UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Estimating Parkinson's Disease Stages from Accelerometer Data

Vitor Marques Lobo

Mestrado em Engenharia Informática
Especialização em Engenharia de Software

Dissertação orientada por:
Prof. Doutor Tiago João Vieira Guerreiro

2022

Acknowledgments

First I would like to thank my parents and my sister, whose unwavering support and reassurance was fundamental towards the completion of this work.

I would also like to express my gratitude towards my advisor Tiago Guerreiro for all his teachings and opportunities throughout this journey. His patience in clarifying all of my doubts, sometimes repeatedly, and availability for numerous meetings and questions were instrumental for the development of this project. In this manner, I also thank Diogo Branco, who went beyond any responsibility to help me, and whose readiness to help and provide feedback was extremely important in directing this work.

Finally, and as importantly as those above, I would like to thank all of my friends and family. Your continued support and encouragement were extremely important for me, and without them, I can confidently say I wouldn't have gotten this far. Through this unexpectedly long period of my life, through all of the highs and lows, I am happy to have shared it all with you.

Resumo

A doença de Parkinson é uma doença neurodegenerativa que afeta o sistema nervoso central, e é mais prevalente em homens adultos, raramente ocorrendo em idades inferiores aos 50 anos. Apesar da causa desta doença ser desconhecida, a continua investigação sobre os seus mecanismos tem permitido algum progresso no tratamento dos sintomas, e levado à identificação de diversos fatores ambientais e genéticos de risco aumentado. Sabemos agora que esta doença causa a degeneração progressiva de neurónios específicos responsáveis pela regulação da dopamina no cérebro, um neurotransmissor importante envolvido em funções motoras e de memória. Este processo degenerativo leva á emergência dos sintomas tradicionalmente associados à doença de Parkinson. O acrónimo TRAP é uma abreviatura usada recorrentemente na literatura para descrever os sintomas: Tremor em repouso, Rigidez, Akinesia ou lentidão de movimentos, também designada Bradykinesia, e instabilidade Postural. Apesar de estes serem os sintomas mais frequentemente associados à doença de Parkinson, outros sintomas menos conhecidos incluem o 'freezing' of gait, que consiste em pequenas hesitações durante a fase inicial ou o decorrer da marcha, e uma série de sintomas neuropsiquiátricos que incluem depressão, ansiedade e demência, e ainda outros sintomas não motores tais como distúrbios de sono e gastrointestinais. Apesar de não existir uma cura para esta doença, existem tratamentos recomendados para a amenização dos sintomas motores. Entre estes, o mais comum é o uso de medicamentos à base de levodopa, que permitem o alívio rápido do tremor e rigidez. No entanto, o uso destes tratamentos a longo prazo, pode levar a outras complicações motoras, caracterizadas pelo aparecimento de movimentos bruscos involuntários (dyskinesia). A intensidade destas complicações está relacionada com o desgaste da medicação após o consumo, levando a flutuações ao longo do dia que foram categorizadas como estados 'ON' e 'OFF', descrevendo se a medicação está a ter efeito ou já desgastou respetivamente. Atendendo a estas flutuações, e à natureza degenerativa da doença que leva a uma progressão de sintomas ao longo do tempo, a monitorização constante do estado da doença é extremamente importante para informar a acão clínica, e permitir ajustes contínuos às abordagens para tratamento.

Atualmente, este acompanhamento é maioritariamente baseado em avaliações clínicas periódicas durante as quais os profissionais de saúde usam diversos exercícios, escalas e

questionários para aferir o estadio da doença e quantificar os diversos sintomas motores e não motores. Entre estes, a Escala Universal para a Doença de Parkinson, especificamente a versão revista pela Sociedade de Distúrbios do Movimento (MDS-UPDRS), é a abordagem mais usada para a avaliação de sintomas. Esta escala encontra se dividida em 4 partes, sendo que as partes 1 e 2 constituídas por 13 itens cada abordam o impacto dos sintomas na vida diária dos pacientes, enquanto as últimas duas partes, constituídas por 33 itens cada, focam-se na quantificação dos sintomas motores, requerendo o desempenho de vários exercícios por parte do paciente sobre observação clínica. A avaliação clínica através deste questionário é ainda frequentemente acompanhada de outras escalas como a 'Unified Dyskinesia Rating Scale' ou a 'Non-Motor symptom Scale'. Apesar de valiosas para a monitorização e acompanhamento da doença de Parkinson, estas escalas partilham alguns defeitos inerentes a este tipo de avaliação. Para além da subjetividade destas avaliações, que leva a variabilidade entre avaliação do mesmo paciente por profissionais de saúde diferentes, a maior desvantagem do uso destas escalas é o requerimento de deslocações por parte dos pacientes a clínicas ou hospitais para este efeito, que atendendo às complicações motoras causadas pela doença representam um fardo para o paciente e cuidadores. Adicionalmente, estas avaliações periódicas não capturam as flutuações de sintomas ao entre períodos de avaliação, informação clinicamente relevantes que muitas vezes acaba por ser transmitida apenas através de relatos dos pacientes.

Ao longo das últimas décadas, o aparecimento de sensores inerciais pouco intrusivos e de reduzido custo, representou uma mudança de paradigma na literatura relacionada com a monitorização da doença de Parkinson. Surgiram múltiplos estudos e trabalhos relacionados com a quantificação dos diversos sintomas motores, como uma maneira de colmatar os defeitos das avaliações tradicionais para este efeito. Assim, o uso destes sensores vestíveis popularizou-se para a classificação de estados 'ON' e 'OFF', a quantificação de tremores de repouso e rigidez, a deteção de episódios de 'freezing' durante a marcha entre outras abordagens para a monitorização da doença através dos seus sintomas. Para além destas abordagens, o estudo de algoritmos de aprendizagem automática para a quantificação e monitorização da doença também produziu resultados promissores, tendo sido estudados todo o tipo de modelos e variáveis para a estimação automática de, por exemplo, itens das diferentes escalas para a avaliação da doença de Parkinson, ou métricas que com estas se correlacionam. Ainda assim, e apesar dos resultados promissores, os métodos tradicionais continuam a ser favorecidos para a avaliação clínica devido a uma falta de consenso e foco nas metodologias para a recolha e tratamento de dados, que levam a que muitas destas abordagens necessitem de validação adicional, ou não sejam usáveis para uso fora do laboratório, no 'mundo real'. Recentemente, o foco na marcha e derivação de suas características através deste tipo de sensores demonstrou que estas propriedades da marcha, como a velocidade, ritmo e variabilidade, têm potencial como marcadores relacionados com a progressão da doença de Parkinson.

Especificamente, vários autores demonstraram não só a possibilidade de extrair estas e outras características como o número e largura de passos, mas também a correlação significativa entre a inibição destas características e estadios mais avançados da doença. Estes factores combinados com a relativa simplicidade de identificar períodos de marcha automaticamente representam uma oportunidade para uma monitorização verdadeiramente continua e objectiva da doença de Parkinson.

O estado da arte da literatura relacionada com este tópico começou recentemente a usar técnicas de 'deep learning' para este efeito, prevendo automaticamente a pontuação agregada da parte 3 do MDS-UPDRS para monitorização da progressão da doença usando dados recolhidos durante a marcha, e outras atividades. Os autores destes estudos, demonstraram uma metodologia para a predição desta métrica com significado clínico estabelecido, que têm a vantagem de ser facilmente interpretáveis por profissionais de saúde e pacientes, e com margens de erro relativamente baixas. No entanto, e como é recorrente na literatura relacionada, estas abordagens diferem significativamente em metodologia em termos de protocolos de recolha e processamento de dados, assim como nos modelos usados. Isto leva uma vez mais à necessidade de validação destes resultados, e à oportunidade de usar diferentes abordagens e modelos para o mesmo efeito, de modo a perceber o impacto das diferentes abordagens e aferir a possibilidade de melhorar os resultados.

Este trabalho pretende testar diversas variáveis relativas à recolha e processamento de dados para a monitorização da progressão da doença através desta métrica da parte 3 do MDS-UPDRS. Para esse efeito, foram testados 4 modelos de aprendizagem usando dados de sensores montados nas costas e no punho de 74 pacientes, e diferentes conjuntos e de features e métodos para a sua extracção do sinal recolhido por estes acelerómetros. Após uma descrição da literatura relacionada e dos dados recolhidos, este trabalho apresenta uma comparação dos modelos testados com o atual estado da arte, e uma breve discussão do efeito das diferentes variáveis no resultado. Apesar de os modelos testados não terem alcançado a performance atingida pelos modelos de 'deep learning' usados no estado da arte, os resultados demonstram que estes modelos são viáveis para a estimação desta métrica, existindo algumas possibilidades para a melhoria do seu desempenho. Para além disso, a discussão do efeito de cada variável testada resultou na identificação de um conjunto de possibilidades para trabalho futuro, que constituem trabalho importante para a monitorização objetiva da doença, e devem ser testadas antes de descartar estes modelos para o estadiamento objetivo da doença de Parkinson.

