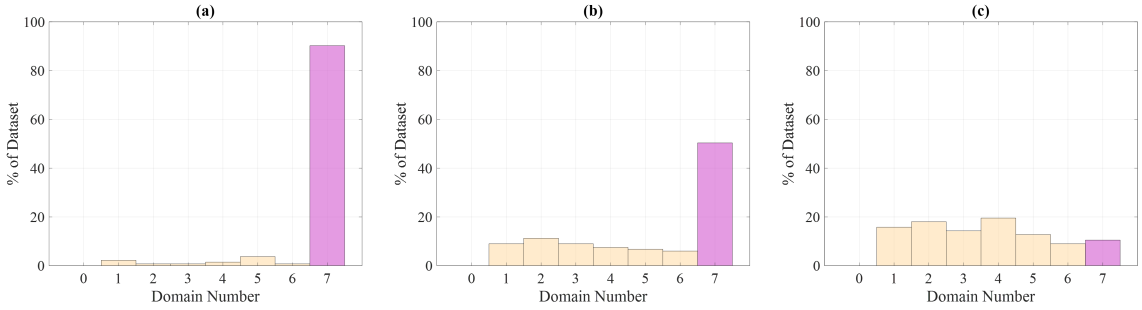


Figure 8.2: Example domain distributions for a three source dataset when (a) $m = 10$ (b) $m = 50$ and (c) $m = 90$. The complete domain in this example is Domain Number 7 which always contains (100-m)% of the participants. The remaining m% of participants are randomly distributed between the remaining incomplete domains such as to induce MCAR source data.

8.3.3 Missing Data Strategies

Complete Dataset Learning

Consistent with the previous chapter, only the participants with complete source data are used for model development and validation in complete dataset learning. The remaining participants, who are missing any data whilst, are discarded - otherwise known as *case deletion imputation* [351, 352]. This results in $(100 - m)$% of participants being used for model development in complete dataset learning at any given simulation. Intuitively, as the percentage of participants with simulated-missing data increase, there are fewer participants available for the model training and validation procedure. This simulates the very common and widely used missing-data scenario as was observed in the full mPower dataset of the previous chapter [232, 237]. The data preparation of complete dataset learning is visualised in Figure 8.3.

The benefits of this technique is that all models are developed and validated on

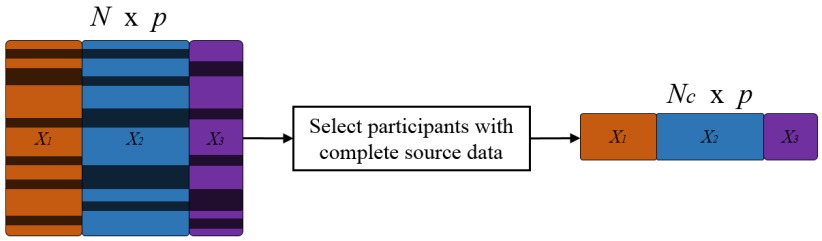


Figure 8.3: Dataset preparation for Complete Dataset Learning wherein only the N_c participants with complete source data are used for model development.

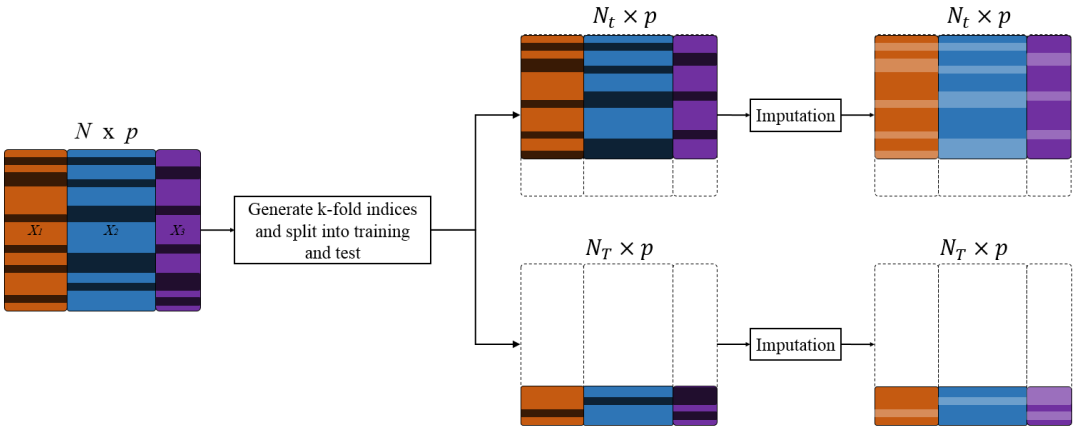


Figure 8.4: Data preparation for Source-wise Imputation Learning. All participants are split into training and validation sets containing N_t and N_T participants respectively. The two sets undergo imputation separately prior to model development.

participants with complete source data, thus removing the need for more complex models whilst also avoiding the need to perform imputation. The limitation of this method is that it often results in a large proportion, potentially the majority, of participants being discarded [344].

Source-Wise Imputation Learning

Chapter 5 demonstrated that the sources in the OxQUIP dataset contain many inter-source correlations. Conversely, Chapter 7 demonstrated that the sources in the mPower dataset do not contain any inter-source correlations. Through performing source-wise imputation learning on both datasets, it is possible to determine the effects of inter-source correlations during the imputation process. This allows investigation into the potential limitations of source-wise imputation proposed in the previous chapter and in existing literature [344].

In this implementation of source-wise imputation learning, two imputation techniques are used:

- Mean imputation - wherein a missing feature value is imputed as the mean of the non-missing values of that feature. This is the simplest imputation technique and will act as a benchmark for the other learning and imputation strategies.
- k-nearest neighbour (kNN) imputation - wherein a missing feature value is imputed using the k closest other observations who are not missing the feature. This approach assumes that a missing feature can be approximated using the underlying structure between other observations and other features [341].

The imputation pipeline is summarised in Algorithm 2 and the data preparation is visualised in Figure 8.4.

Algorithm 2 Procedure for simulating source-wise missing data and the preparation of data for implementing source-wise imputation learning.

```
Complete Dataset
for m : 0 → 100 do
  for 1 → iter do
    Balance disease groups (50%:50%)
    Randomly assign m% of participants into incomplete source domains
    for 1 → R do
      generate k-fold cross validation indices
      for 1 → k do
        Impute Training
        Impute Validation
        Model Development and Validation
      end for
    end for
  end for
end for
```

Note that imputation occurs after the definition of the training and validation sets such as to prevent data leakage.

In all simulations, the source-wise imputation approaches maintain a 100% participant retention rate for model development and validation. However, as

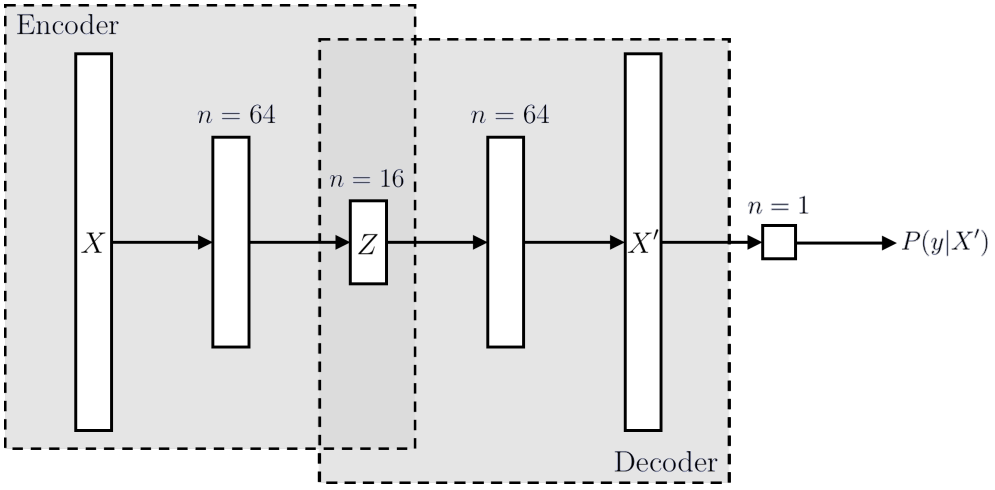


Figure 8.5: The architecture of the DMMAE including the stacked logistic regression classification neuron. All layers of the autoencoder undergo $\ell_2 - regularisation$.

more participants are induced with source-wise missing data, a larger proportion of features will be needed to be imputed.

Deep Multimodal Autoencoding

The most recently proposed method for compensating for source-wise missing data is that of the Deep Multimodal Autoencoder (DMMAE) [272, 273, 353]. The DMMAA entails a special usage of the traditional denoising autoencoder (§3.3.3) wherein the autoencoder reconstructs missing data by treating it as a form of noise.

In order for an autoencoder to know which values are missing, a unique pre-processing procedure is required. The initial data preparation is the same as that used during the imputation approaches (Figure 8.4). Within each of the cross-validation folds, the training and validation sets undergo zero-mean unit-variance normalization with respect to the training set. For the feature sets to be eligible for use in the DAE they must also undergo range normalization such that each feature is within the range [0,1]. This is performed separately on the training and validation sets. All missing values are ignored during both normalization processes. Unique to DMMAE pre-processing, all missing values are replaced with a 'special value' after the range normalisation. The special value is selected such that it is impossible for it to be present in the true feature set. As such, the special value is selected to have any value outside the range of [0,1]; in this work the special value is set as -1. The

setting of all missing values to a value outside the possible range of the true data is analogous to inserting noise into feature set, thus the DMMAE functions under the same principals as a denoising autoencoder. This process is described in Algorithm 3.

Algorithm 3 Procedure for simulating source-wise missing data and the preparation of data for implementing a DMMAE.

```
Complete Dataset
for m : 0 → 100 do
  for 1 → iter do
    Balance disease groups (50%:50%)
    Randomly assign m% of participants into incomplete source domains
    for 1 → R do
      generate k-fold cross validation indices
      for 1 → k do
        Zero-mean unit variance normalisation
        Range normalisation [0,1]
        Special value insertion
        DMMAE Training and Validation
      end for
    end for
  end for
end for
```

The architecture of the DMMAE is shown in Figure 8.5. The network consists of three hidden layers with 64, 16, and 64 units respectively . Each layer undergoes ℓ_2-norm regularisation and all units having a logistic transfer function. The logistic transfer function is the most appropriate as the output range is bounded such that it is the same as the range the feature set has been constrained to. Finally, classification is performed via stacking a single neuron with a logistic transfer function to the final layer of the autoencoder.

An additional benefit of the DMMAE is the inherent ability to perform feature selection in tandem with classification. Stacking the autoencoder (feature construction) and a logistic regression unit (classification) allows traditional back-propagation optimization to be applied as when using a deep neural network for classification purposes only [272].

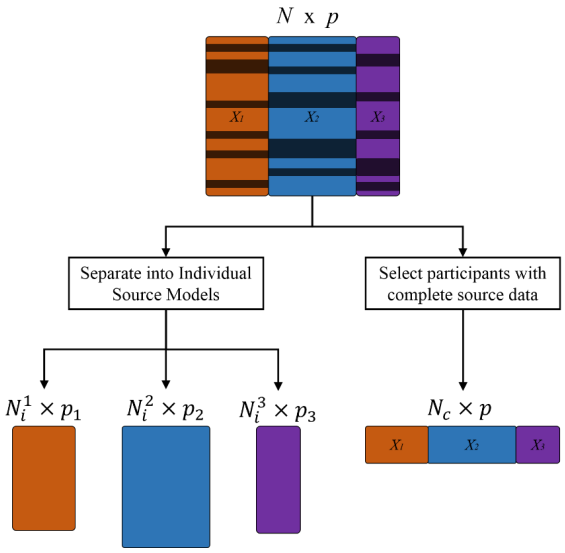


Figure 8.6: Data preparation for M-SEL. From the participants with missing data, N_i^n are used to developed and validate the n^{th} individual source model. There are N_c participants with complete source data who are used as a test set for the individual source models.

Multi-Source Ensemble Learning

This is the novel methodology proposed in the previous chapter. After simulating source-wise missing data, all participants with missing source data are assigned as training participants and are used to develop the individual source models. The participants with complete source data are reserved to act as the test group. Each individual source model can then be applied to the test participants, and the outcomes of each undergo a majority voting ensemble. Note that the test participants in the multi-source ensemble learning approach are the same participant group used to develop and validate the models in the complete dataset learning approach of §8.3.3. This process is visualised in Figure 8.6.

8.3.4 Evaluation & Comparison of Strategies

A summary of each of the three methods' strengths and weaknesses are summarised in Table 8.1.

For all learning approaches, logistic regression models are developed using repeated stratified 10-fold cross validation. Prior to being separated into folds,

	100% Participant Retention	Imputation
Complete Dataset Learning	No	No
Source-wise Imputation	Yes	Yes
DMMAE	Yes	Yes
Multi-Source Ensemble Learning	Yes	No

Table 8.1: Summary of the participant retention vs. need for imputation trade-off for each of the techniques.

all data undergoes minority class balancing and then undergoes the missing data simulation, as shown in Algorithm 2. In the case of M-SEL, a logistic regression model is developed for each of the individual source models. When implemented on the test set, the resulting outcomes of each individual source model are fused through majority voting. The binary classification accuracy of all models is reported at each simulation instance.

Feature Selection

LASSO feature selection is utilised in all approaches apart from the DMMAE. In complete dataset learning, the features selected from the training set are applied to the validation set during model evaluation. During the imputation approaches, feature selection is performed on the training set after the imputation, with the selected features being transferred to the imputed validation set. In M-SEL, feature selection is performed on each of the individual source models and then applied to the test set.

8.3.5 Imputation & Autoencoding Errors

As the data being estimated during the imputation and autoencoding methods has been simulated, and therefore the true values are known, it is possible to determine the error of the approaches at each simulation. This facilitates quantitative comparison between the suitability of performing imputation and DMMAE on each of the two datasets.