Palavras-chave: Doença de Parkinson; Marcha; Machine Learning

Abstract

Current practices for monitoring disease stage and progression in Parkinson's Disease still rely on periodic clinical visitations that can be cumbersome and highly stressful for patients and caretakers. Furthermore, the current gold standard for these evaluations is still reliant on observations made by trained clinicians using clinical scales that, in spite of their repeatedly verified validity, are subject to fluctuations due to intra or inter-rater variability. Over the last decade, technological developments in sensors for data collection and data science algorithms have enabled systems and tools for health and tele-medicine applications, along with a battery of research into the objective and continuous monitoring of Parkinson's Disease. Among such research, gait and it's characteristics have emerged as reliable markers for the progression of PD. As such, studies leveraging these characteristics for the objective monitoring of the disease have become a common trend in the related literature. A limiting factor in several of these studies is the use of scores and outcomes that, in spite of their high correlation to established clinical scales, are different to those usually used by clinicians, making them harder to interpret and adopt for clinical use. To bridge this gap, several studies have attempted to classify or estimate specific parts of the MDS-UPDRS, the most clinically used scale for the assessment of PD. Recently, the automatic estimation of scores for part 3 of the MDS-UPDRS, which focuses the severity of motor symptoms, have leveraged the use of deep learning techniques for this purpose, with promising results. This work presents a comparison of traditional feature engineered models against those current state of the art deep learning approaches for the prediction of this score. Furthermore, an analysis on different approaches to data collection, feature extraction and model parametrization was also performed, to assess the effect of these different variables on the estimation task. Finally, an analysis of the best configurations of machine learning pipelines for this purpose was also performed to direct further studies. While the optimal models in this study failed to match the performance of those state of the art approaches, the identified limitations and recommendations for future research form a solid foundation for future work on this topic.

Keywords: Parkinson's Disease; Gait; Machine Learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Parkinson's Disease (PD) is a neurodegenerative disease of the central nervous system. The cause of PD itself remains uncertain, but genetic factors like the presence of specific mutations, and environmental factors such as exposure to toxins or heavy metals, have been linked to increased susceptibility. Its prevalence is also higher in older adults, rarely occurring earlier than the age of 50, and slightly increased for men. Despite its unknown cause, the effects of the disease on the nervous system have been largely studied. The degeneration of dopaminergic neurons results in a loss of dopamine, a neurotransmitter involved in functions like movement and memory. This process leads to the loss of a significant percentage of these specific neurons before the most notable symptoms start manifesting [1]. PD symptoms are usually split into motor and non-motor categories. Motor symptoms include the three cardinal symptoms: tremor, slow limb movement (bradykinesia), and postural or limb rigidity, along with gait impairments that mostly manifest as lower walking stability, symmetry, and episodes of freezing of gait (FoG) [2]. Besides these, non-motor and non-dopaminergic symptoms can manifest at different stages of the disease, ranging from neuropsychiatric symptoms like depression, anxiety, and dementia, to sleep or gastrointestinal disorders, among others [3]. Although a cure for PD is far from becoming a reality, a battery of pharmacological and surgical interventions have been developed throughout the years in an attempt to ameliorate its symptoms. Different therapies are usually used in conjunction to address the needs of each patient, given the stage of their disease. Of these, the most commonly prescribed drugs are levodopa based pharmaceuticals, given their capacity to rapidly alleviate tremor, akinesia, and rigidity. Even though this seems like a favorable outcome, long term levodopa treatments lead to additional motor complications, characterized by involuntary convoluted movements (dyskinesia). The varying intensity of this side effect, along with the medication's wear off, leads to fluctuations in motor function throughout the day. These have been labeled as 'ON' and 'OFF' stages, describing whether the medication has worn 'OFF' or is still in effect ('ON') [4], [5]. Given our current inability to cure PD, the rapid and objective assessment of symptoms to enable an informed, swift intervention becomes a critical task.

Traditionally, the evaluation process required a visit to the clinic, but the popularization and improvement of wearable devices have shifted the way parkinsonian symptoms are assessed. The ability to keep track of patients' condition while they perform their activities of daily living (ADL) paints a clearer picture of their health. While monitoring of all symptoms is an essential task, gait performance and specific gait-related disabilities have been strongly correlated with the progression of the disease [4]. Some of these approaches are further discussed in the following sections of this report, with an increased focus on the assessment of disease state based on data collection systems.

1.1 Motivation

Over the last decade, a lot of research has arised focusing on objective monitoring of PD symptom severity using inertial data. While some works address individual symptoms, like tremor [6] and bradykinesia [7], others have attempted more general approaches that assess global motor impairments in PwP [8] [9]. Even though a lot of different methods and techniques have been suggested, a truly remote and clinically compliant system for continuous monitoring of disease progression has yet to be widely accepted and deployed. Some of the limitations that have contributed to this standstill over time are:

- The estimation of 'mobility' or 'disease' scores that show some degree of correlation with clinically used scales, but lack validation and transparency in their methods.
- The use of data collection systems that are cumbersome for the patient, restricting at-home evaluation.
- The blackbox nature of some ML techniques more recently used for this purpose.
- A lack of consensus on data collection and signal processing methods for continuous symptom monitoring.

1.2 Research Objectives

While some of the limitations described in the previous section are difficult to overcome, previous work has shown that there are several open challenges that could aid research towards true remote monitoring of PD. As such, the following research objectives were set in order to bridge the gap towards remote, continuous monitoring of symptom severity in PD:

(i) Develop a reproducible pipeline for data collection, processing, and MDS-UPDRS III estimation, and compare its' results with current state of the art approaches

(ii) Compare different machine learning models and data collection and processing variables for the purpose of MDS-UPDRS III estimation

(iii) To assess the contribution and possible clinical significance of features extracted from inertial data for MDS-UPDRS III estimation and disease monitoring

1.3 Document Structure

This document is structured as follows:

- Chapter 2 – Concepts and Background
- Chapter 3 – Related Work
- Chapter 4 - Methods
- Chapter 5 - Results
- Chapter 6 - Discussion
- Chapter 7 - Conclusion

Chapter 2

Concepts and Background

This chapter is an introduction to Parkinson's Disease (PD) and the fundamental concepts of machine learning required for an understanding of current and developing methods for the continuous assessment of disease stage and monitoring of symptom fluctuations.

2.1 Fundamentals of Parkinson's Disease

Parkinson's Disease is a neurodegenerative disease that affects the central nervous system. Its prevalence is higher in adults older than 50, and slightly higher in men. While its cause remains uncertain, the pathophysiology of PD has been widely studied and linked to the continuous degeneration of dopaminergic neurons in the *substantia nigra*. This causes a lower uptake of dopamine in the *basal ganglia*, a group of structures that is responsible, in part, for the communication between the brain and muscles through the neurotransmitter dopamine. As these neurons decay and dopamine levels drop motor dysfunctions may start manifesting, leading to complex motor symptoms that are commonly associated with the disease. The most evident symptoms usually manifest long after this process starts, highlighting the importance of early-onset detection and monitoring in order to plan clinical interventions and minimize the impact on the patients' quality of life.

While they are more commonly associated with the disease, motor manifestations are not the only consequences of its progression. Nonmotor symptoms often emerge alongside motor symptoms and can vary widely, ranging from several forms of sleep disorders to other neuropsychiatric symptoms like depression or anxiety, and even constipation, sexual dysfunction, or pain. Although these often manifest alongside motor symptoms, they are most common in later disease stages and may not present themselves at all for some patients. This lead to a focus on motor symptoms for the purposes of initial diagnosis, namely the four cardinal symptoms often abreviated as TRAP: Tremor at rest, rigidyty, akinesia and postural instability. These and other common motor symptoms are described in the following paragraphs, along with some notes on their assessment and treatment.

Bradykinesia

Bradykinesia is one of the fundamental symptoms of PD and must be present alongside tremor or rigidity in order for a diagnosis to be considered. It refers to a slowness of movements that stems from difficulties in planning, initiating and executing motor tasks. In the early stages of the disease, bradykinesia will mostly manifest as lethargic movements and reduced reaction times, impairing fine motor tasks like writing or buttoning a shirt. Its effect on facial and neck muscles can also lead to loss of facial expression, drooling and impaired speech. As with other symptoms of PD bradykinesia can be affected by the patients emotional state meaning excited patients may be able to briefly move or react faster than normal. External stimuli have also been found to affect bradykinesia, a phenomenon that has been exploited by researchers through the use of audiovisual [10] and haptic [11] cues in an attempt to ameliorate gait impairments. Assessment of this symptom is usually done at the clinic, by asking patients to perform rapid and repetitive movements while observing their slowness and decrementing amplitude. [12]

Tremor

Tremor is one of the most widely recognizable symptoms of PD. It can manifest in different ways, and is broadly classified in 2 categories. The most prevalent type of tremor is resting tremor, specifically in the 4-6Hz frequency band, which occurs during periods of relaxation and is mainly present unilaterally in the patients' hands, although it can also manifest in the lips, jaw and legs. Inversely, action tremors are those that manifest during the patients' motor functions, whether while holding a pose or performing a specific task. In both cases, assessment is usually made during routine clinical visits through the clinicians observation of the patient's state during resting and standardized activities.