The errors induced by the DMMAE and imputation techniques is calculated for each training and validation set separately. The root mean squared error is

calculated between the imputed/autoencoded feature vector, $\boldsymbol{x}'$, and the true feature vector, $\boldsymbol{x}$, for each participant. It is important to note that the number of features within each source is different and varies significantly. As such, when a source containing many features is missing, a larger error would be expected than if a source with few features were missing. Furthermore, the error would be zero for all of the features that have not been imputed which would add an error bias towards instances where the sources with large number of features have not been imputed. Therefore the use of a normalised Root Mean Squared Error (RMSE) is necessary to avoid adding this bias into the error calculation. For each participant, all the features that were *not* missing are removed from both $\boldsymbol{x}'$ and $\boldsymbol{x}$. The RMSE for each participant is subsequently calculated using only the imputed or autoencoded values, with the normalising factor set as the number of missing features for that participant. For the n^{th} participant, this error is given by:

$$\epsilon_n = \frac{||\tilde{\boldsymbol{x}}_{\boldsymbol{n}} - \tilde{\boldsymbol{x}}'_n||^2}{n(\tilde{\boldsymbol{x}})} \tag{8.2}$$

where $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}'$ are the vectors containing only the imputed or autoencoded features and their corresponding true values respectively and $n(\tilde{\boldsymbol{x}})$ is the cardinality of $\tilde{\boldsymbol{x}}$ ($n(\tilde{\boldsymbol{x}}) = n(\tilde{\boldsymbol{x}}')$). The RMSE over a set of N participants is then given by:

$$\epsilon_{rms} = \sqrt{\sum_{n=1}^{N} \frac{\epsilon_n}{N}} \tag{8.3}$$

This is calculated for the training and validation sets on both datasets whilst varying the percentage of participants with missing data.

8.4 Results

OxQUIP

The classification results of all learning approaches on the OxQUIP dataset, with various degrees of missing data, are shown in Figure 8.7.

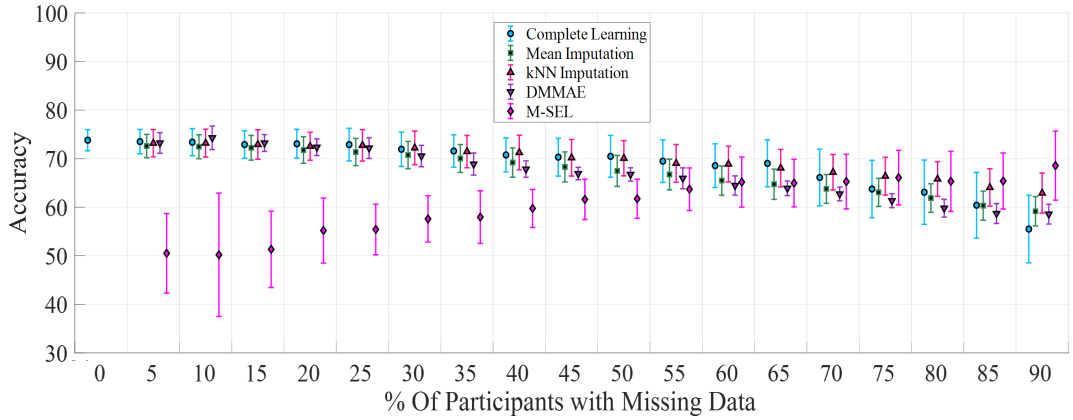


Figure 8.7: The performance of disease classification, as performed by multiple learning strategies, at different percentages of missing source data using the OxQUIP dataset. These results correspond to the methods outlined in §8.3.3.

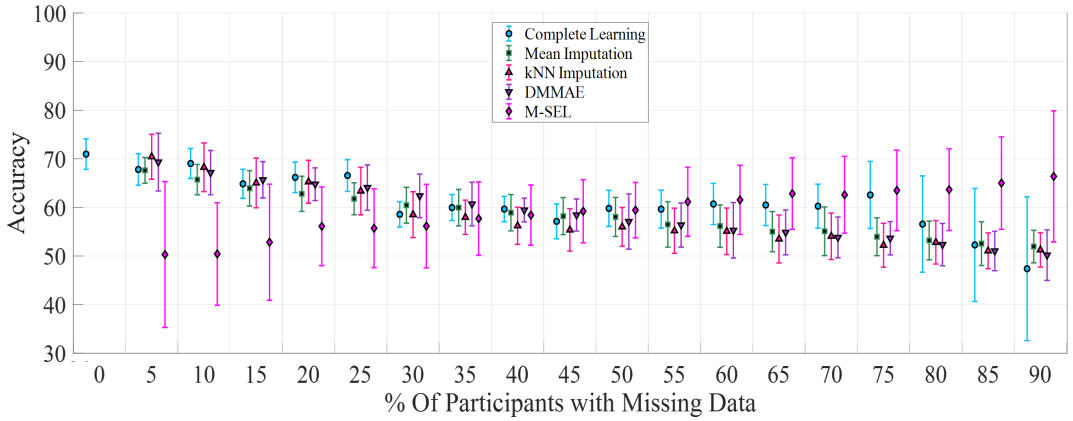


Figure 8.8: The performance of disease classification, as performed by multiple learning strategies, at different percentages of missing source data using the mPower dataset. These results correspond to the methods outlined in §8.3.3.

mPower

The results of all learning approaches on the mPower dataset, with various degrees of missing data, are shown in Figure 8.8.

Imputation Error Rates

Figure 8.9(a) and (b) shows the RMSE induced into the feature sets as a result of mean and kNN imputation. Figure 8.9(c) shows the corresponding errors induced by the DMMAE.

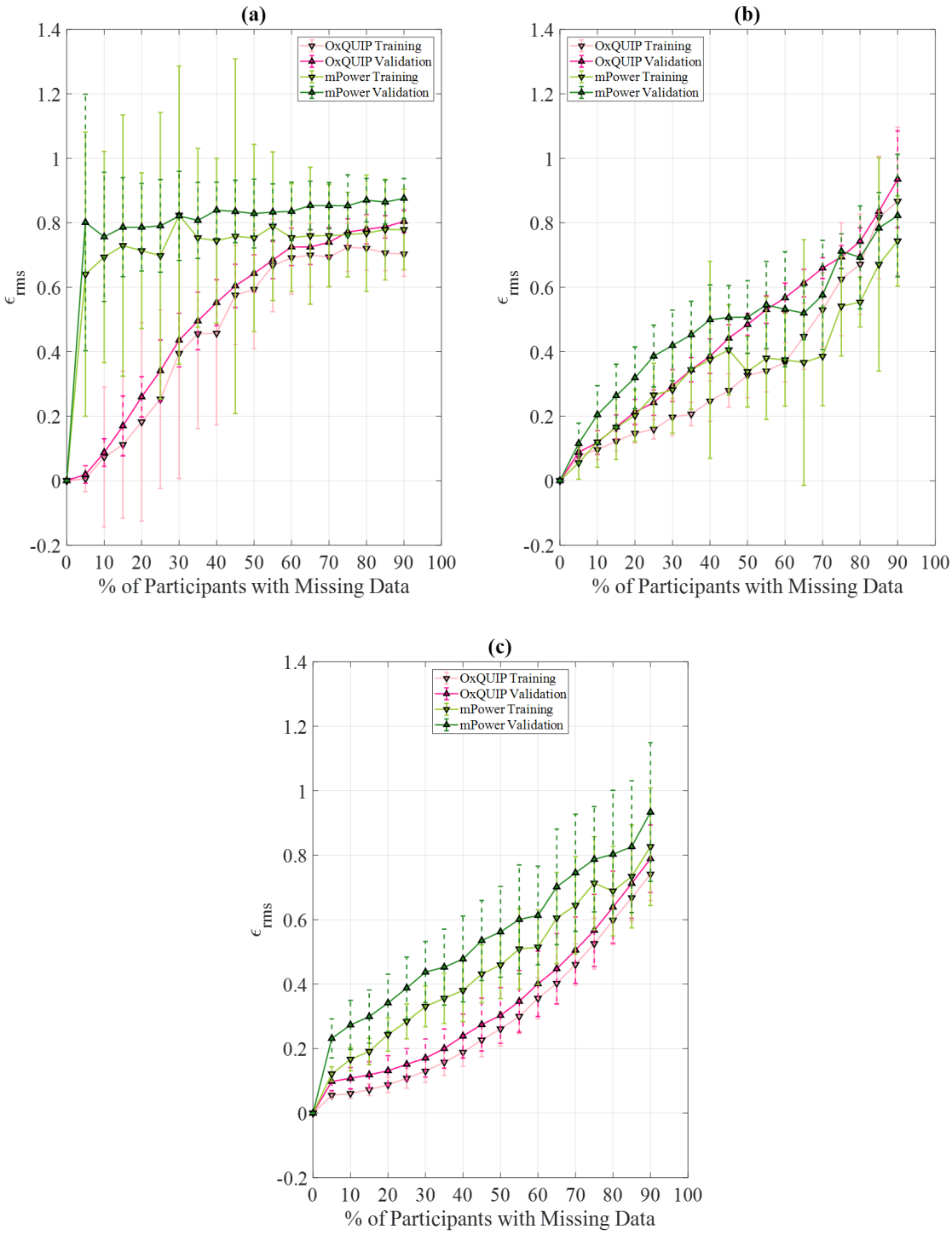


Figure 8.9: The error experienced by the training and validation sets in both datasets whilst undergoing (a) mean imputation, (b) kNN Imputation and (c) DMMAE. These results correspond to the methods outlined in §8.3.5.

8.5 Discussion

In this chapter, a comparison of multiple missing data strategies have been evaluated on two source-wise missing datasets. Via inducing variable amounts of MCAR source-wise data, the strength of each strategy was measured in terms of their ability to perform binary classification of Parkinson's disease as well as the magnitude of error induced by imputation. Furthermore, this analysis enables this thesis' novel Multi-Source Ensemble Learning technique to be compared against the current and state-of-the-art alternative strategies for overcoming source-wise missing data.

Figure 8.7 shows the classification performance of each strategy at various degrees of missing data when implemented on the OxQUIP dataset. Intuitively, the accuracy of complete dataset learning monotonically decreases as the quantity of missing data increases. This is because less data is available for model development. However, the decrease in accuracy is gradual and decreases by a total of 18.3%. This is gradual decrease is attributed to the cleanliness of the OxQUIP feature set. Even when a large number of observations are missing data, feature selection can still be successfully performed. However, the standard deviation of the complete dataset learning accuracy also increases as less participants are used to develop models; indicating a lower confidence in model performance.

The accuracy of the imputation learning approaches closely match that of complete dataset learning. This implies that source-wise imputation is being successfully performed; confirming that the separate sources of the OxQUIP dataset contain mutual information facilitating this successful source-wise imputation. This finding can be explained as a direct extension of the findings in 5. It was shown that the Walk and TUG sources yield the best classification accuracy when used alone, whereas the Sway test generally yields a lower classification accuracy. As such, if a participant is missing the Walk source, but possesses the TUG and Sway sources, many of the important Walking features can be successfully imputed from the TUG features.

The performance of M-SEL increases with the number of participants with missing data. This is because as more participants are induced with missing data,

more participants are assigned to develop the individual source models. Interestingly, MSEL only outperforms the alternative methods when a very high percentage of participants possess missing data.

Figure 8.8 shows the classification performance of each strategy at various degrees of missing data when implemented on the mPower dataset. The behaviour of all strategies follow very similar trends to that seen in the OxQUIP dataset with several important differences. Firstly, the rate at which the complete dataset learning and source-imputation learning accuracies decrease is much faster. This is accompanied by very large standard deviations of model predictions. Unlike in the OxQUIP dataset, source-imputation often yields lower classification accuracies when compared to complete dataset learning once roughly 25% of participants are missing data. This is due to the lack of inter-source feature correlations between sources making the imputed values likely to be poorly representative of their true values.

The interaction between M-SEL and the alternative approaches differ between the datasets. In the OxQUIP dataset, the classification accuracy of M-SEL only outperforms the other approaches when a very large $(75 - 90\%)$ of the participants have missing data. Conversely, in the mPower dataset, M-SEL outperforms the alternative methods with far fewer missing participants $(40 - 60\%)$. A further comment on the use of M-SEL in this analysis is that it is being used in its most basic form. In the previous chapter it was demonstrated that performing majority voting at the source level with only a logistic regression classifier yielded substantially lower classification accuracies than when ensemble methods at the classifier level were also employed. If multiple classifiers were to be used to develop each individual source model (such as a RF and a DNN), classifier fusion could be implemented as was suggested beneficial to the classification accuracy in the previous chapter.

Figure 8.9 provides quantitative error rates for both imputation techniques and the DMMAE when applied to the datasets. The mean imputation error (Figure 8.9(a)) reflects the quantity of noise in the mPower dataset. Mean imputation is commonly used when small amounts of data is missing, and the imputation aims to estimate the population mean from the non-missing samples. However, if a large

quantity of data is missing, or the remaining samples do not accurately reflect the population mean, imputation error rates will be large. As such, even when a small percentage of participants are missing in the mPower dataset the imputation error is large. Conversely, as the OxQUIP data is clean, the sample mean gradually deviates from the true mean with an increase in missing data, reflected by the pseudo-steady rate of increasing imputation error. As would be expected, the imputation error is larger in the validation set in both datasets on account of there being fewer samples.

Figure 8.9(b) shows the kNN imputation error rates of both datasets. As kNN exploits relationships between features (and therefore sources in this study), the kNN imputation error in the OxQUIP dataset is significantly lower than in the mPower dataset due to the aforementioned existence of inter-source correlations. Figures 8.9(a) and Figure 8.9(b) demonstrate that mean imputation is more applicable to clean datasets as imputation error rises quickly in the presence of missing data in noisy datasets whilst the error rate increases slower in clean datasets. Further, kNN is the more robust to the influence of noise of the two imputation techniques and, when presented with inter-source correlated features, presents the slowest rate of error increase.

Finally, the error introduced by the DMMAE is shown in Figure 8.9(c). The heteroscedastic response of the error due to missing data in the OxQUIP dataset is more favourable than the responses seen by either the mean of kNN approaches. A previously suggested benefit of utilising a DMMAE is the ability to optimise the classification procedure whilst also performing feature selection that is accounting for missing data. This approach is not possible in the other techniques as imputation, feature selection, and classification are all performed independently. It is the ability of the DMMAE to perform these tasks simultaneously that the slowest rate of error increase is attributed to in both datasets.

A consistent finding across both datasets is that the use of M-SEL is most appropriate when a majority of the participants in a dataset is missing data. However, the foremost conclusion of this analysis is that source-wise missing data strategy selection should be governed by the underlying nature of the multi-source

dataset. If a dataset is found to contain correlated-sources, and a small to moderate amount of missing data is present, then imputation or DMMAE techniques can yield similar results to if no missing data is present whilst maintaining a 100% data retention rate. Conversely, if weak correlations exist between sources, M-SEL is more appropriate than imputation or DMMAE provided a large quantity of data is missing.

There are several avenues for future work. Firstly, having compared the missing data strategies whilst using a single neuron classifier, more advanced classifiers and classification techniques could be utilised. Each missing data technique should be implemented and the resulting feature set utilised by several classifiers whose responses undergo a classifier ensemble. Secondly, an alternative missing data simulation strategy could be employed that better represents that observed in real life settings. This could include missing data in the form of 'Missing Not At Random' as governed by the domain distribution seen in §7.3.2. However, this would only be appropriate for constraining the use of these approaches to a specific dataset whose domain distribution is already known.