Rigidity

Rigidity manifests as an increased resistance during passive movement of the limbs and neck, usually associated with pain and the ”cogwheel” phenomenon. Rigidity associated shoulder pains are one of the most frequent manifestations for onset PD. It is usually assessed by clinicians by passively moving the patients' limbs.

Gait and Postural Instability

Postural instability is one of the later manifestations of PD. It is affected by other parkinsonian symptoms, and usually associated with gait instability and falls, although it is not the only factor for these side effects. Assessment is usually done through the pull test, in which clinicians perform a quick pull backwards on the patients shoulders while observing their reaction.

Freezing

Although it’s not the most common symptom in PD, freezing is considered one of the most disabling symptoms for patients. Freezing more commonly manifests in the legs during gait, but it may also affect the patients’ arms and eyelids. The more common episodes of freezing of gait (FoG) can manifest as hesitation during gait initiation, the characteristic shuffling walk, and sudden pauses during walks in specific situations, which can lead to falls. Freezing usually manifests in the later stages of the disease and is usually assessed at the clinical during standardized walk tests.

Treatment induced motor complications

While bradykinesia and other motor symptoms usually respond to levodopa based medication, long term usage of these drugs can lead to other motor complications. The fluctuations induced by this long term usage are described as ON and OFF states, describing periods where the medication is effectively ameliorating symptoms or failing to do so. This long term usage may also lead to involuntary abnormal movements, denominated as dyskinesia.[13] The assessment of these fluctuations to enable clinical intervention is a challenge due to the punctual nature of traditional clinical assessment, requiring other approaches like the use of patient diaries, which have their inherent flaws, to track fluctuations over long periods of time.

2.2 Clinical assessment of PD

Currently, the progression of many symptoms in PD is monitored through visual assessment during periodic clinical visits, usually guided by one of many clinical scales. These are based on patient filled questionnaires or clinician-led interviews and can address one or more symptoms with varying specificity. One such scale, the current hallmark for assessment in both clinical and academic settings, is the Unified Parkinson’s Disease Rating Scale. Specifically, the version revised by the Movement Disorder Society, which is now commonly referred to as the MDS-UPDRS. This version is split into four parts, each addressing different aspects of the disease. Parts 1 and 2 comprise 13 items each, addressing Non-Motor and Motor Experiences of Daily Living respectively. These give some insight into how the symptoms affect the patient during their ADL. On the other hand, Parts 3 and 4, respectively titled Motor Examination and Motor Complications, comprise 33 items each in an attempt to measure the severity of these symptoms. These parts of the scale are usually assessed through the performance and observation of three tasks: sit to stand (S2S) where the patient is asked to repeatedly stand up from a sittting position, leg agility where the patients repeatedly and alternatingly stomps their feet from a sitting position, and the gait task, which usually consists on up to three straight line 10 meter walks. The last item of Part 3 is the Hoehn and Yahr staging scale [14], which categorizes five stages of progression in PD, according to the progress of the motor disability. Assessments made using the MDS-UPDRS can also be complemented by the use of many other scales, like the Unified Dyskinesia Rating Scale, which focuses motor complications that arise from extended treatment, or the Non-Motor Symptom Scale, which comprises a 30 item interview to thoroughly evaluate NMS [15]. Although they're still valuable in improving the disease's evaluation, scales like these share some major problems with the MDS-UPDRS and traditional assessment in general.

Shortcomings in current evaluation methods can be attributed to different factors. For starters, the periodic nature of traditional approaches is a major barrier for the assessment of symptom fluctuations, and although the use of patient journals has been studied as a solution, the responsibility it puts on patients to accurately log and measure these events is a major fault. Other concerns relate to the subjective nature of visual assessment, which, even when performed by trained professionals, can fail in identifying subtle changes or lead to intra and inter-rater variability [16] [17]. The ongoing shift towards ubiquitous computing, has allowed researchers to address these problems, by developing applications for the continuous monitoring of symptoms, to extract objective and ecologically valid metrics. Specifically, the increasing availability of wearable sensors has a lot of potential to precisely measure motor features both in, and out of the lab, making them a valuable tool for the evaluation of symptom progression.

2.3 Machine Learning Fundamentals

Machine learning algorithms have become a valuable tool for scientists and clinicians for the assessment, detection, and monitoring of several diseases. From computer vision algorithms that can detect anomalies in medical imaging to detection systems for medically significant events like falls or worsening symptoms, these tools are now seeing widespread research and deployment towards a more objective and informed clinical monitoring. This section intends to describe some of the fundamentals of machine learning techniques as an introduction to the following chapters describing specific applications for the assessment and monitoring of PD.

2.3.1 Machine Learning Models

The amount of fields that have adopted and contributed to ML research has grown rapidly over the last decade. To keep this summary within the scope of the project, this subsection will mostly focus models and approaches that have been used in clinical research, with a focus on those that are more common in research relating to the objective monitoring of PD. These can be categorized according to the way they 'learn' from data and the tasks they were designed to accomplish.

Supervised Learning

Models that fall under this category are usually used in problems that require the estimation of a target value, from a vector of features that corresponds to a single observation of a large dataset. These can be further categorized according to the type of value that this target variable assumes. If the value to be estimated is continuous, like the price of a stock for example, they are referred to as regression problems whereas if it is binary or discrete, usually corresponding to classes or labels, it is considered a classification problem. In spite of this distinction, these models function largely the same. They learn from pairs of feature vectors and corresponding known target values, their performance is then gauged by estimating values for some left out samples for which the target value is also known, and if it falls within an acceptable scope for a given application they can be used to make predictions from new, unseen data. [18]

The Random Forest (RF) algorithm, originally described by Leo Breiman in 2001 [19], has become a popular option among models using supervised learning across several fields of research due to it's relative simplicity in training and tuning [20]. In it's simplest form, and as the name suggests, RF models are trained by randomly splitting the training data into several subsets, a process called Bootstrap Aggregation, and training decision trees using randomly selected features from each subset to make several predictions of the target variable. For classification problems, these predictions from each tree are then used in a majority vote to get the final prediction from the ensemble, while in regression problems each predicted value is averaged into a final result. In PD related research, RF have been tested for several different challenges like the detection of PD in onset patients through gait analysis[21], or the prediction of FoG events in affected patients[22].

Among supervised learning approaches, and besides tree based models, Support Vector Machines (SVM) are also commonly used in literature pertaining to the objective monitoring of PD. Originally described by 1991 by Cortes and Vapnik [23] for binary classification, SVM's are based on the concept of Support Vectors that describe an Hyperplane that linearly splits an n-dimensional space in two, each 'side' representing one class of the classification data. This model was then built upon by the original authors and others since, enabling the definition of non-linear boundaries through the use of different kernel functions, multi-class classification by splitting the task into several binary classification tasks, and even accommodating regression tasks by adjusting the error function that determines the optimal vectors[20]. In PD related research, SVMs have been extensively used, for example to discern parkinsonian tremor from essential tremors [24] and to quantify the severity of motor symptoms like tremor, bradykinesia, and dyskinesia [25].

Unsupervised Learning

Unsupervised learning models are those that require no labeled data. Instead of leveraging these labels or target values to make predictions, these models are usually used to find structure in large datasets, by clustering data points or finding associations between features that would be hard to identify otherwise [26]. As an example, while a supervised model trained to classify pictures of cats and dogs would require a labeled dataset consisting of several pictures and their corresponding class, an unsupervised model for the same purpose would be able to identify patterns in the data that can split it into two groups, without any notion of what they represent. While not as common as their supervised counterparts, such models have seen some use in PD related research. Clustering has been used to diagnose PD with great accuracy, by finding patterns in voice recordings of healthy subjects and PwPD [27]. Another study from 2017 used data from surveys on the severity of several motor and non motor symptoms of PD to discern several sub-types of the disease according to the expression and severity of these symptoms [28].

Reinforcement Learning

Reinforcement learning can be summarized as the training of autonomous agents to take actions, based on the context of their environment, in order to maximise a set reward or goal [29]. This type of model is often associated to it's use in the video games industry or the development of autonomous vehicles, however, there has been extensive research for their use in the medical field. A recent example is the work of Kim *et al.* who developed a pharmacological recommendation model for the treatment of symptoms experienced by PwPD at various stages and states of the disease, achieving a higher reduction in symptom severity when compared to the recommendations made by clinicians[30]. While it has proven useful in numerous scenarios, this category of ML models has a much narrower scope for application than those previously mentioned, and given the (relatively) small amount of research on the use of such models in the medical field, falls outside the scope of this brief introduction.

Neural Networks

Neural networks are a special case when it comes to machine learning models. They can be designed to perform all of the aforementioned tasks, however, they do so in a very different way. These models usually consists on a collection of 'neurons' which are linked by layers. Each 'neuron' of each layer is a unit that applies some function to the data that is received from other layers, and transmits the result of this operation to the following layer. The firs and last layers are known as input and output layers respectively, while all others in between are commonly known as the hidden layers. This is a very general definition, but it encompasses most of the architectures of NN that are currently used, the differences arise according to the desired application of each model. The type of functions performed by each 'neuron', the 'weight' of each connection and the amounts of layers in a model are all part of these differences, and their variations form what is commonly called the model's architecture. Given the versatility of these models, numerous architectures have emerged over the years to respond to the necessities of different fields of research. Recently, research using these models has produced novel methods for the objective monitoring of PD using established clinical scales. Specifically, the work of Rehman et al. describes the use of a Convolutional Neural Network for this purpose using motion data collected during gait [9], while Hssayeni et al. similarly predicted the same score using a different ensemble model composed of three Long Term Short Memory (LSTM) neural networks and trained on motion data collected during different tasks [8]. While the low level characteristics of the models used in these studies fall outside of the scope of this introduction, their contributions are further discussed in the 'Related Work' chapter.

2.3.2 Knowledge as Features

As the name suggests, in ML features are distinctive characteristics or aspects of data that encode information that is relevant for a given task. The process of identifying, extracting and selecting relevant features for any machine learning problem is known as feature engineering [31].

Feature engineering is one of the first steps when designing a machine learning pipeline. While this process is highly sensitive to the type of data and models being used, it usually starts with an exploratory analysis of the data relying on domain-specific knowledge, data processing techniques and intuition in order to extract and select variables that encode relevant information for the task at hand. While the extraction process is highly reliant on the type of data being used, feature selection methods have been the object of intense research over the years, as it is considered one of the most important steps for ML pipeline.

Feature Selection

Feature selection in machine learning pipelines is mostly used to decrease the redundancy, and thus size, of the feature space, and to increase it's relevance or descriptive power for a given task. These goals can be achieved through several approaches, the most basic of which is using domain-specific knowledge to select features that are known to be highly relevant to the task at hand. Returning to the previous examples, someone like a clinician could point to features like age, weight and family history as the most relevant for predicting incidence of a disease, which would allow for the remaining features to be discarded, improving model performance, and decreasing the computational power required to train it [32]. There are also algorithmic methods to achieve these goals which have been an intense area of research in the field, and can be grouped into three categories .

Filters

Filter based methods are those that reduce the feature space by ranking them according to some metric of relevance, and selecting a user specified number of the most relevant features. These methods can be useful regardless of the chosen model, and have the advantage of being significantly less computationally expensive than other methods, which makes them a popular option for use with larger datasets. The number of features to select is set by the user, which makes the testing of varying feature spaces an important step to optimize the results.

The metrics used for feature ranking often rely on some statistical or probabilistic features of the used features, like the correlation between features and target variables, or the mutual information of the feature set, however other more complex methods have become widely used for this purpose. Relief based methods for example, are built upon the work of Kira and Rendell, who suggested a feature ranking algorithm that attributes weights to each feature by iterating over all of the training instances in the dataset, and for a given feature, iteratively update it's weight according to the distance between the feature values of a given instance, and it's nearest neighbour instances that have the same class (nearest hit), and a different class (nearest miss) [33, 34]. As this description implies, the algorithm originally contemplated only binary classification, but it has since been extended by several authors to address this, and other issues, originating an entire family of feature selection algorithms, among which one of the most popular is RelieF, which finds *k* neighbours instead of only two in order to accomodate problems other than binary classification, and deal with noisy, incomplete datasets. The popularity of this family of algoriths is mostly based on their capabilities of detecting feature dependencies, while maintaining a relatively high performance for larger datasets and feature spaces [35]

Wrappers

Wrapper based feature selection leverages the results of any given model to build the feature space iteratively. These methods are often more exhaustive than filters, which leads to a massive and sometimes prohibitive increase in computational requirements, particularly for larger feature sets [32]. Popular examples that fall under this category are Sequential Feature Selection, which consists on iteratively adding or removing features from the feature space while observing the performance of models using these subsets, and Recursive Feature Elimination, which hinges on the same concept of iteratively testing feature subsets, but aims to increase the relevance of features being used instead of the model's performance. Another example of a wrapper method, which illustrates the worst case complexity of this approach, is Exhaustive Feature Selection, where all possible combinations of features are tested and compared to select the optimal subset.

Embedded methods

While wrappers and filters are considered additional steps to be used before model training, embedded methods are those performed by some specific ML models during the training phase. Some common examples are decision trees and random forests, which have the ability to determine feature importance based on criteria like the Gini gain or the model's accuracy.

Overall, all of these methods have been extensively used and compared, with the consensus among researchers being that each have their advantages and disadvantages, often requiring purpose specific testing in order to determine the best method(s) for any given application.

2.3.3 Evaluating Model Performance

The final step of most ML pipleines is the validation of the model's results. For supervised algorithms, this usually means using the trained models to evaluate some left out samples in order to assess their performance on unseen data.

Model performance is usually evaluated by computing a distance metric between actual and predicted values. In classification problems this process often starts by quantifying the number of true and false positives (TP/FP) and negatives (FP/FN) in order to compute the following metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

As for regression tasks, performance evaluation usually relies on simpler metrics like the Mean Absolute Error or Mean Squared Error [20]:

$$MAE = \sum_{i=1}^{D} |x_i - y_i|$$

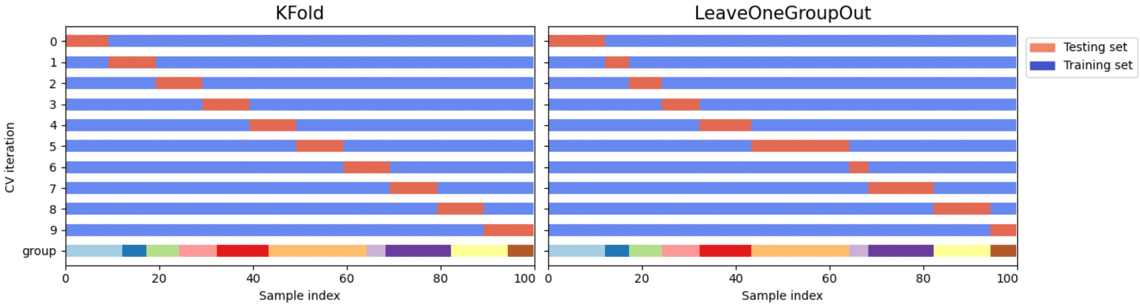

Figure 2.1: Illustration of data splits for two popular CV methods. The strategy on the left (K-Fold) is unaffected by the origin of data, while Leave one Group Out ('Group' is used instead of 'Subject' for abstraction) creates subsets of data that allow for subject-independent analysis. (Adapted from: `https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html`)

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2$$

Result Validation

As discussed, model performance is usually assessed using a set of samples that were unseen by the models during training, usually referred to as the testing set and created by splitting the dataset into two uneven sets. This simple method can adequately portray the model's performance, but in some cases, there may be a risk that the testing set is significantly different or too similar to the training set, leading to misleading performance levels that don't generalize well to novel data that the model will be used for. The aim of cross validation (CV) is to address this issue by splitting the data several times, performing the training and testing tasks using the different originating subsets, and computing an overall performance metric over all subsets. For this purpose, different strategies can be adopted for defining the way a dataset is split.

K-Fold cross validation is a common method for cross validation, where the train/test split procedure is repeated a number (k) of time, and each split (fold) is used to evaluate the model's performance, resulting in the cross validated performance metrics.

Another method, particularly common in clinical applications, is Leave One Subject Out (LOSO) Cross Validation, wherein the models are tested with samples from each of the patients on the dataset, and trained on the remaining samples, which allows for a subject independent analysis of model performance.[36]

Chapter 3

Related Work

The widespread availability of affordable, inertial measuring wearable devices has allowed researchers to study the relationship between collected motion data, and the presence or status of several motor manifesting diseases. For PD specifically, it has enabled research on the early diagnosis of the disease, objective monitoring of symptom progression and response to medication, and continuous remote assessment of symptom fluctuations.

3.1 Objective Monitoring of PD using inertial sensors

Data driven approaches for monitoring motor manifestations in PD have seen great developments over the last decade. Although different types of data like voice recordings [37] and typing patterns [38] have been studied for monitoring and detection of PD, motion data specifically has emerged as the predominant option in related research, mainly due to the wide availability, low cost, and versatility of data collected using body worn monitors (BWM) that measure acceleration and/or angular velocity. These devices can passively collect data during the patients' clinical evaluations or activities of daily living (ADL) on different parts of the body, enabling the extraction of a wide array of metrics and features of movement that can correlate with motor symptoms in PD, or upon processing, clinical scales used for the assessment of the disease.

The DataPark project is one example of these emerging applications. The web platform combines motion data collected with triaxial accelerometers at home or during clinical evaluations with clinical results and annotations about the patients' condition to generate personalized reports and visualizations of relevant clinical metrics like physical activity or energy expenditure levels. Preliminary results revealed that clinicians benefited from using the system, reporting a better understanding of the patients condition, and that the patients themselves were more engaged with therapy when faced with these reports for awareness and discussion of their condition [39].