The architecture of the DMMAE utilised in this work is based on that proposed by [272] who found that the use of a DNN in conjunction with the DMMAE achieved a higher classification accuracy as to when being used with a logistic regression classifier. As such, it is hypothesised that the performance of the DMMAE could be improved via the adoption of alternative classifiers that are also capable of inherently performing feature selection. The use of a DMMAE alongside a DNN utilising regularisation would provide an additional opportunity for the network to perform feature selection. Indeed, the effect of dropout regularisation and alternative activation functions are also yet to receive significant attention during the optimisation of DMMAE for compensating for missing data and should be investigated during any future implementations.

From a more clinical perspective, this is the first study to the authors knowledge to implement consistent experimental/classification protocols on two independent datasets for the purpose of PD classification. This approach has been called for by the research community [80] such as provide a fair platform to compare

demographics and test types. As additional PD multi-source datasets are made available, the techniques outlined in this chapter can be further assessed in future work to further understand the effects of the number of sources and quantity of source-wise missing data on disease classification.

8.6 Conclusion

In this chapter, a comparison has been made regarding the ability of five algorithms to compensate for source-wise missing data. This was implemented on the clinical OxQUIP dataset in addition to the remote mPower dataset and presents the first instance of a consistent classification procedure being applied to multiple datasets for the purpose of Parkinson's disease classification. Furthermore, this chapter enabled the comparison of the novel Multi-source Ensemble Learning (M-SEL) algorithm, as proposed in the previous chapter, with the current state-of-the-art techniques including that of a deep multi-modal autoencoder. It was found that source-wise imputation techniques are more appropriate in the OxQUIP dataset than in the mPower dataset due to the existence of inter-source correlations. The use of M-SEL in both datasets is found to be most appropriate when a majority of the participants possess source-wise missing data on account of more participants being included in the training of the classifier models. A quantitative analysis of the errors introduced by the imputation and autoencoding strategies revealed the influence of noise on their performance. Errors were systematically higher in the noisy mPower dataset with mean imputation and the DMMAE proving the least and most robust methods for both datasets respectively. The findings highlight the importance that a source-wise missing data strategy should be selected based on the underlying structure of the dataset (inter-source correlations), the proportion of participants with missing data, and the degree of noise in the dataset.

9 Conclusions and Future Work

This thesis has explored the ability of digital sensors to perform automatic, objective, and quantitative assessment of Parkinson's disease (PD). As the current assessment of PD is performed on an infrequent basis and entirely in a clinical setting, much of this thesis has been dedicated to unearthing the role of digital sensors in performing remote disease assessment. In this concluding chapter, a summary of results from each chapter is provided highlighting their respective novelties and contributions to the field. Finally, remote monitoring of PD is still in its adolescence and accordingly new findings will inevitably lead to more questions. As such, the findings of this thesis are discussed in relation to their limitations and possible directions of future work with an emphasis on how digital sensors can be further leveraged to improve disease assessment.

9.1 Summary of Contributions

The majority of the previous attempts to quantify PD have been in a clinical environment. Although many of such studies successfully identified a range of features able to quantify disease presence and severity, they were limited by two primary factors. Firstly, low subject numbers alongside inconsistent test protocols made for poor generalisation of their findings. Secondly, few studies went beyond classical confirmatory tests to implement machine learning tasks for disease classification and severity regression. As such, chapter 5 was intended to overcome these limitations and demonstrate the ability of wearable sensors to differentiate disease groups whilst also performing disease classification and regression as part of routine clinical care on a large and diverse cohort.

A multitude of features arising from three clinical tests were found to show significant difference between disease groups whilst also showing strong correlations with clinically assigned MDS-UPDRS scores. An important aspect of these features was that they were derived from many different locations of the body; highlighting the importance of utilising multiple sensors unlike the experimental setup of previous studies. Utilising these features, disease classification and regression were performed with the walking test proving to be the single most capable clinical test of performing classification (75.8% accuracy) whilst the sway and timed-up-and-go tests also showed promising ability with accuracies of 62.8% and 73.3% respectively. When the features of all tests are used in conjunction, classification accuracy was maximised (77.9%). Disease severity prediction also proved successful achieving an average error rate of 3.85%.

Chapter 5 also presented the first classification and regression tasks that incorporated longitudinal data. Firstly, novel characteristics of feature variation were highlighted which, when included into the classification task, proved to significantly outperform the corresponding cross-sectional classifications with classification accuracies exceeding 85%. Similarly, objective longitudinal disease tracking was shown capable on a participant-to-participant level for participants at a wide range of disease severities.

However, clinical disease assessment is infrequent and only provides an instantaneous snap-shot of a patient's symptoms. As such, the role of smart-phones to perform frequent testings of multiple symptoms in remote environments is now being explored as an alternative assessment platform. Chapter 6 investigated how features collected from a motor and a non-motor test could be used to gain insight into the efficacy of performing longitudinal disease monitoring using remotely collected data originating from smart-phones. Firstly, the total number of taps completed by 548 participants during the common Alternate Finger Tapping (AFT) test were monitored on a longitudinal basis. This revealed previously unseen longitudinal behaviour wherein a larger proportion of PD participants were found to show transient and variable longitudinal performance. Further, PD participants

were found to take longer to reach a steady performance which was subsequently found to be more representative of their severity as determined through correlation analyses with self-reported MDS-UPDRS. The benefit of using a smart-phone for data collection is that in enables multiple types of data to be collected using a single device. As such, the same analysis was implemented using a non-motor test which similarly revealed variable longitudinal performance in both disease groups. Subsequently, the longitudinal behaviour of the motor and non-motor symptoms were found to be independent of one another whilst also proving to be independent of medication state (ON/OFF) and sex.

A secondary objective of chapter 6 was to highlight the characteristics of the data within the mPower dataset. As tests were conducted under variable environmental conditions and not under clinical supervision, the subsequent features were found to suffer from noise. Indeed, de-noising techniques were employed such as to ensure robust extraction of the features whose longitudinal behaviour were being investigated; a requirement not present in the corresponding OxQUIP analysis. Additionally, when inspecting the self-reported MDS-UPDRS scores, it was found that they showed significantly weaker correlations with the AFT test than when the test is performed in a clinical environment further suggesting noise is present in the test and severity scoring data. When comparing the behaviour of the motor test with the non-motor test, it was found that only a small subset of participants had contributed both tests longitudinally, demonstrating data was contributed inconsistently across the different tests. Finally, the participant retention rate was found to be very low with only 6.8% of 8,003 participants completing 20 or more instances in the most popular test: the AFT.

Having demonstrated the ability to perform disease classification in a clinical environment in chapter 5, the logical next step was to perform remote disease classification. However, chapter 6 revealed the mPower dataset as being heavily corrupted by noise and to contain large amounts missing data thus making traditional classification techniques highly inappropriate. Accordingly, the novel Multi-Source Ensemble Learning (M-SEL) methodology was presented in Chapter 7. During the

implementation of M-SEL, a dataset decomposition technique was combined with ensemble strategies that operate at both the classifier and source level. The source-wise ensemble strategies enabled 100% of participants to be used for developing machine learning models (even those with source-wise missing data) without the need to perform imputation whilst the classifier ensemble strategies were used to overcome noise. It was found that 91.2% of the participants possessed at least one missing source of data at baseline all of whom would have had to be omitted from model development if traditional missing data strategies had been employed.

Included in this implementation of M-SEL was the use of the state-of-the-art Deep Neural Networks and Convolutation Neural Networks (CNNs) classifiers. The latter of these classification techniques capitalised on the large quantity of raw times-series data available and was found to outperform all traditional feature based classifiers. Additionally, the inclusion of the CNN also proved highly beneficial when introduced into the ensemble strategies with classification accuracy being found to exceed 80%.

As four test types were present during the classification tasks of chapter 7, the relationship between tests and the effect of different feature selection strategies were also investigated. This demonstrated that the tests posses very little inter-source correlations which highlighted the inappropriateness of imputation techniques. The effect of sample size on feature distributions, and therefore feature selection strategies, further motivated the use of M-SEL as it was found that feature distributions varied significantly at lower sample sizes due to the noise within the mPower dataset.

The purpose of chapter 8 was two-fold. Firstly, as a novel missing data methodology had been presented in chapter 7, it was necessary to expose its strengths and weaknesses in relation to existing missing data strategies. Secondly, this could be performed using both the OxQUIP and mPower datasets, thus presenting an opportunity to perform a consistent analysis protocol on two independent datasets for the purpose of PD classification. Having extensively studied both datasets throughout the thesis, their underlying characteristics (such as noise, individual test performances, and inter-source correlations) were already known,

thus allowing conclusions to be drawn regarding the applicability of each strategy on different dataset types.

Via simulating source-wise missing data in each dataset, the capability of each source-wise missing data strategy was assessed under varying amounts of missing data. Similar trends were observed in both datasets, namely that as more data is missing from a dataset the less robust traditional imputation becomes. Conversely, as the amount of missing data increases in a dataset, the applicability of M-SEL increases as more participants are available for model development and thus providing a unique alternative methodology to existing techniques.

The primary conclusion of chapter 8 was that the technique used to compensate for source-wise missing data should not only be governed by the amount of missing data, but also by the underlying relationship of the sources. If the sources are known to have a degree of inter-source correlations (as in OxQUIP), imputation techniques can be appropriate. Conversely, if no inter-source correlations exist (as in mPower), imputation is inappropriate and alternative methods such as M-SEL are more appropriate. As such, when a dataset contains source-wise missing data, one must compromise between data retention and the error induced by imputation.

9.2 Future Work

Several technical and clinical avenues of future work have been motivated by the findings and limitations of this thesis which are discussed in this concluding section.

Given the demonstrated benefit of incorporating longitudinal data into the classification tasks of Chapter 5, the most logical progression of the work presented in this thesis would be to implement a similar longitudinal classification task on the mPower dataset. Chapter 7 performed classification using the first set of test instances from each participant. However, the results of the preceding chapter revealed that many participants exhibit variable longitudinal behaviour that deviate significantly from their first test performance. It would be therefore be recommended to determine the effect of this longitudinal variation on classification performance. However, a number of challenges face longitudinal classification using the mPower

dataset. Missing data, irregular contributions, and variable retention rates will all need to be accounted for. Nonetheless, incorporation of the mPower longitudinal data would increase the number of test instances by over an order of magnitude than that presented in Chapter 7. This presents the opportunity to either utilise multi-source ensemble learning on a longitudinal basis or to adopt alternative machine learning approaches such as multi-task Gaussian processes which are well suited to the study of irregularly sampled longitudinal data with missingness.

On a similar note, when utilising the mPower dataset, this thesis initially presented a methodology for monitoring longitudinal behaviour and secondly a methodology for disease classification. Remote severity regression, particularly for longitudinal severity tracking, has not been addressed in this thesis and presents an additional opportunity for future work. However, the self-reported MDS-UPDRS was completed by a significantly smaller subset of participants on a longitudinal basis and is heavily corrupted by subjectivity and therefore is an unreliable dependent variable. An alternative approach to tracking disease severity is the formation of a composite severity score whose formulation is based entirely on quantitative measures. If a probabilistic approach was taken to disease classification, the probability of having the disease could be interpreted as a proxy for disease severity i.e. $P(y = 1|\boldsymbol{x}) = 1$ is extremely severe and $P(y = 1|\boldsymbol{x}) = 0$ is completely healthy.

This thesis has identified and quantified the noise within a remotely collected dataset (Chapter 8). The translational invariance property of convolutional neural networks enabled the raw signals from multiple sources to be used for classification without the need to perform pre-processing. An alternative area of future work would be to develop algorithms capable of robust signal segmentation and determination of signal quality indices. These type of algorithms are essential if the vast quantities of passively collected data in remote environments are to be used for disease assessment. Passive data is composed of multiple activity types and therefore the continued development of activity recognition algorithms will allow applicable segments (such as periods of walking) to be segmented and inspected for disease specific features.

Medication response remains an active area of research, and has been the focus of most studies utilising remotely collected datasets. What these studies have failed to address is the longitudinal relationship between medication and symptom severity. Levodopa-induced dyskinesia effects a large proportion of patients receiving dopaminergic treatment and negatively impacts their quality of life. Remote datasets facilitate the longitudinal influence of dopaminergic treatment on many symptoms, on a high frequency basis and, given that each participant contributes many instances, on a personalised level. Personalised medication response directed towards detecting the progression of levodopa-induced dyskinesia should be prioritised by future works investigating medication response.

Finally, as the remote assessment of PD is a nascent field of research there remains plenty of room for growth regarding study design. Future studies should adopt data collection strategies enabling the fusion of clinically and remotely collected data. It does not seem unviable for a smart-phone application to be used to collect test data during routine clinical visits whilst also being used to collect remotely measured test data between clinical visits. During the clinical usage of such a system, the clinically assigned MDS-UPDRS score could be recorded (overcoming the limitations of the currently used self-reported MDS-UPDRS) whilst also collecting measurement data under clinical supervision (thus ensuring correct test implementation). Further, the incentive of frequently contributing remote data would be increased for a patient if the review of remotely collected data formed part of routine clinical care.

Digital sensors and smart-phones present the most viable platforms for transitioning the assessment of Parkinson's disease from a purely clinical environment to a clinical-remote hybrid. Achieving frequent, objective, and interpretable measures of multiple disease symptoms in a remote environment could guide clinical decisions and ultimately lead to improved disease management and a better quality of life for those afflicted.

A Appendix

A.1 OxQUIP: Mobility Lab Feature Sets

All of the clinical features automatically extracted from the OxQUIP TUG, sway, and walking tests by the Mobility Lab software are given in Tables A.1, A.2, and A.3 respectively. The mean and standard deviation of all features listed were extracted and used in all analyses.

Figure A.1 provides a visual interpretation of many of the features in Table A.3.

A.2 mPower: Multi-Source Feature Description

Table A.4 provides a description for the features selected during the multiple feature selection of §7.3.6 and reported in Table 7.9. For a more comprehensive description of the full feature set, including the derivation technique of each feature, the reader is directed to [39, 40, 184].

Table A.1: The automatically generated Mobility Lab features from the OxQUIP TUG Test.

Feature	Description
Duration (s)	Total duration of test
Sit to Stand - Duration (s)	Duration of sit-to-stand transition
Sit to Stand - Lean Angle (degs)	Peak (95%) angle in the sagittal plane during sit-to-stand transition
Sit to Stand - N	Number of sit-to-stand attempts
Stand to Sit - Duration (s)	Duration of stand-to-sit transition
Stand to Sit - Lean Angle (degs)	Peak (95%) angle in the sagittal plane during stand-to-sit transition
Stand to Sit - N	Number of stand-to-sit attempts
Turns - Angle (degs)	Peak (95%) angle during the turn before the stand-to-sit transition
Turns - Duration (s)	Duration of the turn before the stand-to-sit transition
Turns - N	Number of turn attempts before the stand-to-sit transition
Turns - Turn Velocity (degs/s)	Peak (95%) velocity during the turn before the stand-to-sit transition

Table A.2: The automatically generated Mobility Lab features from the OxQUIP SWAY Test.