Another widely deployed system that has arised from such research is the Parkinson's

Kinetigraph (PKG). The PKG is a wrist worn accelerometer that uses proprietary algorithms to produce bradykinesia and dyskinesia scores [7], estimate time spent with tremor [6], and extract other clinically significant metrics. Bradykinesia and dyskinesia scores (BKS/DKS) showed high correlation with the UPDRS III (BKS: p¡0.0005—r=0.64) and the Abnormal Involuntary Movement Scale (DKS: $p<0.0001$—r=0.80), while the computed metric of percent time with tremor achieved 88.7% sensitivity and 89.5% selectivity for tremor detection. Despite the promising results and later validating studies that showed beneficial outcomes for PD monitoring using this system [40], difficulties in detecting symptoms in non-sensed limbs or finer tremors that manifest at the fingertips have been pointed as limitations of the PKG. Furthermore, a review of the use of technology in PD by the MDS also highlighted the lack of transparency and open analytical methods of such systems as an open challenge for further validation [41].

In order to address issues like sensor placement, several studies have suggested different setups for data collection. A 2019 study demonstrated the use of a wrist worn device with a finger mounted accelerometer for the classification of tremor related UPDRS items with 99.24% certainty given a margin of error of one point, highlighting the potential of this setup to assess finer resting tremors which often manifest at the fingertips[42]. Similarly, Manzanera *et al.* used wrist, finger and toe mounted gyroscopes to assess bradykinesia using SVM classifiers, estimating MDS-UPDRS scores for bradykinesia with errors below inter-rater variability [43]. While sensor placements in the upper limbs have demonstrated high potential for the assessment of symptoms like bradykinesia and tremor, data collected from the lower limbs can also provide information for the assessment of PD.

Although cumbersome and difficult to use in free living settings, the use of body sensor networks (BSN) can inform research on the contribution of various sensor placements for the assessment of different manifestations of PD. Parisi *et al.* developed a method using three IMU, two mounted on the patient's thighs and on in the chest, to automatically classify the leg agility, sit to stand and gait tasks of the MDS-UPDRS. For each of the tasks, spatio-temporal signal and derived features were extracted to train three classifiers in order to determine the optimal combination of sensor placement, extracted features, and used models. The k-Nearest Neighbors (kNN) classifier attained the best classification results for all tasks, with an accuracy of 43% for the LA and S2S tasks and 63% for the gait task. The authors note that despite the low accuracy, 94% of the cases presented a classification lesser than 1, which is similar to inter-rater variability and according to the authors ”accurate” enough to mimic classification by trained professionals [44]. A follow-up to this work was published focusing on characterization and automatic classification of the gait task of the MDS-UPDRS using the same data collection system. In this study, the authors estimated the same spatio-temporal gait parameters derived from heel strike (HS) and toe off (TO) events to use as features for the selected classifiers. Once again, the best performing classifier was a kNN model which yielded 53% accuracy, and an error lesser than 1 for 98% of the samples. In their analysis of the used features, the authors noted a strong correlation between spectral power and gait impairments, which could represent a good avenue to further extend the automatic assessment of PD from gait data [45].

While several methods for data collection and processing have been tested there is still a lack of consensus on the optimal practices for symptom assessment in PD. However, a common factor in recent research is the use of gait tasks for data collection and feature extraction to enable the objective monitoring of the disease. The next section highlights recent research using gait tasks for objective monitoring of PD, and the emergence of gait as a biomarker for PD progression.

3.2 Gait as a biomarker for PD

Research on objective monitoring of PD is naturally progressing towards the passive sensing of patients during their ADL in order to collect data that can provide information on the patient's state and severity of their symptoms. The monitoring of gait tasks has emerged as a valuable option for this purpose due to a combination of several factors. The possibility of automatically detecting gait instances, demonstrated by research on activity classification, offers a way to achieve this objective that is unobtrusive and completely passive, requiring no extra burden on the patient. Furthermore, gait disturbances are one of the most common manifestations of PD, and while they are more prevalent in later disease stages, studies have shown that subtle gait impairments can be observed from disease onset [46].

Over the years, several symptoms have been assessed through the use of gait tasks. Naturally, one of the most common goals when collecting data during gait has been the detection of FoG episodes. For this purpose, Tripotli *et al.* suggested a method using 6 IMU's mounted on the patients trunk, wrists, and thighs to collect data for this purpose. The authors computed the entropy of 1 second windows on each axis of each device to use as features for several different classifiers, in order to assess the effect of sensor placement on the classification task, and the optimal model for FoG detection. The Random Forest (RF) model achieved the highest accuracy for all sensor configurations, and the highest using data from all devices. Despite the promising results, several limitations of the study require further research, like the low sample size (6 PwPD and FoG), the amount of extracted features for classification, and the cumbersome data collection setup that would make it difficult do deploy this system in true free living conditions [47]. Recently, related research has focused on detecting not only FoG, but also the periods that precede it with the goal of developing systems for the automatic prevention of these episodes through the use of rhythmic audio or visual cues. A recent 2020 study instrumented patients with a single accelerometer mounted on the lower back to collect gait data that included FoG episodes, labeled by trained neurologists. The authors segmented the signal, and considered the segments preceding FoG episodes as pre-FoG segments for classification. For the detection of pre-FoG instances, four models were developed in order to maximize accuracy, and minimize latency for the classification task. The optimal model achieved 77.9% accuracy for detecting these pre-FoG periods, lower than the aformentioned study, but using a single sensor, which is promising for future research pertaining to FoG prevention in FL. Furthermore, the authors noted that the duration of the classified periods preceding FoG, correlated strongly with the disease stage and duration [48]. This correlation between gait characteristics and disease stage or progression, has been a widely discussed topic in a lot of research pertaining to gait in PD.

In 2015, a study on the characteristics of gait in PD produced a set of algorithms for the extraction of seventeen gait characteristics from older and young healthy adults using a single accelerometer mounted on the patient's lower back. Upon validation, the researchers found that the extracted metrics were in agreement with laboratory references, but noted the need for refinement of some of these algorithms for further applicability [49]. These refinements were included on a follow up study in the following year, validating fourteen core gait characteristics that were also viably measured in PwPD. These characteristics addressed step time, length and velocity, stance time and swing time, and while their mean value demonstrated excellent agreement with reference laboratory measures, asymmetry and variability metrics only achieved poor agreement. The authors note that this may be due to the intrinsic limitation of comparing these systems that measure different properties, and were confident that the use of a single BWM mounted at the lower back was a good option for measuring these metrics [50]. The authors also conducted a further study on the impact of PD and the environment where data was collected for the extraction of these metrics. For this purpose, they collected data using the same device and placement from healthy subjects and PwPD both in the lab, and in free living conditions over a period of 7 days. Their findings suggested that gait performance was worse for both groups in free living conditions, and that these exacerbated the difference between groups compared to data collected at the lab. Furthermore, the authors observed that the extracted gait characteristics changed with the length of the analysed walks, and that the difference between both groups was more noticeable in longer walks [4].

Research using these gait characteristics as a marker for PD has demonstrated potential for monitoring the disease in several ways. A 2019 study demonstrated their use for discriminating PwPD from healthy subject, using these characteristics and machine learning models to determine the optimal configuration. The random forest model achieved the highest classification accuracy, at 97,14% and the most relevant features for classification were the mean values for step length, width and velocity, and the variability of step width and length [51]. For this study however, the gait characteristics were extracted from data collected through instrumented floor mats, which are less error prone than those extracted from accelerometry data. Furthermore, only k-fold cross validation was performed, which may not be ideal to test the generalization power of the suggested models. Considering these limitations, the authors point towards the use of accelerometers for extraction of the gait characteristics and the validation of these results using external datasets as a possible avenue for further research. Another study, recently published by Branquinho *et al.*, describes the development and use of a custom built wearable for the collection, processing and storage of gait characteristics in real time, during gait. These data was then compared between healthy subjects and PwPD, finding significant discrepancies between groups for asymmetry, variability, rhythm and pace metrics, which is in agreement with the aforementioned study. Additionally, the authors explored the correlation between disease progression and quality of life, measured by MDS-UPDRS III scores and the PDQ-39 scale respectively, with the extracted characteristics. For both scales, the authors noted higher values in the asymmetry and variability domains, and reduced values in the pace domain, and significant correlations between disease progression and the computed metrics, once again supporting gait as a marker for disease progression. In closing, the authors point to the possibility of combining machine learning models with the developed method for the automatic estimation of disease progression [52].

While the use of these gait characteristics has become a popular approach for monitoring PD, novel research has started to analyze signal processing metrics that could also be of use for this purpose. In their 2019 paper, Rehman *et al.* analysed the contribution of signal based features and gait characteristics for the classification of PD. Accelerometry data was collected during an oval walk using a single accelerometer mounted in the lower back, and the segmentation of the collected signals was enabled by data collected from a GAITRite pressure sensing mat. 25 gait characteristics were computed using the GAITRite data, based on previous work already discussed in this section, and 185 accelerometry signal features were extracted from the spectral density, regularity, magnitude and complexity domains. Partial least square regression combined with discriminatory analysis was then used to develop several models to discern between PwPD and healthy control subjects. The findings revealed that the signal characteristics had better discriminatory power than the spatiotemporal gait characteristics, with accuracies of 87.32% and 70.42%. The change in these values when both feature sets were used for classification was negligible, but increased slightly with the addition of demographic related features [53]. This is in agreement with a longitudinal study published in the same year, where a similar albeit smaller set of gait and signal characteristics were analysed for their potential as markers for the progression of PD. The data for this study was similarly obtained from a single sensor mounted on the lower back during a circular gait task, although here, angular velocity was collected besides accelerometry data. By comparing gait and signal characteristics between patients and healthy subjects over a period of 5 years, the authors found that 5 of the 24 previously extracted characteristics were good markers for progression in early PD, and 3 others were best fit to monitor progression in later stages, with most of these 8 being signal features [54].

The findings in these studies support the use of gait data collected from accelerometers as a valid option for the monitoring of PD. Current state of the art research has focused on the use of gait to automatically stage PD, supported by the previously discussed literature on gait and signal characteristics as a biomarker for the disease. The following section contains a summary of relevant literature pertaining to the use of inertial data collected during gait for this purpose.

3.3 Estimating disease stage from motion data

Although not completely novel, research relating to the estimation of scores for clinically used scales has gained traction over the last few years as a way to monitor the progression of PD in a more comprehensive way. The use of established clinical scales opposed to research oriented metrics and scores lowers the barrier of entry for the use of technology in PD, making it easier for clinicians and patients to understand and use the outcomes of such systems for clinical intervention and monitoring.

The Hoehn and Yahr (HY) scale is used during regular MDS-UPDRS assessments to stage the functional disability associated with PD. In a recent study, Mirelman *et al* discussed the possibility and contribution of several mobility features extracted form inertial data during different gait tasks to discriminate a healthy cohort (HC) from PwPD, and different stages of PD as measured by this scale. The authors compared the contribution of sensor location, task complexity, and extracted feature domains in an attempt to determine the optimal configuration of these variables for monitoring of disease progression. Sensitivity and specificity of the classification task between healthy subjects and patients classified as HYI were the highest, at 83% and 80% respectively, and decreased significantly for later stages of the disease down to 74% and 69% for discriminating between stages HYII and HY III. The findings reaffirmed the potential of more complex gait tasks to highlight subtle motor dysfunctions, and revealed that while asymmetry related features showed better discriminating power for earlier disease stages, the remaining feature domains were not stage specific. As for sensor location, upper limb and trunk mounted sensors performed better at discriminating earlier stages of the disease, but 40% of the most relevant features were collected from sensors in the lower limbs, suggesting that sensor placement should be adjusted to disease progression. Finally, the authors note that further research is needed to validate these findings, and assess the effect of symptom fluctuations and dominant side of symptom manifestations which were mostly unaccounted for [55].

Another emerging method to stage PD is the use of total scores of the entire MDS-

UPDRS or sub parts of the scale. Specifically, MDS-UPDRS III scores have been empirically demonstrated as a good metric for monitoring progression of PD [17]. As such, several studies have focused on the prediction of this score to monitor disease progression.

A recent example of this approach for the monitoring of PD progression is a 2021 study that leveraged a convolutional neural network (CNN) model trained using inertial data collected from the lower back during gait to estimate MDS-UPDRS III scores. The authors used the common sliding window method to segment these data into 5 second segments and compute signal vector magnitude to train the model and gridsearch to optimize its hyperparameters, achieving a clinically significant mean absolute error (MAE) of 6.29 and strong correlation (r=0.82) between estimated and actual scores. While these results are promising, the authors suggest that a comparison with traditional feature engineered machine learning models could be an avenue for future work, towards the deployment of such technologies for continuous monitoring of PD. Furthermore the longitudinal nature of this project that used data collected from PwPD over 3 years implies that the model may have learned some patient specific characteristics, which could harm generalization [9]. A similar study published in the same year used data collected simulated activities of daily living (ADL) from wrist and ankle worn inertial measuring units (IMU) to predict this score, addressing some of the previously mentioned limiatations. The authors proposed an ensemble of three deep learning models using hand crafted features, raw angular velocity signal, and time-frequency data for UPDRS-III estimation. This method resulted in an even lower MAE of 5.95 while maintaining strong correlation, even after Leave One Subject Out cross validation to ensure generalizability of the developed model. In order to compare the result of the proposed ensemble model with other common methods, the authors used the same data to train a gradient tree boosting model that performed significantly worse (MAE = 7.85) [8]. The used features and low sample size (n=24 PwPD) may have negatively effected these results, meaning that once again further work is required to validate the use of traditional feature engineered models for estimation of MDS-UPDRS III.

As discussed, the automatic staging of PD using established clinical scales is still an emerging area of research. The high variability of data collection and processing pipelines used for this purpose requires a comparison and validation of different variables in order to establish a set of optimal practices towards this goal. The next section discusses some open challenges for the objective monitoring of PD, along with possible approaches towards convergence on established and validated methods for this purpose.

3.4 Discussion

The previous section has laid out the current state of the art on the objective monitoring of PD, including some of it's limitations, emerging approaches, and future research directions.

One of the most recurrently mentioned limitation in this field of research, is the lack of consensus on data processing and machine learning pipelines for monitoring the disease's progression. While this has been a limiting factor for some time, a common factor has emerged for the assessment and monitoring of PD's motor manifestations. The use of motion data collected during gait has gained popularity and been increasingly used by research in this field. This convergence can be attributed to several factors, the most important of which are the established relationship between disease progression and gait impairment, and the relative ease of detecting gait in free-living for automatic and remote monitoring of the disease [4][51].

The processing and extraction of motion features from the wide array of data collection devices employed in research pertaining to continuous disease monitoring has also been the aim of extensive research. For this purpose, the most popular approach in recent research has been the collection of accelerometry and angular velocity data from wearable devices mounted in varying positions, favouring unobtrusive data collection systems that have potential for use in free-living conditions. While the extraction of gait characteristics from such data has been studied and successfully employed for some studies, recent efforts have shifted to the use of instrumented mats for the extraction of these characteristics as a way to avoid error propagation in machine learning pipelines, and validate the characteristics extracted from inertial data [50] [53]. Conversely, recent studies have posed the option of using features relating to the spectral, harmonic and magnitude characteristics of collected signals to find correlations with disease stage. Most recently, research using raw data to feed complex neural network based algorithms for disease monitoring has become a common approach for this purpose, with some authors pointing to the need for comparison between these, and traditional feature engineered machine learning models [9].

As for the outcomes of these objective monitoring systems, the use of several research specific scores and scales for this purpose has been extensively explored. However in recent years, there has been a convergence towards the use of clinically established scales or their subparts for direct disease stage estimation, leveraging the popularity and widespread knowledge of these metrics among clinicians and patients dealing with PD to produce results that directly relate to clinical practice and dismiss additional information or training. [17][8][53].

In summary, past research has revealed that it is possible to estimate PD progression using gait data collected with accelerometers. However, the relative efficacy and effect of different approaches to data collection and processing, and machine learning pipeline design still lack consensus and clear comparisons that could help inform future research in this field. Towards this goal, this project aims to compare the effect of some of these options, and establish some baseline data processing and machine learning practices for the estimation of disease progression using techniques based on the current literature, or approaches suggested as future research directions in this field.

Chapter 4

Methods

The MDS-UPDRS III estimation task was performed using different approaches to data collection, signal processing and using different machine learning pipelines. This chapter describes which steps were taken towards this task along with the several variables for each step, in order to enable a comparison between different design decisions and their effect on the estimation of disease stage. An illustration of these steps and their sequence is also included in Figure 4.1 to aid readability.

4.1 Research Questions

Attending to the previously discussed open challenges in research pertaining to the automatic staging of PD, the current work aims to answer the following research questions:

- How do traditional feature engineered machine learning models compare with state of the art deep learning techniques for the estimation of disease stage in PD?
- How do factors like sliding window length, sensor placement, and model selection affect this estimation task?
- How do the used features contribute for the estimation task?

For this purpose, a set of techniques and machine learning models were selected from related literature for comparison. The following sections describe all the steps taken, from data collection to MDS-UPDRS III estimation in order to study and compare the effect of these different variables for the objective monitoring of PD.