Feature	Description
95% Ellipse Axis 1 Radius (m/s^2)	1st Radius encapsulating 95% of the acceleration path
95% Ellipse Axis 2 Radius (m/s^2)	2nd Radius encapsulating 95% of the acceleration path
95% Ellipse Radius Average (m/s^2)	Average Radius encapsulating 95% of the acceleration path
95% Ellipse Sway Area (m^2/s^4)	Sway area, computed as the area in the sway trajectory per unit of time
Centroidal Frequency (Hz)	Frequency of sway from the centroid of the sway power spectrum
Centroidal Frequency (Coronal) (Hz)	Frequency of Coronal sway from the centroid of the sway power spectrum
Centroidal Frequency (Sagittal) (Hz)	Frequency of Sagittal sway from the centroid of the sway power spectrum
Frequency Dispersion (AD)	Frequency dispersion Total
Frequency Dispersion (Coronal) (AD)	Frequency dispersion Total
Frequency Dispersion (Sagittal) (AD)	Frequency dispersion Total
Jerk (m^2/s^5)	Smoothness of sway from the time derivative of the lumbar's acceleration Normalized (both planes)
Jerk (Coronal) (m^2/s^5)	Jerk in the Coronal Plane
Jerk (Sagittal) (m^2/s^5)	Jerk in the Sagittal Plane
Mean Velocity (m/s)	Average velocity
Mean Velocity (Coronal) (m/s)	Average Coronal velocity
Mean Velocity (Sagittal) (m/s)	Average Sagittal velocity
Path Length (m/s^2)	Overall path length
Path Length (Coronal) (m/s^2)	Coronal path length
Path Length (Sagittal) (m/s^2)	Sagittal path length
RMS Sway (m/s^2)	Average Root Mean Squared Sway
RMS Sway (Coronal) (m/s^2)	Coronal Root Mean Squared Sway
RMS Sway (Sagittal) (m/s^2)	Sagittal Root Mean Squared Sway
Range (m/s^2)	Total range of the lumbar's acceleration trajectory
Range (Coronal) (m/s^2)	Total range of the lumbar's acceleration trajectory
Range (Sagittal) (m/s^2)	Total range of the lumbar's acceleration trajectory
Sway Area Radius (Coronal) (degs)	Radius encapsulating 95% of coronal sway degs
Sway Area Radius (Sagittal) (degs)	Radius encapsulating 95% of sagittal sway degs
Sway Area Rotation (degs)	Average sway degree rotation
Sway Area (degs ^{2})	Total sway degree area
RMS Sway (degs)	Average Root Mean Square of sway degs
RMS Sway (Coronal) (degs)	Average Coronal Root Mean Square of sway degs
RMS Sway (Sagittal) (degs)	Average Sagittal Root Mean Square of sway degs

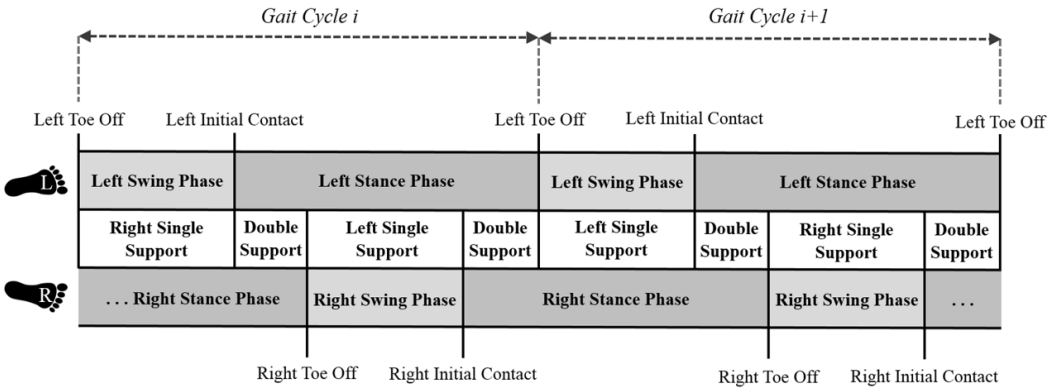


Figure A.1: A visualisation of the lower limb clinical features extracted by Mobility Lab.

Table A.3: The automatically generated Mobility Lab features from the OxQUIP Walk Test. Please note that units of % indicate the percentage of the total gait cycle. APA: Anticipatory Postural Adjustment, RoM: Range of Motion,LL: Lower Limb, UL: Upper Limb, Lum: Lumbar, Tnk: Trunk

Feature	Description
APA Duration (s)	APA Duration
First Step Duration (s)	Duration of the first step
First Step RoM (degs)	RoM of first step
Forward APA Peak (m/s^2)	Peak (95%) coronal APA acceleration during
Lateral APA Peak (m/s^2)	Peak (95%) sagittal APA acceleration during
Duration (s)	Total duration of test
LL Cadence L (steps/min)	Average left cadence
LL Cadence R (steps/min)	Average right cadence
LL Double Support L (%)	Average duration of left lower limb double stance
LL Double Support R (%)	Average duration of right lower limb double stance
LL Elevation at Midswing L (cm)	Average height of left foot during midswing
LL Elevation at Midswing R (cm)	Average height of right foot during midswing
LL Gait Cycle Duration L (s)	Average left lower limb gait cycle duration
LL Gait Cycle Duration R (s)	Average right lower limb gait cycle duration
LL Gait Speed L (m/s)	Average left lower limb gait speed
LL Gait Speed R (m/s)	Average right lower limb gait speed
LL Lateral Step Variability L (cm)	Average variability of left lower limb step length
LL Lateral Step Variability R (cm)	Average variability of right lower limb step length
LL Circumduction L (cm)	Average circumduction of left lower limb
LL Circumduction R (cm)	Average circumduction of right lower limb
LL N (steps)	Number of lower limb steps
LL Foot Strike Angle L (degs)	Average left lower limb foot strike angle
LL Foot Strike Angle R (degs)	Average right lower limb foot strike angle
LL Toe Off Angle L (degs)	Average left lower limb toe off angle
LL Toe Off Angle R (degs)	Average right lower limb toe off angle
LL Single Limb Support L (%)	Average duration of left lower limb single support
LL Single Limb Support R (%)	Average duration of right lower limb single support
LL Stance L (%)	Average duration of left limb stance
LL Stance R (%)	Average duration of right limb stance
LL Step Duration L (s)	Average duration of left limb step
LL Step Duration R (s)	Average duration of right limb step
LL Stride Length L (m)	Average length of left limb stride
LL Stride Length R (m)	Average length of right limb stride
LL Swing L (%)	Average duration of left lower limb swing
LL Swing R (%)	Average duration of right lower limb swing
LL Terminal Double Support L (%)	Average duration of left lower limb terminal double support
LL Terminal Double Support R (%)	Average duration of right lower limb terminal double support
LL Toe Out Angle L (degs)	Average left lower limb toe out angle
LL Toe Out Angle R (degs)	Average right lower limb toe out angle
Lum Coronal RoM (degs)	Average lumbar coronal RoM
Lum Sagittal RoM (degs)	Average lumbar sagittal RoM
Lum Transverse RoM (degs)	Average transverse coronal RoM
Tnk Coronal RoM (degs)	Average trunk coronal RoM
Tnk Sagittal RoM (degs)	Average trunk lumbar sagittal RoM
Tnk Transverse RoM (degs)	Average trunk transverse coronal RoM
UL Arm Swing Velocity L (degs/s)	Average velocity of left upper limb
UL Arm Swing Velocity R (degs/s)	Average velocity of right upper limb
UL Arm RoM L (degs)	Average RoM of left upper limb
UL Arm RoM R (deg)	Average RoM of right upper limb
Turns - Angle (degs)	Peak (95%) angle during all turning periods
Turns - Duration (s)	Average duration of all turning periods
Turns - N (n)	Total number of turns
Turns - Turn Velocity (degs/s)	Average turning velocity
Turns - Steps in Turn (n)	Average number of steps in a turn

Table A.4: A description of the features selected during the feature selection tasks in Chapter 7 as presented in Table 7.9.

Feature Name	Source of Origin	Description
std-MFCC-10th coef	Voice	The standard deviation of the 10th Mel Frequency Cepstral
Coefficient yGyro-Skewness	Walking	The skewness of the y-axis gyroscope waveform
xGyro-Mean	Walking	The mean of the x-axis gyroscope waveform
xAccel-STD	Tapping	The standard deviation of the x-axis accelerometer waveform
std-8th delta	Voice	The standard deviation of the first temporal derivative of the 8th Mel Frequency Cepstral Coefficient
GNE-NSR-TKEO	Voice	The Noise to Signal Ratio of the Taeger-Kaiser energy operator of the Glottal to Noise Excitation ratio
xGyro-DFA	Walking	The Dominant Frequency Amplitude of the x-gyroscope waveform
xAccel-Skewness	Walking	The skewness of the x-axis accelerometer waveform
Shimmer-F0mean	Voice	The mean shimmer (amplitude) of the vocal fundamental frequency
Shimmer-F0-prc75	Voice	The 75th percentile of the mean shimmer (amplitude) of the vocal fundamental frequency
zAccel-Median	Walking	The median of the z-axis accelerometer waveform
mean-MFCC-8th coef	Voice	The mean of the 10th Mel Frequency Cepstral Coefficient
zAccel-Skewness	Tapping	The skewness of the z-axis accelerometer waveform
yAccel-Kurtosis-STD	Walking	The standard deviation of the fourth central moment (kurtosis) in the y-accelerometer waveform
zAccel-AR-Lag-1	Tapping	The 1st coefficient of a 10th order autoregressive model on the z-accelerometer waveform
mean-delta delta 0th	Voice	The mean second temporal derivative of the mean of the 0th Mel Frequency Cepstral Coefficient
IQR of Tap Distance	Tapping	The interquartile range of the distance between consecutive taps (in pixels)
mean- delta-delta 12th	Voice	The mean second temporal derivative of the mean of the 12th Mel Frequency Cepstral Coefficient
zAccel-Kurtosis	Tapping	The kurtosis of the z-accelerometer gyroscope waveform
HNR-mean	Voice	The mean of the harmonics to noise ratio
GNE-SNR-TKEO	Voice	The Signal to Noise Ratio of the Taeger-Kaiser energy operator of the Glottal to Noise Excitation ratio
Shimmer-F0-prc5	Voice	The 5th percentile of the mean shimmer (amplitude) of the vocal fundamental frequency
Shimmer-F0-FM	Voice	The frequency modulation of the vocal fundamental frequency shimmer (amplitude)
mean-11th delta	Voice	The mean of the first temporal derivative of the 11th Mel Frequency Cepstral Coefficient
yAccel-DFA-STD	Walking	The standard deviation of the Dominant Frequency Amplitude of the y-gyroscope waveform
mean-MFCC-8th coef	Voice	The mean of the 8th Mel Frequency Cepstral Coefficient
zGyro-Median-STD	Walking	The standard deviation of the median of the z-axis gyroscope
waveform mean-MFCC-5th coef	Voice	The mean of the 5th Mel Frequency Cepstral Coefficient
zGyro-AR-Lag-1-STD	Walking	The standard deviation of the 1st coefficient of a 10th order autoregressive model on the z-gyroscope waveform

References

[1] J. Prince, S. Arora, and M. De Vos. "Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes". In: *Physiological Measurement* 39.4 (2018), p. 044005.

[2] J. Prince and M. De Vos. "A Deep Learning Framework for the Remote Detection of Parkinson's Disease Using Smart-phone Sensor Data". In: *Engineering in Medicine and Biology Society (EMBC), 2018 40th Annual International Conference of the IEEE*. 2018.

[3] J. Prince, F. Andreotti, and M. De Vos. "Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-wise Missing Data". In: *IEEE Transactions on Biomedical Engineering* (2018), pp. 1–10.

[4] J. Prince, F. Andreotti, and M. De Vos. "Effects of Source-wise Missing Data Strategies on Classifier Performance". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[5] T. Montine and K. Montine. "Precision medicine: Clarity for the clinical and biological complexity of Alzheimer's and Parkinson's diseases". In: *The Journal of Experimental Medicine* 212.5 (2015), pp. 601–605.

[6] O. Tysnes and A. Storstein. "Epidemiology of Parkinson's disease". In: *Journal of Neural Transmission* 124.8 (2017), pp. 901–905.

[7] S. Von Campenhausen et al. "Prevalence and incidence of Parkinson's disease in Europe". In: *European Neuropsychopharmacology* 15.4 (2005), pp. 473–490.

[8] The National Health Service Accessed 25 August 2018. *NHS Disorders: Parkinson's Disease https://www.nhs.uk/conditions/parkinsons-disease/*. 2016.

[9] A. Hughes et al. "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases". In: *Journal of neurology, neurosurgery, and psychiatry* 55 (1992), pp. 181–184.

[10] P. Martinez-Martin et al. "The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease". In: *Movement Disorders* 26.3 (2011), pp. 399–406.

[11] S. Fahn. "Description of Parkinson's Disease as a Clinical Syndrome". In: *Annals of the New York Academy of Sciences* 991.1 (2003), pp. 1–14.

[12] C. Davie. "A review of Parkinson's disease". In: *British Medical Bulletin* 86.1 (2008), pp. 109–127.

[13] W. Poewe. "The natural history of Parkinson's disease". In: *Journal of Neurology* 253.SUPPL. 7 (2006), pp. 2–6.

[14] M. Hely et al. "Sydney Multicenter Study of Parkinson's disease: Non-L-dopa-responsive problems dominate at 15 years". In: *Movement Disorders* 20.2 (2005), pp. 190–199.

[15] W. Maetzler et al. "Quantitative wearable sensors for objective assessment of Parkinson's disease". In: *Movement Disorders* 28.12 (2013), pp. 1628–1637.

[16] M. Naghavi et al. “Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013”. In: *The Lancet* 385.9963 (2015), pp. 117–171.

[17] P. McCrone, L. Allcock, and D. Burn. “Predicting the cost of Parkinson’s disease”. In: *Movement Disorders* 22.6 (2007), pp. 804–812.

[18] P. Lindgren et al. “Cost of Parkinson ’ s disease in Europe”. In: *European Journal of Neurology* 12 (2005), pp. 68–73.

[19] E. Dorsey, R. Constantinescu, and J. Thompson. “Projected number of people with Parkinson disease in the most populous nations”. In: *Neurology* january 30.68 (2007), pp. 384–6.

[20] E. Rovini, C. Maremmani, and F. Cavallo. “How wearable sensors can support Parkinson’s disease diagnosis and treatment: A systematic review”. In: *Frontiers in Neuroscience* 11.1 (2017), p. 555.

[21] J. FitzGerald et al. “Quantifying Motor Impairment in Movement Disorders”. In: *Frontiers in Neuroscience* 12.4 (2018), pp. 1–7.

[22] A. Schrag et al. “How valid is the clinical diagnosis of Parkinson’s disease in the community?” In: *Journal of neurology, neurosurgery, and psychiatry* 73.5 (2002), pp. 529–34.

[23] B. Post et al. “Unified Parkinson’s Disease Rating Scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?” In: *Movement Disorders* 20.12 (2005), pp. 1577–1584.

[24] T. Melzer et al. “Tracking Parkinson’s disease over one year with multimodal magnetic resonance imaging in a group of older patients with moderate disease”. In: *PLoS ONE* 10.12 (2015), pp. 1–14.

[25] J. Parkinson. “An essay on the shaking palsy”. In: *Journal of Neural Transmission* 14.2 (2002), p. 223 236.

[26] P. Kempster, B. Hurwitz, and A. Lees. “A new look at James Parkinson’s Essay on the Shaking Palsy”. In: *Neurology* 69.5 (2007), pp. 482–485.

[27] J. Jankovic. “Parkinson’s disease: clinical features and diagnosis”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 79.4 (2008), pp. 368–376.

[28] G. Cotzias. “L-Dopa for Parkinsonism”. In: *New England Journal of Medicine* 278 (1968), p. 630.

[29] S. Fahn and J. Jankovic. “Principles and practice of movement disorders.” In: *Journal of Neuro-Ophthalmology* 29.3 (2009), p. 255.

[30] J. Fearnley and A. Lees. “Ageing and Parkinson’s Disease: Substantia Nigra Regional Selectivity”. In: *Brain* 114.5 (1991), pp. 2283–2301.

[31] M. Alvarez, V. Evidente, and E. Driver-Dunckley. “Differentiating Parkinson ’ s Disease from Other Parkinsonian Disorders”. In: *Seminars in Neurology* 1.212 (2007), pp. 356–362.

[32] A. Rajput et al. “Epidemiology of Parhnsonism: Incidence , Classlfication, and Mortality”. In: *American Neurological Association and the Child Neurology* 16.3 (1984), pp. 278–282.

[33] N. Singh, V. Pillay, and Y. Choonara. "Advances in the treatment of Parkinson's disease". In: *Progress in Neurobiology* 81.1 (2007), pp. 29–44.

[34] D. Rye. "A population perspective on the IWG-2 research diagnostic criteria for Alzheimer's disease". In: 13.June (2014), pp. 532–534.

[35] A. Rajput et al. "Significance of Parkinsonian Manifestations in Essential Tremor". In: *Canadian Journal of Neurological Sciences* 20.2 (1993), pp. 114–117.

[36] R. Dewey. *Clinical features of Parkinson's disease. Parkinson's Disease and Movement Disorders: Diagnosis and Treatment Guidelines for the Practicing Physicians*, pp. 297–319.

[37] A. Rajput, B. Rozdilsky, and A. Rajput. "Accuracy of Clinical Diagnosis in Parkinsonism - A Prospective Study". In: *Canadian Journal of Neurological Sciences* 18.03 (1991), pp. 275–278.

[38] L. Hartelius and P. Svensson. "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey." In: *Folia phoniatrica et logopaedica* 46.1 (1994), pp. 9–17.

[39] A. Tsanas et al. "Accurate telemonitoring of Parkinsons disease progression by noninvasive speech tests". In: *IEEE Transactions on Biomedical Engineering* 57.4 (2010), pp. 884–893.

[40] A. Tsanas et al. "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity". In: *Journal of the Royal Society Interface* November 2010 (2010), pp. 842–855.

[41] T. Quatieri et al. "Neurophysiological Vocal Source Modeling for Biomarkers of Disease". In: *MIT Open Access Articles* (2017).

[42] O. Rascol et al. "Square wave jerks in parkinsonian syndromes." In: *Journal of neurology, neurosurgery, and psychiatry* 54.7 (1991), pp. 599–602.

[43] C. Antoniades et al. "The relationship between abnormalities of saccadic and manual response times in Parkinson's disease". In: *Journal of Parkinson's Disease* 3.4 (2013), pp. 557–563.

[44] E. Perry et al. "Cholinergic correlates of cognitive impairment in Parkinson's disease : comparisons with Alzheimer's disease". In: *Journal of Neurology, Neurosurgery & Psychiatry* 48.5 (1985), pp. 413–421.

[45] P. Remy et al. "Depression in Parkinson's disease: Loss of dopamine and noradrenaline innervation in the limbic system". In: *Brain* 128.6 (2005), pp. 1314–1322.

[46] K. Chaudhuri et al. "The Parkinson's disease sleep scale: A new instrument for assessing sleep and nocturnal disability in Parkinson's disease". In: *Journal of Neurology Neurosurgery and Psychiatry* 73.6 (2002), pp. 629–635.

[47] S. Carpi et al. "Non-motor symptoms in Parkinson's disease". In: *Current Neurobiology* 4 (2013), pp. 53–65.

[48] S. O'Sullivan et al. "Nonmotor symptoms as presenting complaints in Parkinson's disease: A clinicopathological study". In: *Movement Disorders* 23.1 (2008), pp. 101–106.

[49] G. Levy et al. “Memory and executive function impairment predict dementia in Parkinson’s disease”. In: *Movement Disorders* 17.6 (2002), pp. 1221–1226.

[50] P. LeWitt and S. Fahn. “Levodopa therapy for Parkinson disease: A look backward and forward”. In: *Neurology* 86.14 (2016), S3–S12.

[51] A. Björklund and S. Dunnett. “Dopamine neuron systems in the brain: an update”. In: *Trends in Neurosciences* 30.5 (2007), pp. 194–202.

[52] C. Marsden and J. Parkes. “Success and Problems of Long-Term Levodopa Therapy in Parkinson’s Disease”. In: *The Lancet* 309.8007 (1977), pp. 345–349.

[53] P. Lewitt. “Levodopa therapy for Parkinson’s disease: Pharmacokinetics and pharmacodynamics”. In: *Movement Disorders* 30.1 (2015), pp. 64–72.

[54] N. Keijsers et al. “Detection and Assessment of the Severity of Levodopa-Induced Dyskinesia in Patients With Parkinson’s Disease by Neural Networks”. In: *Movement Disorders* 15.6 (2000), pp. 1104–1111.

[55] C. Curtze et al. “Levodopa Is a Double-Edged Sword for Balance and Gait in People With Parkinson’s Disease”. In: *Movement Disorders* 30.10 (2015), pp. 1361–1370.

[56] S. Lee et al. “A novel method for assessing the severity of levodopa-induced dyskinesia using wearable sensors”. In: *EEE Engineering in Medicine and Biology Society* 2015-Novem.August (2015), pp. 8087–8090.

[57] W. Poewe, A. Lees, and G. Stern. “Low-dose L-dopa therapy in Parkinson’s disease : A 6-year follow-up study”. In: *Neurology* 36 (1983), pp. 1528–1530.

[58] A. Schrag and N. Quinn. “Dyskinesias and motor fluctuations in Parkinson’s disease: a community- based study”. In: *Brain* 123.11 (2000), pp. 2297–2305.

[59] M. Rodriguez-Oroz et al. “Bilateral deep brain stimulation in Parkinson’s disease: A multicentre study with 4 years follow-up”. In: *Brain* 128.10 (2005), pp. 2240–2249.

[60] R. Iansek, J. Rosenfeld, and F. Huxham. “Deep brain stimulation of the subthalamic nucleus in Parkinson’s disease”. In: *Medical Journal of Australia.* 177.3 (2002), pp. 142–146.

[61] G. Deuschl et al. “A Randomized Trial of Deep-Brain Stimulation for Parkinson”. In: *Lancet Neurology* (2010), pp. 581–591.

[62] J. Levin et al. “The Differential Diagnosis and Treatment of Atypical Parkinsonism”. In: *Deutsches Arzteblatt International* 113.5 (2016), pp. 61–69.

[63] J. Cooper et al. “Cognitive Impairment in Early , Untreated Parkinson’s Disease and Its Relationship To Motor Disability”. In: *Brain* 114 (1991), pp. 2095–2122.

[64] K. Chaudhuri, D. Healy, and A. Schapira. “Non-motor symptoms of Parkinson’s disease: diagnosis and management”. In: *The Lancet Neurology* 5.3 (2006), pp. 235–245.

[65] C. Haaxma et al. “Gender differences in Parkinson’s disease”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 78.8 (2007), pp. 819–824.

[66] J. Obeso et al. “Missing pieces in the Parkinson’s disease puzzle”. In: *Nature Medicine* 16.6 (2010), pp. 653–661.

[67] T. Buter et al. “Dementia and survival in Parkinson’s disease: a 12-years population study.” In: *Neurology* 70 (2008), pp. 1017–1022.

[68] M. Hoehn and M. Yahr. "Parkinsonism : onset, progression, and mortality Parkinsonism: onset, progression, and mortality". In: *Neurology* 17.5 (1967), pp. 427–442.

[69] C. Goetz et al. "Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations". In: *Movement Disorders* 19.9 (2004), pp. 1020–1028.

[70] R. Schwab and A. England. "Projection Technique for Evaluating Surgery in Parkinson's Disease". In: *Third Symposium on Parkinson's Disease. E&S Livingstone* (1969), pp. 152–157.

[71] C. Goetz et al. "Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results". In: *Movement Disorders* 23.15 (2008), pp. 2129–2170.

[72] T. Boonstra et al. "Gait disorders and balance disturbances in Parkinson's disease: clinical update and pathophysiology." In: *Current opinion in neurology* 21.4 (2008), pp. 461–471.

[73] A. Hughes et al. "Clinicopathologic Study of 100 Cases of Parkinson's Disease". In: *Archives of neurology* 50.2 (1993), pp. 140–148.

[74] S. Das et al. "Quantitative measurement of motor symptoms in Parkinson's disease: a study with full-body motion capture data." In: *IEEE Engineering in Medicine and Biology Society.* (2011), pp. 6789–6792.

[75] C. Goetz and G. Stebbins. "Assuring interrater reliability for the UPDRS motor section: Utility of the UPDRS teaching tape". In: *Movement Disorders* 19.12 (2004), pp. 1453–1456.

[76] C. Ramaker et al. "Systematic evaluation of rating scales for impairment and disability in Parkinson's disease". In: *Movement Disorders* 17.5 (2002), pp. 867–876.

[77] A. Abdolahi et al. "Potential reliability and validity of a modified version of the Unified Parkinson's Disease Rating Scale that could be administered remotely". In: *Parkinsonism and Related Disorders* 19.2 (2013), pp. 218–221.

[78] B. Bot et al. "The mPower study, Parkinson disease mobile data collected using ResearchKit". In: *Scientific Data* 3 (2016), p. 160011.

[79] A. Zhan et al. "High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection". In: *arXiv preprint* (2016), pp. 1–12. arXiv: `1601.00960`.

[80] Á. Sánchez-Ferro et al. "New methods for the assessment of Parkinson's disease 2005 to 2015: A systematic review". In: *Movement Disorders* 31.9 (2016), pp. 1283–1292.

[81] R. Drillis. "Objective Recording and Biomechanics of Pathological Gait". In: *Annals of the New York Academy of Sciences* 74.1 (1958), pp. 86–109.

[82] A. Salarian et al. "Gait assessment in Parkinson's disease: Toward an ambulatory system for long-term monitoring". In: *IEEE Transactions on Biomedical Engineering* 51.8 (2004), pp. 1434–1443.

[83] G. Lewis and S. Phillips. "Mobile Health Technologies". In: *Methods in molecular biology* 1256 (2015), pp. 213–29.

[84] B. Galna et al. “Progression of gait dysfunction in incident Parkinson’s disease: Impact of medication and phenotype”. In: *Movement Disorders* 30.3 (2015), pp. 359–367.

[85] A. Espay et al. “Technology in Parkinson’s disease: Challenges and opportunities”. In: *Movement Disorders* 31.9 (2016), pp. 1272–1282.

[86] Q. Oung et al. “Technologies for Assessment of Motor Disorders in Parkinson’s Disease: A Review”. In: *Sensors* 15.9 (2015), pp. 21710–21745.

[87] T. Nooritawati and H. Manap. “Parkinson Disease Gait Classification based on Machine Learning Approach”. In: *Journal of Applied Sciences* 12.2 (2012), pp. 180–185.

[88] G. Cavagna, F. Saibene, and R. Margaria. “A three-directional accelerometer for analyzing body movements”. In: *Journal of Applied Physiology* 16 (1961), p. 191.

[89] R. LeMoyne et al. “Accelerometers for Quantification of Gait and Movement Disorders: a Perspective Review”. In: *Journal of Mechanics in Medicine and Biology* 08.02 (2008), pp. 137–152.

[90] G. Ciuti et al. “MEMS sensor technologies for human centred applications in healthcare, physical activities, safety and environmental sensing: A review on research activities in Italy”. In: *Sensors* 15.3 (2015), pp. 6441–6468.

[91] “The Major Determinants in Normal and Pathological Gait”. In: *Surgery* 35.3 (1953), pp. 543–558.

[92] J. Barth et al. “Subsequence dynamic time warping as a method for robust step segmentation using gyroscope signals of daily life activities”. In: *IEEE Engineering in Medicine and Biology Society* April (2013), pp. 6744–6747.

[93] L. Palmerini et al. “Quantification of motor impairment in Parkinson’s disease using an instrumented timed up and go test”. In: *IEEE transactions on neural systems and rehabilitation engineering* 21.4 (2013), pp. 664–673.

[94] F. Parisi et al. “Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, sit-to-stand, and Gait tasks in Parkinson’s disease”. In: *IEEE Journal of Biomedical and Health Informatics* 19.6 (2015), pp. 1777–1793.

[95] S. Frenkel-Toledo et al. “Treadmill walking as an external pacemaker to improve gait rhythm and stability in Parkinson’s disease”. In: *Movement Disorders* 20.9 (2005), pp. 1109–1114.

[96] J. Stamatakis et al. “Gait feature extraction in Parkinson’s disease using low-cost accelerometers.” In: *IEEE Engineering in Medicine and Biology Society. Annual Conference* 2011 (2011), pp. 7900–3.

[97] N. Keijsers, M. Horstink, and S. Gielen. “Ambulatory motor assessment in Parkinson’s disease”. In: *Movement Disorders* 21.1 (2006), pp. 34–44.