4.2 Tools and Software

All data was collected using Axivity AX3 devices, a commercial version of the open source logging accelerometer developed by the OpenMovement project at Newcastle University. These sensors have been validated [56, 57] and extensively used for research relating to physical activity monitoring and motor impairment assessment. Fig. 4.2 displays

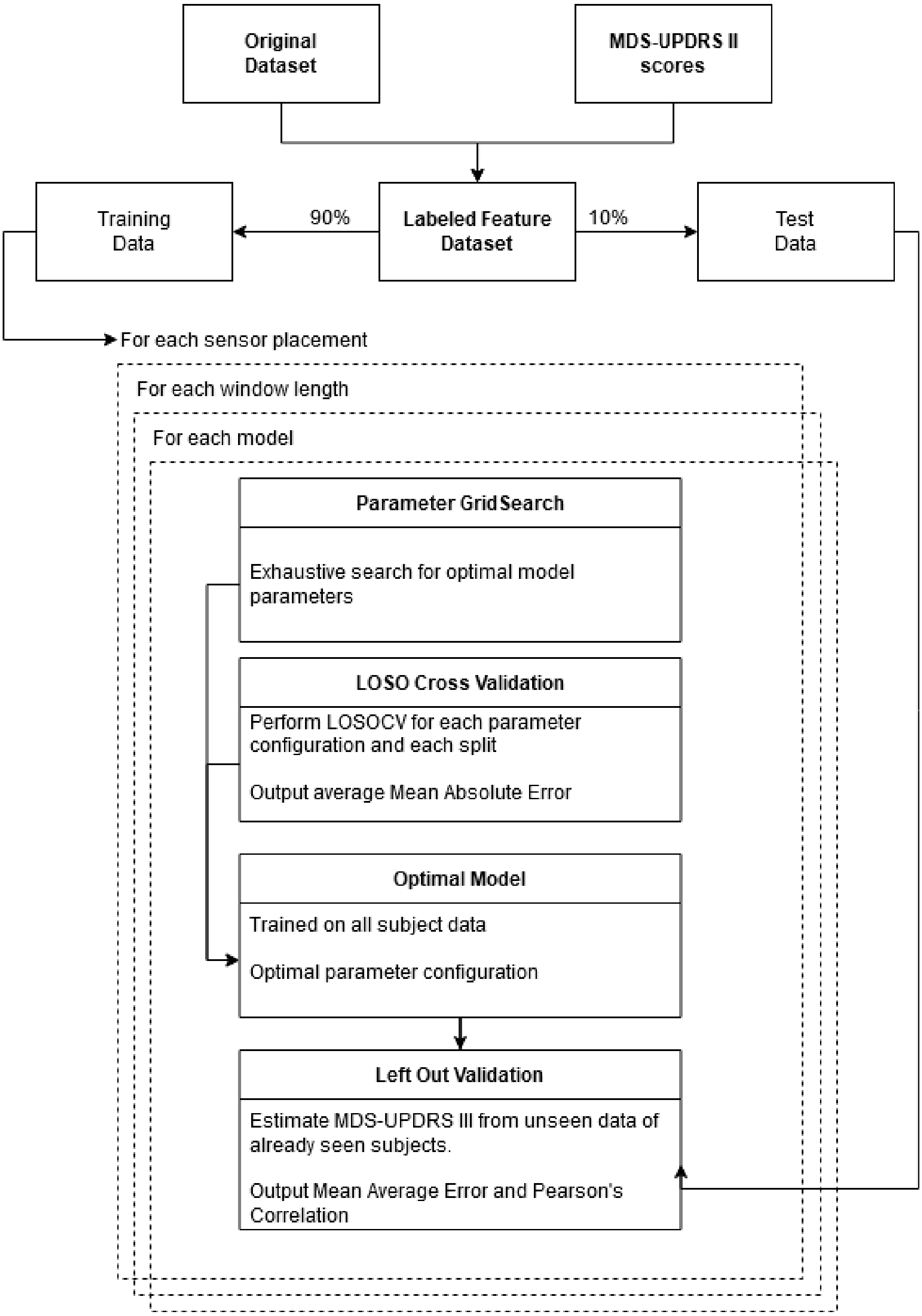
Original Dataset
MDS-UPDRS II scores
Training Data
90%
Labeled Feature Dataset
10%
Test Data
For each sensor placement
For each window length
For each model
Parameter Grid Search
Exhaustive search for optimal model parameters
LOSO Cross Validation
Perform LOSOCV for each parameter configuration and each split
Output average Mean Absolute Error
Optimal Model
Trained on all subject data
Optimal parameter configuration
Left Out Validation
Estimate MDS-UPDRS III from unseen data of already seen subjects.
Output Mean Average Error and Pearson's Correlation

Figure 4.1: Sequence of steps taken towards MDS-UPDRS III estimation.

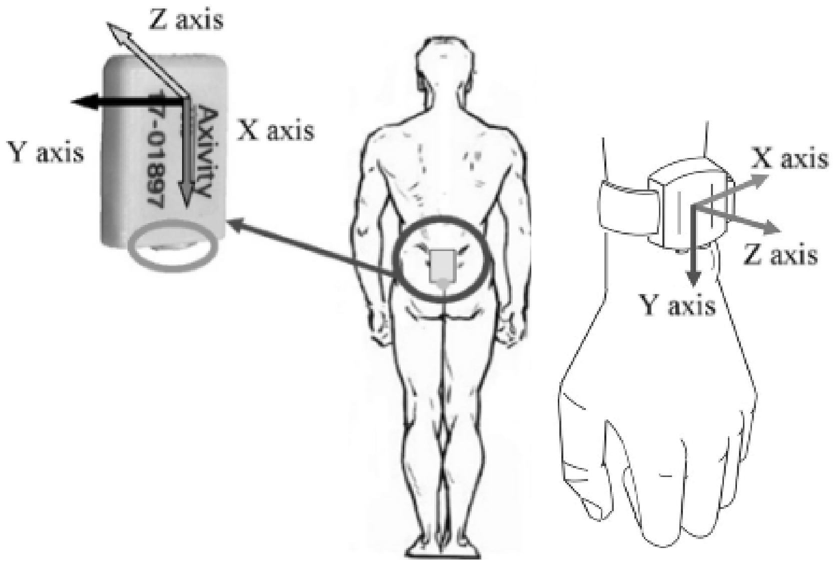

Figure 4.2: Device placement and orientation.

the mounting configuration and orientation of the devices used at CNS, which collected the data used in this study.

The collected data was extracted and converted from its original binary format using both the omconvert library and the Open Movement Graphical User Interface (OMGUI). A detailed description of the data collection process can be found in the next section. All data processing was done using python (version 3), with the aid of the following libraries:

- Time Series Feature Extraction Library (TSFEL) for Time and Frequency domain feature extraction [58]
- Sklearn for the built in machine learning models and cross validation methods [59]
- Numpy, Scipy for data processing
- XGboost for the implementation of the gradient boosted trees model [60]
- SKrebate for the implementation of the SURF and RelieF algorithms for feature selection [61]

All tasks related to training, cross validation and hyper-parameter space search were performed on a remote high performance cluster made available by the LASIGE research group at FCUL. The availiability of this high performance hardware was essential for the training of all models given the compounded computational cost of the grid search procedure and LOSO CV scheme. The gradient boosted trees model specifically saw the largest speedup through GPU acceleration which was not available for the implementation used for the remaining models.

4.3 Dataset Description

Data for this project was acquired from PD patients at the Campus Neurológico Sénior, a tertiary specialised movement disorders center in Portugal that employs the DataPark platform for subjective and objective data collection and visualisation. This section describes all the steps taken for data collection and processing towards the resulting consolidated dataset.

4.3.1 Data Collection

The dataset used in this project originated from routine periodic evaluations of patients at CNS conducted by trained physiotherapists. During these evaluations, patients are first asked about their current state and symptoms during a clinical interview. Then, depending on the patient's specific disease and conditions, therapists use one or multiple assessment tools like the MDS-UPDRS or the Mini-BESTest to understand how symptoms have progressed since the patient's last assessment and observe their general condition. Finally, the therapists may ask patient's to perform some additional standardized tests like the Timed Up and Go (TuG) test or the 10 meter walk test in order to observe the patient's performance and keep track of their condition. All of the information obtained from the clinical interview, along with the results for the clinical scales, timestamps for each activity and therapists' observations are then compiled using a custom built application for the DataPark platform, creating a session file that is appended to the patient's digital backlog. Additionally, accelerometry data is recorded during each of these sessions using two AX3 accelerometers mounted at the patient's lower back and wrist, set to record inertial data at 100 Hz. The resulting data files are also uploaded to the DataPark platform to enable the automatic generation of useful clinical visualization and extraction of other relevant metrics to monitor the patient's symptom progression. [39]

Due to its longitudinal nature and scope, the DataPark platform has collected large amounts of data beyond the specific focus on gait in PD that was set for this project. As such, a preliminary step for data selection was necessary to exclude unwanted or unrelated data. The first step towards this data selection process was to exclude all patients who had no PD diagnosis. To enable the feature extraction process and following machine learning pipeline for MDS-UPRS III estimation, a complete MDS-UPDRS evaluation was required for inclusion, along with the completion of at least one of the various 10 meter walk exercises that are usually performed during the clinical evaluations. The remaining reports and motion data files were then processed and prepared for feature extraction.