[98] F. Parisi et al. “Inertial BSN-Based Characterization and Automatic UPDRS Evaluation of the Gait Task of Parkinsonians”. In: *IEEE Transactions on Affective Computing* 7.3 (2016), pp. 258–271.

[99] S. Arora et al. “Detecting and monitoring the symptoms of Parkinson’s disease using smartphones: A pilot study”. In: *Parkinsonism & Related Disorders* 21.6 (2015), pp. 650–653.

[100] F. Bugané et al. "Estimation of spatial-temporal gait parameters in level walking based on a single accelerometer: Validation on normal subjects by standard gait analysis". In: *Computer Methods and Programs in Biomedicine* 108.1 (2012), pp. 129–137.

[101] J. Myerson et al. "Balance and Gait Represent Independent Domains of Mobility in Parkinson Disease". In: *Physical Therapy* 9.2 (1998), pp. 1364–1371.

[102] A. Weiss et al. "Toward automated, at-home assessment of mobility among patients with Parkinson disease, using a body-worn accelerometer." In: *Neurorehabilitation and neural repair* 25.9 (2011), pp. 810–8.

[103] E. Grimpampi et al. "Estimate of lower trunk angles using gyroscope data in pathological gait". In: *Biosystems and Biorobotics* 1 (2013), pp. 747–751.

[104] C. Curtze et al. "Objective Gait and Balance Impairments Relate to Balance Confidence and Perceived Mobility in People With Parkinson's Disease." In: *Physical therapy* 96.11 (2016), pp. 1734–1743.

[105] S. Lord et al. "Executive dysfunction and attention contribute to gait interference in 'off' state Parkinson's Disease". In: *Gait and Posture* 31.2 (2010), pp. 169–174.

[106] M. Demonceau et al. "Contribution of a trunk accelerometer system to the characterization of Gait in patients with mild-to-moderate Parkinson's disease". In: *IEEE Journal of Biomedical and Health Informatics* 19.6 (2015), pp. 1803–1808.

[107] B. Bloem et al. "Falls and freezing of Gait in Parkinson's disease: A review of two interconnected, episodic phenomena". In: *Movement Disorders* 19.8 (2004), pp. 871–884.

[108] N. Giladi and A. Nieuwboer. "Understanding and treating freezing of gait in Parkinsonism, proposed working definition, and setting the stage". In: *Movement Disorders* 23.2 (2008), pp. 423–425.

[109] J. Han et al. "Gait analysis for freezing detection in patients with movement disorder using three dimensional acceleration system". In: *IEEE Engineering in Medicine and Biology* (2003), pp. 1863–1865.

[110] T. Morris et al. "Parkinsonism and Related Disorders A comparison of clinical and objective measures of freezing of gait in Parkinson's disease". In: *Parkinsonism and Related Disorders* 18.5 (2012), pp. 572–577.

[111] J. Hausdorff et al. "Impaired regulation of stride variability in Parkinson's disease subjects with freezing of gait." In: *Experimental brain research* 149.2 (2003), pp. 187–94.

[112] S. Vercruysse et al. "Freezing Beyond Gait in Parkinson's Disease: a review of current neurobehavioral evidence". In: *Neuroscience & Biobehavioral Reviews* 43.1 (2013), pp. 213–227.

[113] A. Snijders et al. "Clinimetrics of freezing of gait". In: *Movement Disorders* 23.2 (2008), pp. 468–474.

[114] A. Mirelman, N. Giladi, and J. Hausdorff. "Body-fixed sensors for Parkinson disease". In: *Journal of the American Medical Association* 314.9 (2015), pp. 873–874.

[115] M. Mancini et al. “Trunk accelerometry reveals postural instability in untreated Parkinson’s disease”. In: *Parkinsonism and Related Disorders* 17.7 (2011), pp. 557–562.

[116] M. Mancini et al. “Postural sway as a marker of progression in Parkinson’s disease: A pilot longitudinal study”. In: *Gait and Posture* 36.3 (2012), pp. 471–476.

[117] C. Baston et al. “Effects of Levodopa on Postural Strategies in Parkinson’s disease”. In: *Gait and Posture* 46 (2016), pp. 26–29.

[118] S. Mellone et al. “Hilbert-huang-based tremor removal to assess postural properties from accelerometers”. In: *IEEE Transactions on Biomedical Engineering* 58.6 (2011), pp. 1752–1761.

[119] D. Borja-Cacho and J. Matthews. “A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains”. In: *IEEE Engineering in Medicine and Biology Society* 22.3 (2014), p. 603.

[120] T. Chen et al. “Postural sway in idiopathic rapid eye movement sleep behavior disorder: A potential marker of prodromal Parkinsons disease”. In: *Brain Research* 1559 (2014), pp. 26–32.

[121] A. Frenklach et al. “Excessive postural sway and the risk of falls at different stages of parkinson’s disease”. In: *Movement Disorders* 24.3 (2009), pp. 377–385.

[122] L. Palmerini et al. “Feature selection for accelerometer-based posture analysis in Parkinsons disease”. In: *IEEE Transactions on Information Technology in Biomedicine* 15.3 (2011), pp. 481–490.

[123] M. Mancini and F. Horak. “Potential of APDM Mobility Lab for the monitoring of the progression of Parkinson’s disease”. In: *Expert review of medical devices* 13.5 (2016), pp. 455–462.

[124] J. Han et al. “Gait detection from three dimensional acceleration signals of ankles for the patients with Parkinson’s disease”. In: *IEEE The International Special Topic Conference on Information Technology in Biomedicine* 2628 (2006), pp. 1–4.

[125] M. Lewek et al. “Arm swing magnitude and asymmetry during gait in the early stages of Parkinson’s disease”. In: *Gait and Posture* 31.2 (2010), pp. 256–260.

[126] X. Huang et al. “Both coordination and symmetry of arm swing are reduced in Parkinson’s disease”. In: *Gait and Posture* 35.3 (2012), pp. 373–377.

[127] C. Cho et al. “A vision-based analysis system for gait recognition in patients with Parkinson’s disease”. In: *Expert Systems with Applications* 36.3, Part 2 (2009), pp. 7033–7039.

[128] Y. Fatmehsari and F. Bahrami. “Assessment of Parkinson’s disease: Classification and complexity analysis”. In: *17th Iranian Conference of Biomedical Engineering (ICBME)* (2010).

[129] M. Yoneyama et al. “Accelerometry-based gait analysis and its application to Parkinson’s disease assessment - part 2: a new measure for quantifying walking behavior.” In: *IEEE Engineering in Medicine and Biology Society* 21.6 (2013), pp. 999–1005.

[130] J. Klucken et al. “Unbiased and Mobile Gait Analysis Detects Motor Impairment in Parkinson’s Disease”. In: *PLoS ONE* 8.2 (2013).

[131] D. C. Dewey et al. “Automated gait and balance parameters diagnose and correlate with severity in Parkinson disease”. In: *Journal of the Neurological Sciences* 345.1 (2014), pp. 131–138.

[132] B. Scanlon et al. “An accelerometry-based study of lower and upper limb tremor in Parkinson’s disease”. In: *Journal of Clinical Neuroscience* 20.6 (2013), pp. 827–830.

[133] N. Ghassemi et al. “Combined accelerometer and EMG analysis to differentiate essential tremor from Parkinson’s disease”. In: *IEEE Engineering in Medicine and Biology Society* 2016.1 (2016), pp. 672–675.

[134] M. Budisic et al. “Distinguishing Parkinson’s disease and essential tremor with transcranial sonography”. In: *Acta Neurologica Scandinavica* 119.1 (2009), pp. 17–21.

[135] B. Harel, M. Cannizzaro, and P. J. Snyder. “Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: A longitudinal case study”. In: *Brain and Cognition* 56.1 (2004), pp. 24–29.

[136] K. Niazmand et al. “Quantitative evaluation of Parkinson’s disease using sensor based smart glove”. In: *24th International Symposium on Computer-Based Medical Systems* (2011), pp. 1–8.

[137] R. Dunnewold, C. Jacobi, and J. Van Hilten. “Quantitative assessment of bradykinesia in patients with Parkinson’s disease”. In: *Journal of Neuroscience Methods* 74.1 (1997), pp. 107–112.

[138] B. Printy et al. “Smartphone application for classification of motor impairment severity in Parkinson’s disease”. In: *IEEE Engineering in Medicine and Biology Society* (2014), pp. 2686–2689.

[139] A. Tavares et al. “Quantitative measurements of alternating finger tapping in Parkinson’s disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation”. In: *Movement Disorders* 20.10 (2005), pp. 1286–1298.

[140] M. Picillo et al. “Learning more from finger tapping in Parkinson’s disease: up and down from dyskinesia to bradykinesia”. In: *Movement Disorders Clinical Practice* 3.2 (2016), pp. 184–187.

[141] M. M et al. “Mobility lab to assess balance and gait with synchronized body-worn sensors”. In: *J Bioeng Biomed Sci* (2011), pp. 1–15.

[142] V. Ruonala et al. “EMG signal morphology and kinematic parameters in essential tremor and Parkinson’s disease patients”. In: *Journal of Electromyography and Kinesiology* 24.2 (2014), pp. 300–306.

[143] I. Hwang, C. Lin, and P. Wu. “Tremor modulation in patients with Parkinson’s disease compared to healthy counterparts during loaded postural holding”. In: *Journal of Electromyography and Kinesiology* 19.6 (2009), pp. 520–528.

[144] M. Alhamid, A. Alamri, and A. El Saddik. “Measuring hand-arm steadiness for post-stroke and Parkinson’s disease patients using SIERRA framework”. In: *IEEE International Workshop on Medical Measurements and Applications* (2010), pp. 6–9.

[145] D. Vaillancourt and K. Newell. “The dynamics of resting and postural tremor in Parkinson’s disease”. In: *Clinical Neurophysiology* 111.11 (2000), pp. 2046–2056.

[146] K. Niazmand et al. "A measurement device for motion analysis of patients with Parkinson's disease using sensor based smart clothes". In: *Pervasive Computing Technologies for Healthcare* (2011), pp. 9–16.

[147] Y. Zhou et al. "The measurement and analysis of Parkinsonian hand tremor". In: *IEEE International Conference on Biomedical and Health Informatics* (2016), pp. 414–417.

[148] G. Deuschl, P. Bain, and M. Brin. "Consensus Statement of the Movement Disorder Society on Tremor". In: *Movement Disorders* 13.3 (2008), pp. 2–23.

[149] A. Hossen et al. "Discrimination of parkinsonian tremor from essential tremor by implementation of a wavelet-based soft-decision technique on emg and accelerometer signals". In: *Biomedical Signal Processing and Control* 5.3 (2010), pp. 181–188.

[150] A. Butt et al. "Biomechanical parameter assessment for classification of Parkinson's disease on clinical scale". In: *International Journal of Distributed Sensor Networks* 13.5 (2017), p. 1550147717707417.

[151] C. De Frias et al. "Intraindividual variability in neurocognitive speed: A comparison of Parkinson's disease and normal older adults". In: *Neuropsychologia* 45.11 (2007), pp. 2499–2507.

[152] Y. Sano et al. "Quantifying Parkinson's disease finger-tapping severity by extracting and synthesizing finger motion properties". In: *Medical and Biological Engineering and Computing* 54.6 (2016), pp. 1–13.

[153] M. Tanaka et al. "Determinants of tapping speed in normal control subjects and subjects with Parkinson's disease: Differing effects of brief and continued practice". In: *Movement Disorders* 15.5 (2000), pp. 843–849.

[154] R. Okuno et al. "Measurement system of finger-tapping contact force for quantitative diagnosis of Parkinson's disease". In: *IEEE Engineering in Medicine and Biology* (2007), pp. 1354–1357.

[155] J. Stamatakis and B. Macq. "Finger Tapping clinimetric score prediction in Parkinson ' s Disease using low-cost accelerometers". In: *Computational intelligence and neuroscience* (2013), pp. 1–7.

[156] T. Arroyo-Gallego et al. "Detection of motor impairment in Parkinson's disease via mobile touchscreen typing". In: *IEEE Transactions on Biomedical Engineering* 64.9 (2017), pp. 1994–2002.

[157] Á. Jobbágy et al. "Analysis of finger-tapping movement". In: *Journal of Neuroscience Methods* 141.1 (2005), pp. 29–39.

[158] D. Heldman et al. "Clinician versus machine: Reliability and responsiveness of motor endpoints in Parkinson's disease". In: *Parkinsonism and Related Disorders* 20.6 (2014), pp. 590–595.

[159] M. Braybrook et al. "An Ambulatory Tremor Score for Parkinson's Disease". In: *Journal of Parkinson's Disease* 6.4 (2016), pp. 723–731.

[160] G. Rigas et al. "Assessment of tremor activity in the parkinsons disease using a set of wearable sensors". In: *IEEE Transactions on Information Technology in Biomedicine* 16.3 (2012), pp. 478–487.

[161] M. Memedi et al. "Automatic and objective assessment of alternating tapping performance in Parkinson's disease." In: *Sensors* 13.12 (2013), pp. 16965–16984.

[162] C. Lainscsek et al. "Finger tapping movements of Parkinson's disease patients automatically rated using nonlinear delay differential equations". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22.1 (2012), pp. 1–13.

[163] S. Patel et al. "Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors". In: *IEEE Transactions on Information Technology in Biomedicine* 13.6 (2009), pp. 864–873.

[164] J. Jankovic. "Parkinson's disease: clinical features and diagnosis". In: *Journal of Neurology, Neurosurgery & Psychiatry* 79.4 (2008), pp. 368–376.

[165] D. Kempler and D. Lancker. "Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease". In: *Brain and Language* 80.3 (2002), pp. 449–464.

[166] L. Hartelius and P. Svensson. "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey." In: *Folia phoniatrica et logopaedica* 46.1 (1994), pp. 9–17.

[167] M. Velasco García et al. "Acoustic analysis of voice in Huntington's disease patients". In: *Journal of Voice* 25.2 (2011), pp. 208–217.

[168] A. Tsanas. "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning". Doctoral Thesis. University of Oxford, 2012, p. 261.

[169] J. Canter. "Characteristics of Patients with Parkinson's Disease: Intensity, Pitch, and Duration". In: *Journal of Speech & Hearing Disorders* 28.3 (1963), pp. 221–229.

[170] A. Ho et al. "Speech Impairment in a Large Sample of Patients with Parkinson's Disease". In: *Behavioural Neurology* 11.3 (1999), pp. 131–137.

[171] E. Metter and W. Hanson. "Clinical and acoustic variability in hypokynetic dysarthria". In: *Journal of communication disorders* 19 (1986), pp. 347–366.

[172] G. Canter. "Speech characteristics of patients with Parkinson's disease: II." In: *Journal of Speech & Hearing Disorders* 30.1 (1965), pp. 44–49.

[173] J. Logemann et al. "Frequency and Cooccurence of Vocal Tract Dysfunctions in the Speech of a Large Sample of Parkinson's Patients". In: *Journal of Speech & Hearing Disorders* 40.3 (1978), pp. 47–57.

[174] P. Zwirner, T. Murry, and G. E. Woodson. "Phonatory function of neurologically impaired patients". In: *Journal of Communication Disorders* 24.4 (1991), pp. 287–300.

[175] A. Bayestehtashk et al. "Fully automated assessment of the severity of Parkinson's disease from speech". In: *Computer Speech and Language* 29.1 (2015), pp. 172–185.

[176] C. Perez et al. "Diagnosis and Tracking of Parkinson's Disease by using Automatically Extracted Acoustic Features". In: *Journal of Alzheimer's Disease & Parkinsonism* 6.5 (2016).

[177] P. Boersma. "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound". In: *Proceedings of the Institute of Phonetic Sciences* 17 (1993), pp. 97–110.

[178] A. De Cheveigné and H. Kawahara. "YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1917–1930.

[179] D. Talkin. "A robust algorithm for pitch tracking". In: *Speech Coding and Synthesis* (1995).

[180] P. Naylor et al. "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.1 (2007), pp. 34–43.

[181] A. Camacho and J. G. Harris. "A sawtooth waveform inspired pitch estimator for speech and music". In: *The Journal of the Acoustical Society of America* 124.3 (2008), pp. 1638–1652.

[182] H. Kawahara et al. "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2008), pp. 3933–3936.

[183] X. Sun. "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing* 1 (2002), pp. 333–336.

[184] A. Tsanas et al. "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity". In: *International Symposium on Nonlinear Theory and its Applications* (2010), pp. 457–460.

[185] J. Schoentgen and R. de Guchteneere. "Time series analysis of jitter". In: *Journal of Phonetics* 23.1-2 (1995), pp. 189–201.

[186] C. Ferrer-Riesgo and E. Gonzalez-moreira. "Progress in Pattern Recognition, Image Analysis and Applications". In: 4225.November (2006).

[187] M. Little et al. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease". In: *IEEE Transactions on Biomedical Engineering* 56.4 (2009), pp. 1015–1022.

[188] K. Polat. "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering". In: *International Journal of Systems Science* 43.4 (2012), pp. 597–609.

[189] M. Hariharan, K. Polat, and R. Sindhu. "A new hybrid intelligent system for accurate detection of Parkinson's disease". In: *Computer Methods and Programs in Biomedicine* 113.3 (2014), pp. 904–913.

[190] F. Ström and R. Koker. "A parallel neural network approach to prediction of Parkinson's Disease". In: *Expert Systems with Applications* 38.10 (2011), pp. 12470–12474.

[191] M. Shahbakhi, D. Far, and E. Tahami. "Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine". In: *Journal of Biomedical Science and Engineering* 07.04 (2014), pp. 147–156.

[192] Z. Bakar et al. "Classification of Parkinson's disease Based on Multilayer Perceptrons, Neural Network and ANOVA as a Feature Extraction". In: *IEEE International Colloquium on Signal Processing and its Applications* (2012), pp. 63–67.

[193] T. Bocklet et al. “Detection of persons with Parkinson’s disease by acoustic, vocal, and prosodic analysis”. In: *IEEE Workshop on Automatic Speech Recognition and Understanding* (2011), pp. 478–483.

[194] D. Aarsland et al. “Mild cognitive impairment in Parkinson disease: A multicenter pooled analysis”. In: *Neurology* 75.12 (2010), pp. 1062–1069.

[195] T. W. Robbins and R. Cools. “Cognitive deficits in Parkinson’s disease: A cognitive neuroscience perspective”. In: *Movement Disorders* 29.5 (2014), pp. 597–607.

[196] A. Pascual-Leone et al. “Procedural learning in Parkinson’s disease and cerebellar degeneration”. In: *Ann.Neurol.* 34.4 (1993), pp. 594–602.

[197] R. Siegert et al. “Is implicit sequence learning impaired in Parkinson’s Disease? A meta-analysis”. In: *Brain and Cognition* 67.3 (2008), pp. 351–359.

[198] M. Nissen and P. Bullemer. “Attentional requirements of learning: Evidence from performance measures”. In: *Cognitive Psychology* 19.1 (1987), pp. 1–32.

[199] D. Muslimović et al. “Motor procedural learning in Parkinson’s disease”. In: *Brain* 130.11 (2007), pp. 2887–2897.

[200] O. Koenig, C. Thomas-Antérion, and B. Laurent. “Procedural learning in Parkinson’s disease: Intact and impaired cognitive components”. In: *Neuropsychologia* 37.10 (1999), pp. 1103–1109.

[201] G. Jackson et al. “Serial reaction time learning and Parkinson’s disease: Evidence for a procedural learning deficit”. In: *Neuropsychologia* 33.5 (1995), pp. 577–593.

[202] H. Krebs et al. “Procedural motor learning in parkinson’s disease”. In: *Experimental Brain Research* 141.4 (2001), pp. 425–437.

[203] G. M. Jackson et al. “Serial reaction time learning and Parkinson’s disease: Evidence for a procedural learning deficit”. In: *Neuropsychologia* 33.5 (1995), pp. 577–593.

[204] E. Arroyo-Anll et al. “Procedural learning of semantic categorization in Parkinson’s disease.” In: *Journal of Alzheimer’s Disease* 45.1 (2015), pp. 205–216.

[205] Newzoo. *Global Mobile Market Report 2017*. Tech. rep. 2017.

[206] Ofcom. *The Communications Market Report*. Tech. rep. 2015, p. 431.

[207] O. Incel, M. Kose, and C. Ersoy. “A Review and Taxonomy of Activity Recognition on Mobile Phones”. In: *BioNanoScience* 3.2 (2013), pp. 145–171.

[208] N. Palmius. “Personalised Modelling of Geographic Movements in Depression”. Doctoral Thesis. University of Oxford, 2017, p. 250.

[209] N. Palmius et al. “Detecting Bipolar Depression From Geographic Location Data”. In: *IEEE Transactions on Biomedical Engineering* 64.8 (2017), pp. 1761–1771.

[210] D. Springer, L. Tarassenko, and G. D. Clifford. “Support Vector Machine Hidden Semi-Markov Model-based Heart Sound Segmentation”. In: *Computing in Cardiology* (2014), pp. 625–628.

[211] D. Dunsmuir et al. “Development of mHealth Applications for Pre-eclampsia Triage”. In: *IEEE Journal of Biomedical and Health Informatics* PP.99 (2014), pp. 1–1.

[212] K. Bobrow et al. "Efficacy of a text messaging (SMS) based intervention for adults with hypertension: protocol for the StAR (SMS Text-message Adherence suppoRt trial) randomised controlled trial." In: *BMC public health* 14 (2014), p. 28.

[213] D. Springer et al. "The SMS-text adherence support (StAR) study". In: *International Conference on Information and Communications Technologies and Development Notes* 2 (2013), pp. 147–150.

[214] *uMotif Inc., Accessed August 7 2018, https://www.umotif.com/.*

[215] A. Brajdic and R. Harle. "Walk detection and step counting on unconstrained smartphones". In: *International joint conference on Pervasive and ubiquitous computing* (2013), p. 225.

[216] S. Mellone, C. Tacconi, and L. Chiari. "Validity of a Smartphone-based instrumented Timed Up and Go". In: *Gait and Posture* 36.1 (2012), pp. 163–165.

[217] F. Sposaro and G. Tyson. "iFall: An android application for fall monitoring and response". In: *IEEE Engineering in Medicine and Biology Society* (2009), pp. 6119–6122.

[218] N. Kosse et al. "Validity and reliability of gait and postural control analysis using the tri-axial accelerometer of the IPod Touch". In: *Annals of biomedical engineering* 43.8 (2015), pp. 1935–46.

[219] R. LeMoyne and T. Mastroianni. "Use of Smartphones and Portable Media Devices for Quantifying Human Movement Characteristics of Gait, Tendon Refl ex Response, and Parkinson's Disease Hand Tremor". In: *Mobile Health Technologies: Methods and Protocols* (2015), pp. 435–458.

[220] S. Mazilu and M. Hardegger. "Online detection of freezing of gait with smartphones and machine learning techniques". In: *Pervasive Computing Technologies for Healthcare* (2012).

[221] L. Pepa et al. "Smartphone based Fuzzy Logic freezing of gait detection in Parkinson's Disease". In: *International Conference on Mechatronic and Embedded Systems and Applications* (2014).

[222] S. Antos, M. Albert, and K. Kording. "Hand, belt, pocket or bag: Practical activity tracking with mobile phones". In: *Journal of Neuroscience Methods* 231 (2014), pp. 22–30.

[223] V. Sharma et al. "Spark: Personalized Parkinson disease interventions through synergy between a smartphone and a smartwatch". In: *Lecture Notes in Computer Science* 8519.3 (2014), pp. 103–114.

[224] H. Zhang et al. "A handheld inertial pedestrian navigation system with accurate step modes and device poses recognition". In: *IEEE Sensors Journal* 15.3 (2015), pp. 1421–1429.

[225] M. Shoaib et al. "Complex human activity recognition using smartphone and wrist-worn motion sensors". In: *Sensors* 16.4 (2016), pp. 1–24.

[226] R. Ellis et al. "A validated smartphone-based assessment of gait and gait variability in Parkinson's disease". In: *PLoS ONE* 10.10 (2015).

[227] N. Kostikis et al. "Smartphone-based evaluation of parkinsonian hand tremor: Quantitative measurements vs clinical assessment scores". In: *IEEE Engineering in Medicine and Biology Society* August 2014 (2014), pp. 906–909.

[228] P. Kassavetis et al. "Developing a tool for remote digital assessment of Parkinson's disease". In: *Movement disorders Clinical Practice* 3 (2016), pp. 59–64.

[229] A. Woods et al. "Parkinson's disease and essential tremor classification on mobile device". In: *Pervasive and Mobile Computing* 13 (2014), pp. 1–12.

[230] C. Lee et al. "A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease". In: *PLoS ONE* 11.7 (2016), pp. 1–11.

[231] R. Okuno et al. "Finger taps movement acceleration measurement system for quantitative diagnosis of Parkinson's disease". In: *IEEE Engineering in Medicine and Biology* (2006), pp. 6623–6626.

[232] A. Zhan et al. "Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity". In: *JAMA Neurology* 21218 (2018).

[233] S. Arora et al. "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study". In: *Parkinsonism and Related Disorders* 21.6 (2015), pp. 650–653.

[234] S. Arora et al. "High Accuracy Discrimination of Parkinson's Disease Participants from Healthy Controls Using Smartphones". In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2014), pp. 10–13.

[235] K. Lan and W. Shih. "Early Diagnosis of Parkinson's Disease Using a Smartphone". In: *Procedia Computer Science* 34 (2014), pp. 305–312.

[236] P. Ginis et al. "Feasibility and effects of home-based smartphone-delivered automated feedback training for gait in people with Parkinson's disease: A pilot randomized controlled trial". In: *Parkinsonism and Related Disorders* 22 (2016), pp. 28–34.

[237] E. Neto et al. "On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: the mPower case study". In: *arXiv preprint* (2017), pp. 1–27. arXiv: `1706.09574`.

[238] E. Neto et al. "Personalized hypothesis test for detecting medication response in Parkinson disease patients using iPhone sensor data". In: *Pacific Symposium on Biocomputing* 21 (2016), pp. 273–84.

[239] M. Capecci et al. "A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease". In: *Gait and Posture* 50 (2016), pp. 28–33.

[240] S. Mazilu et al. "A Wearable Assistant for Gait Training for Parkinson's Disease with Freezing of Gait in Out-of-the-Lab Environments". In: *ACM Transactions on Interactive Intelligent Systems* 5.1 (2015), pp. 1–31.

[241] T. Quatieri et al. "Neurophysiological Vocal Source Modeling for Biomarkers of Disease". In: *MIT Open Access Articles* (2017).

[242] J. Konczak et al. "The perception of passive motion in Parkinson's disease". In: *Journal of Neurology* 254.5 (2007), pp. 655–663.

[243] M. Suzuki, H. Mitoma, and M. Yoneyama. "Quantitative Analysis of Motor Status in Parkinson's Disease Using Wearable Devices: From Methodological Considerations to Problems in Clinical Applications". In: *Parkinson's Disease* 2017.1 (2017).

[244] A. Vauvelle. "Gait Analysis for Objectifying Parkinson's Assessment". Masters Thesis. University of Oxford, 2017, p. 47.

[245] S. Del Din et al. "Free-living monitoring of Parkinson's disease: Lessons from the field". In: *Movement Disorders* 31.9 (2016), pp. 1293–1313.

[246] L. Lin. "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* 45.1 (1989), pp. 255–268.

[247] J. De Winter, S. Gosling, and J. Potter. "Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data". In: *Psychological Methods* 21.3 (2016), pp. 273–290.

[248] S. Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Board of the Foundation of the Scandinavian Journal of Statistics* 6.1 (1978), pp. 65–70.

[249] P. G. James. *Modern Engineering Mathematics*. 2010, p. 1128.

[250] M. Anderson. "A new method for non-parametric multivariate analysis of variance". In: *Austral ecology* 26.2001 (2001), pp. 32–46.

[251] S. Zabell. "On student's 1908 article 'the probable error of a mean'". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 1–7.

[252] H. Lilliefors. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 399–402.

[253] J. Felsenstein. "Confidence limits on phylogenies: an approach using the bootstrap". In: *Evolution* 39.4 (1985), pp. 783–791.

[254] J. Felsenstein. "Confidence Limits on Phylogenies: an Approach Using the Bootstrap". In: *Society for the Study of Evolution* 39.1 (1985), pp. 1–15.

[255] J. Neter. *Applied Linear Statistical Models*. 4. 1996, pp. 1–16.

[256] A. Demaris. "A Tutorial in Logistic Regression". In: *National Council on Family Relations* 57.4 (1995), pp. 956–968.

[257] A. Tikhonov and V. Arsenin. "Solutions of Ill-Posed Problems ." In: *American Mathematical Society* 32.144 (1978), pp. 1320–1322.

[258] A. Hoerl and R. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *American Statistical Association and American Society for Quality* 12.1 (1970), pp. 55–67.

[259] T. Johansen. "On Tikhonov Regularization, Bias and Variance in Nonlinear System Identification". In: *Automatica* 33.2 (1997), pp. 441–446.

[260] I. Guyon and A. Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3.3 (2003), pp. 1157–1182.

[261] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009, p. 491.

[262] J. Tang, S. Alelyani, and H. Liu. "Feature Selection for Classification: A Review". In: *Data Classification: Algorithms and Applications* (2014), pp. 37–64.

[263] L. Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[264] M. Pal and P. Mather. "An assessment of the effectiveness of decision tree methods for land cover classification". In: *Remote Sensing of Environment* 86.4 (2003), pp. 554–565.

[265] C. Strobl et al. "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9 (2008), pp. 1–11.

[266] R. Genuer, J. Poggi, and C. Tuleau-Malot. "Variable selection using random forests". In: *Pattern Recognition Letters* 31.14 (2010), pp. 2225–2236.

[267] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. 2016, p. 800.

[268] D. Ravi et al. "Deep Learning for Health Informatics". In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21.

[269] L. Deng and D. Yu. "Deep Learning: Methods and Applications". In: *Foundations and Trends in Signal Processing* 7.3-4 (2014), pp. 197–387.

[270] J. Schmidhuber. "Deep Learning in neural networks: An overview". In: *Neural Networks* 61 (2015), pp. 85–117.

[271] W. Liu et al. "A survey of deep neural network architectures and their applications". In: *Neurocomputing* 234.December 2016 (2017), pp. 11–26.

[272] N. Jaques et al. "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction". In: *International Conference on Affective Computing and Intelligent Interaction*. 2017, pp. 202–208.

[273] F. Bianchi et al. "Learning representations for multivariate time series with missing data using Temporal Kernelized Autoencoders". In: *arXiv preprint* (2018), pp. 1–16. arXiv: 1805.03473.

[274] P. Vincent and H. Larochelle. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *Journal of Machine Learning Research* 11 (2010), pp. 3371–3408.

[275] A. Karpathy et al. "Large-scale video classification with convolutional neural networks". In: *Computer Vision and Pattern Recognition* (2014), pp. 1725–1732.

[276] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances In Neural Information Processing Systems* (2012), pp. 1–9.

[277] J. Yang, K. Yu, and T. Huang. "Supervised translation-invariant sparse coding". In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 3517–24.

[278] R. Girshick. "Fast R-CNN". In: *EEE International Conference on Computer Vision* 2015 Inter (2015), pp. 1440–1448.

[279] D. Cireşan, U. Meier, and J. Schmidhuber. "Multi-column Deep Neural Networks for Image Classification". In: *arXiv preprint* (2012). arXiv: 1202.2745.

[280] S. Han et al. "Learning both Weights and Connections for Efficient Neural Networks". In: *Advances in neural information processing systems* (2015), pp. 1135–1143.

[281] F. Agostinelli et al. "Learning Activation Functions to Improve Deep Neural Networks". In: *arXiv preprint* 2013 (2014), pp. 1–9. arXiv: 1412.6830.

[282] S. Hochreiter. "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06.02 (1998), pp. 107–116.

[283] Y. Lecun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[284] N. Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958.

[285] L. Bottou. "Large Scale Machine Learning with Stochastic Gradient Descent". In: *Proceedings of COMPSTAT 0210* (2010), pp. 177–186.

[286] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv preprint* (2014), pp. 1–15. arXiv: `1412.6980`.

[287] F. Moreno-Seco et al. "Comparison of classifier fusion methods for classification in pattern recognition tasks". In: *Lecture Notes in Computer Science* 4109 (2006), pp. 705–713.

[288] D. Opitz and R. Maclin. "Popular Ensemble Methods: An Empirical Study". In: *Journal of Artificial Intelligent Research* 11 (1999), pp. 169–198.

[289] D. Ruta and B. Gabrys. "An Overview of Classifier Fusion Methods". In: *Computing and Information Systems* 7.1 (2000), pp. 1–10.

[290] R. Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *International Joint Conference on Artificial Intelligence* March 2001 (2016), pp. 1137–1143.

[291] I. Rahman et al. "Ergothioneine inhibits oxidative stress- and TNF-alpha-induced NF-kappa B activation and interleukin-8 release in alveolar epithelial cells." In: *Biochemical and biophysical research communications* 302.4 (2003), pp. 860–4.

[292] APDM. *Mobility Lab User's Guide*. Tech. rep. APDM Wearable Sensors, 2014.

[293] J. Cooke et al. "Npsnet: Flight Simulation Dynamic Modeling Using Quaternions". In: *Presence* 1.4 (1994), pp. 404–420.

[294] M. Tundo, E. Lemaire, and N. Baddour. "Correcting Smartphone orientation for accelerometer-based analysis". In: *MeMeA 2013 - IEEE International Symposium on Medical Measurements and Applications* (2013), pp. 58–62.

[295] A. Vienne et al. "Inertial Sensors to Assess Gait Quality in Patients with Neurological Disorders : A Systematic Review of Technical and Analytical Challenges". In: *Frontiers in psychology* 8 (2017), pp. 1–12.

[296] J. M. Hausdorff. "Gait variability : methods , modeling and meaning". In: *Journal of NeuroEngineering and Rehabilitation* 2 (2005), pp. 1–10.

[297] J. Verghese et al. "Quantitative gait markers and incident fall risk in older adults". In: *Journals of Gerontology* 64.8 (2009), pp. 896–901.

[298] A. McDonough et al. "The validity and reliability of the GAITRite system's measurements: A preliminary evaluation". In: *Archives of Physical Medicine and Rehabilitation* 82.3 (2001), pp. 419–425.

[299] J. Schlachetzki et al. "Wearable sensors objectively measure gait parameters in Parkinson's disease". In: *PLoS ONE (submitted)* (2017), pp. 1–10.

[300] S. Del Din, A. Godfrey, and L. Rochester. "Validation of an Accelerometer to Quantify a Comprehensive Battery of Gait Characteristics in Healthy Older Adults and Parkinson's Disease: Toward Clinical and at Home Use". In: *IEEE Journal of Biomedical and Health Informatics* 20.3 (2016), pp. 838–847.

[301] R. LeMoyne. "Wearable and wireless accelerometer systems for monitoring Parkinson's disease patients—A perspective review". In: *Advances in Parkinson's Disease* 02.04 (2013), pp. 113–115.

[302] C. F. Pasluosta et al. "An emerging era in the management of Parkinson's disease: Wearable technologies and the internet of things". In: *IEEE Journal of Biomedical and Health Informatics* 19.6 (2015), pp. 1873–1881.

[303] B. Mollenhauer et al. "Monitoring of 30 marker candidates in early Parkinson disease as progression markers". In: *Neurology* 87.2 (2016), pp. 168–177.

[304] S. Fereshtehnejad et al. "New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression". In: *JAMA Neurology* 72.8 (2015), p. 863.

[305] M. Little et al. "Quantifying short-term dynamics of parkinson's disease using self-reported symptom data from an internet social network". In: *Journal of Medical Internet Research* 15.1 (2013), pp. 1–11.

[306] G. Yogev et al. "Gait asymmetry in patients with Parkinson's disease and elderly fallers: When does the bilateral coordination of gait require attention?" In: *Experimental Brain Research* 177.3 (2007), pp. 336–346.

[307] M. Plotnik et al. "Is freezing of gait in Parkinson's disease related to asymmetric motor function?" In: *Annals of Neurology* 57.5 (2005), pp. 656–663.

[308] P. DasMahapatra et al. "Free-Living Physical Activity Monitoring in Adult US Patients with Multiple Sclerosis Using a Consumer Wearable Device". In: *Digital Biomarkers* 2.1 (2018), pp. 47–63.

[309] S. Patel et al. "Monitoring Motor Fluctuations in Patients With Parkinson's Disease Using Wearable Sensors". In: *IEEE Transactions on Information Technology in Biomedicine* 13.6 (2009), pp. 864–873.

[310] S. Lord et al. "Gait variability in Parkinson's disease: An indicator of non-dopaminergic contributors to gait dysfunction?" In: *Journal of Neurology* 258.4 (2011), pp. 566–572.

[311] B. Dijkstra, Y. Kamsma, and W. Zijlstra. "Detection of gait and postures using a miniaturized triaxial accelerometer-based system: Accuracy in patients with mild to moderate Parkinson's disease". In: *Archives of Physical Medicine and Rehabilitation* 91.8 (2010), pp. 1272–1277.

[312] M. Bryant et al. "Effects of levodopa on forward and backward gait patterns in persons with Parkinson's disease". In: *NeuroRehabilitation* 29.3 (2011), pp. 247–252.

[313] M. Pimentel et al. "A review of novelty detection". In: *Signal Processing* 99 (2014), pp. 215–249.

[314] S. Eddy. "Profile hidden Markov models." In: *Bioinformatics* 14.9 (1998), pp. 755–763.

[315] C. Rasmussen and K. Williams. *Gaussian processes for machine learning.* Vol. 14. 2. 2004, pp. 69–106.

[316] K. Kubota, J. Chen, and M. Little. “Machine learning for large-scale wearable sensor data in Parkinson’s disease: Concepts, promises, pitfalls, and futures”. In: *Movement Disorders* 31.9 (2016), pp. 1314–1326.

[317] R. Okuno et al. “Finger taps movement acceleration measurement system for quantitative diagnosis of Parkinson’s disease”. In: *IEEE Engineering in Medicine and Biology* (2006), pp. 6623–6626.

[318] L. Helmuth, U. Mayr, and I. Daum. “Sequence learning in Parkinson’s disease : a comparison of spatial- attention and number-response sequences”. In: *Psychological Research* 38 (2000), pp. 1443–1451.

[319] P. Kraus et al. “Analysis of the course of Parkinson’s disease under dopaminergic therapy: Performance of "fast tapping" is not a suitable parameter”. In: *Movement Disorders* 20.3 (2005), pp. 348–354.

[320] A. Behrman, J. Cauraugh, and K. Light. “Practice as an intervention to improve speeded motor performance and motor learning in Parkinson’s disease”. In: *Journal of the Neurological Sciences* 174.2 (2000), pp. 127–136.

[321] K. Light et al. “Reaction Times and MovementTimes: Benefits of Practice to Younger and Older Adults”. In: *Human kinetics Publishers* (1996), pp. 27–41.

[322] B. Sahakian et al. “A comparative study of visuospatial memory and learning in Alzheimer-type dementia and Parkinson’s disease.” In: *Brain : a journal of neurology* 3 (1988), pp. 695–718.

[323] K. Datla et al. “Differences in dopaminergic neuroprotective effects of estrogen during estrous cycle”. In: *NeuroReport* 14.1 (2003), pp. 47–50.

[324] E. Anderson et al. “Performance of a motor task learned on levodopa deteriorates when subsequently practiced off”. In: *Movement Disorders* 29.1 (2014), pp. 54–60.

[325] A. Antonini et al. “The progression of non-motor symptoms in Parkinson’s disease and their contribution to motor disability and quality of life”. In: *Journal of Neurology* 259.12 (2012), pp. 2621–2631.

[326] T. Pigott. “A Review of Methods for Missing Data”. In: *Educational Research and Evaluation* 7.4 (2001), pp. 353–383.

[327] S. Xiang et al. “Bi-level multi-source learning for heterogeneous block-wise missing data”. In: *NeuroImage* 102 (2014), pp. 192–206.

[328] Q. Yang, S. Pan, and Q. Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359.

[329] A. Maas, A. Hannun, and A. Ng. “Rectifier Nonlinearities Improve Neural Network Acoustic Models”. In: *Proceedings of the 30th International Conference on Machine Learning* 28 (2013), p. 6.

[330] A. Supratak et al. “DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.11 (2017), pp. 1998–2008.

[331] C. Szegedy et al. “Going Deeper with Convolutions”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 7 (2014), pp. 1–9.

[332] K. He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

[333] L. Kuncheva. “A theoretical study on six classifier fusion strategies”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.2 (2002), pp. 281–286.

[334] D. Ruta and B. Gabrys. “Classifier selection for majority voting”. In: *Information Fusion* 6.1 (2005), pp. 63–81.

[335] L. Kuncheva, J. Bezdek, and R. Duin. “Decision templates for multiple classifier fusion”. In: *Pattern Recognition* 34.2 (2001), pp. 299–314.

[336] J. Sammon. “A Nonlinear Mapping for Data Structure Analysis”. In: *IEEE Transactions on Computers* C-18.5 (1969), pp. 401–409.

[337] N. Simon et al. “A sparse-group lasso”. In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 231–245.

[338] R. Little and D. Rublin. “Statistical Analysis with Missing Data”. In: *Wiley, New York.* (1987), p. 381.

[339] M. Huisman. “Imputation of missing item responses: Some simple techniques”. In: *Quality and Quantity* 34.4 (2000), pp. 331–351.

[340] G. Ambler, R. Omar, and P. Royston. “A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome”. In: *Statistical Methods in Medical Research* 16.3 (2007), pp. 277–298.

[341] E. Acuña and C. Rodriguez. “The Treatment of Missing Values and its Effect on Classifier Accuracy”. In: *Classification, Clustering, and Data Mining Applications* 1995 (2004), pp. 639–647.

[342] C. Musil et al. “Techniques for Handling Missing Data”. In: *Western Journal of Nursing Research* 24.7 (2002), pp. 815–829.

[343] J. Schafer and J. Graham. “Missing data: Our view of the state of the art”. In: *Psychological Methods* 7.2 (2002), pp. 147–177.

[344] K. Thung and C. Wee. “A brief review on multi-task learning”. In: *Multimedia Tools and Applications* (2018), pp. 1–21.

[345] P. Allison. “Multiple Imputation for Missing Data: A Cautionary Tale”. In: *Sociological methods & research* 28.3 (2000), pp. 301–309.

[346] B. Efron. “Missing Data, Imputation, and the Bootstrap”. In: *Journal of the American Statistical Association* 89.426 (1994), pp. 463–475.

[347] S. Xiang et al. “Multi-Source Learning with Block-wise Missing Data for Alzheimer’s Disease Prediction”. In: (), pp. 185–193.

[348] D. Bennett. “How can I deal with missing data in my study?” In: *Australian and New Zealand journal of public health* 25.5 (2001), pp. 464–469.

[349] P. Taylor and R. Little. “A Test of Missing Completely at Random for Multivariate Data with Missing Values”. In: *Journal of the American statistical Association* 83.404 (1988), pp. 1198–1202.

[350] E. Silva-Ramírez et al. “Missing value imputation on missing completely at random data using multilayer perceptrons”. In: *Neural Networks* 24.1 (2011), pp. 121–129.

[351] K. Sainani. “Dealing With Missing Data”. In: *PMR* 7.9 (2015), pp. 990–994.

[352] P. Allison. “Missing Data”. In: *Quantitative Applications in the Social Sciences* (2001), p. 104.

[353] C. Williams and C. Nash. “Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case”. In: 2015 (2018), pp. 1–6.