4.3.2 Data Processing

This section describes the steps taken towards processing the raw motion data for later use in feature extraction. All steps are based on data processing techniques used in related

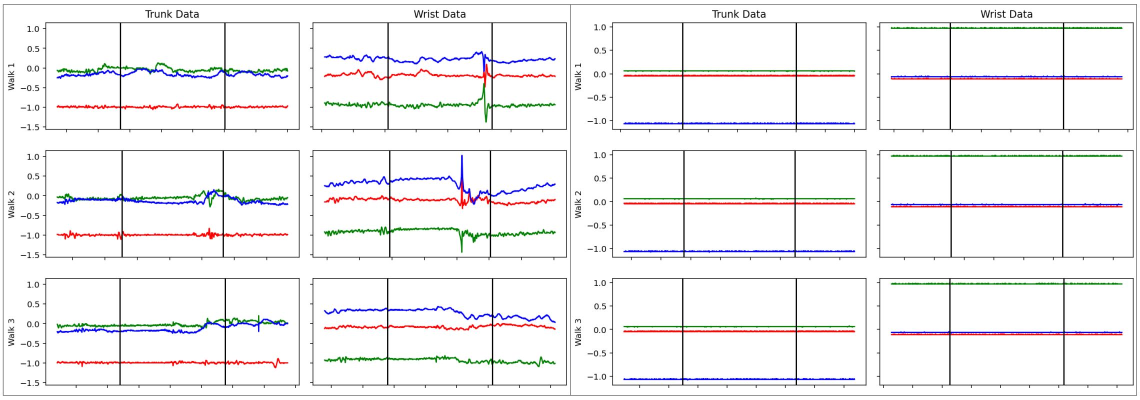

Figure 4.3: Example of discarded data.

works and other accelerometry reliant research.

In order to isolate gait instances that are the focus of this work, the selected data files were segmented using the annotated timestamps for the 10 meter walk exercises, which consisted of three separate 10 meter walks, and adjusted when needed to account for recordings made during daylight saving time. A visualization of each of the segmented gait instances was then created in order to exclude session data that contained sensor failures and misalignment, or mismatched timestamps, as demonstrated in Figures 4.3 and 4.4 comparing discarded and expected data respectively. During this step, the Vector Magnitude (VM) of the accelerometry signal was computed and appended to each segment using the traditional euclidean vector norm formula:

$$VM = \sqrt{x^2 + y^2 + z^2}$$

To avoid the possible temporal drift associated with the process, a resampling step was performed after segmentation to ensure even sampling, as required for the extraction of some of the used Time and Frequency domain features. Finally, all segments were filtered using a fourth order, digital low pass Butterworth filter with a cut-off frequency of 20 Hz in order to remove possible ”machine noise”. [62]

A final overview of the remaining data and it’s characteristics can be found in the next subsection.

4.3.3 Summary of Data

The final subset of data used to train and optimize the machine learning models contained 267 instances of gait from 104 evaluation sessions, collected from 74 different patients.

Among these patients, 49 were male and 23 were female, while the gender for the remaining 2 patients was not reported. The average patient age was 70.4 years with 13.12 standard deviation. The average weight was 71.76±13.89 Kg and the average height

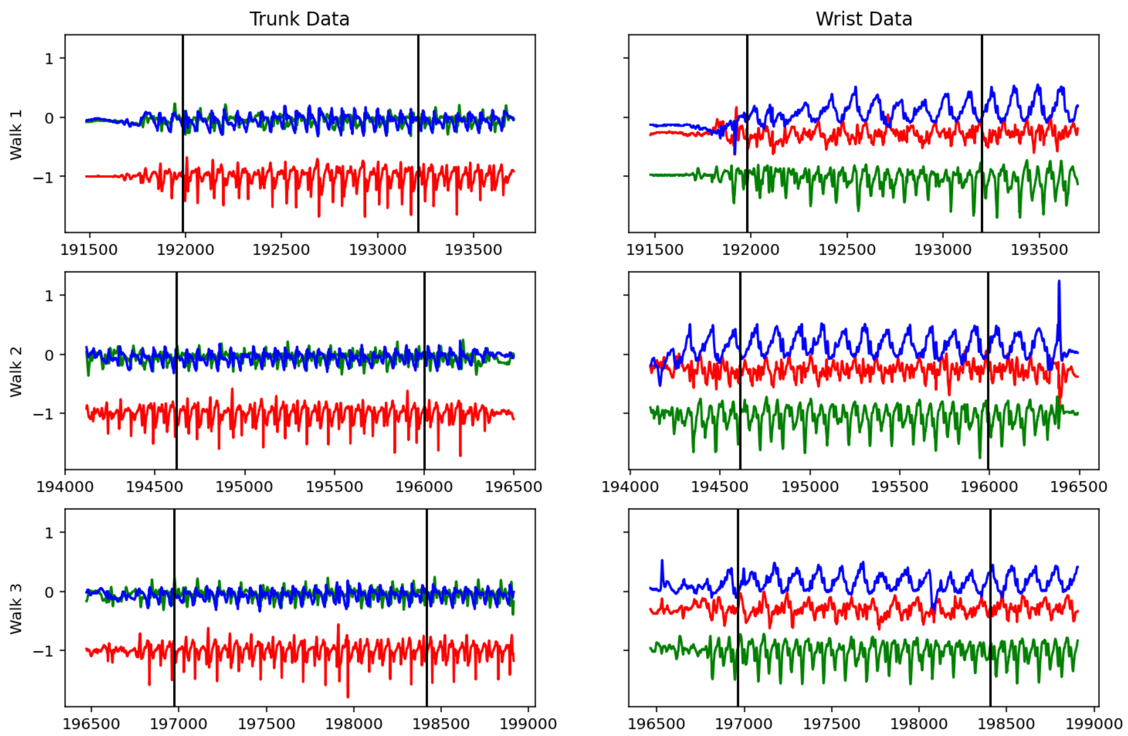

Figure 4.4: Examples of expected walk data.

was 166.49±9.26 cm. Finally, the average MDS-UPDRS III score was 40.92±14.31 and 2.57±0.97 for the H&Y scale. A visualization of the distribution of these scores can be found in Figures 4.5 and 4.6.

4.4 Machine Learning Pipeline

In order to optimize the estimation of MDS-UPDRS III scores a set of commonly used machine learning models and features were selected. This section describes the developed pipeline and accompanying design decisions towards the optimization of the disease stage estimation task.

4.4.1 Feature Extraction

As discussed, signal features have become a common option in research pertaining to the objective monitoring of PD [53, 63, 64]. For this reason, and to ensure a thorough analysis of the potential features for this purpose, a comprehensive set of features from the statistical, temporal and spectral domains were computed from all accelerometry axes and also from the VM. As is common in machine learning applications using time series data, a sliding window technique was used to segment the signal into non-overlapping windows from which the features were extracted. Different feature dataframes were then created using 2.5 and 5 second windows, both of which previously used in the literature

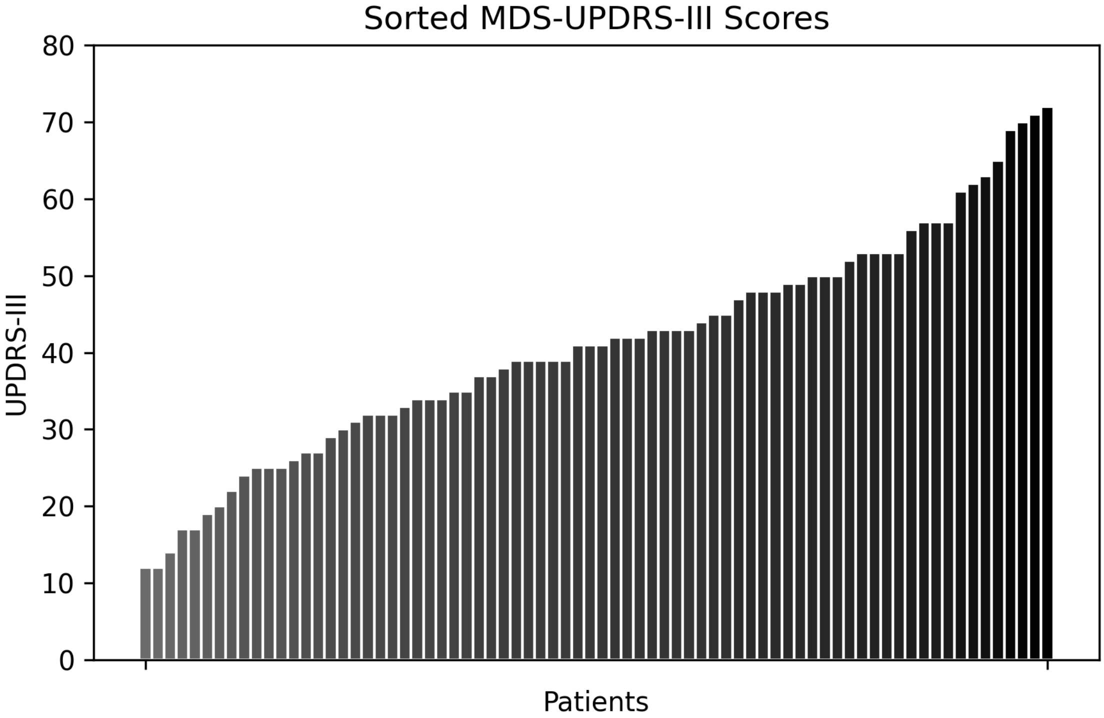

Figure 4.5: Per Patient UPDRS-III score distribution for patients in the dataset.

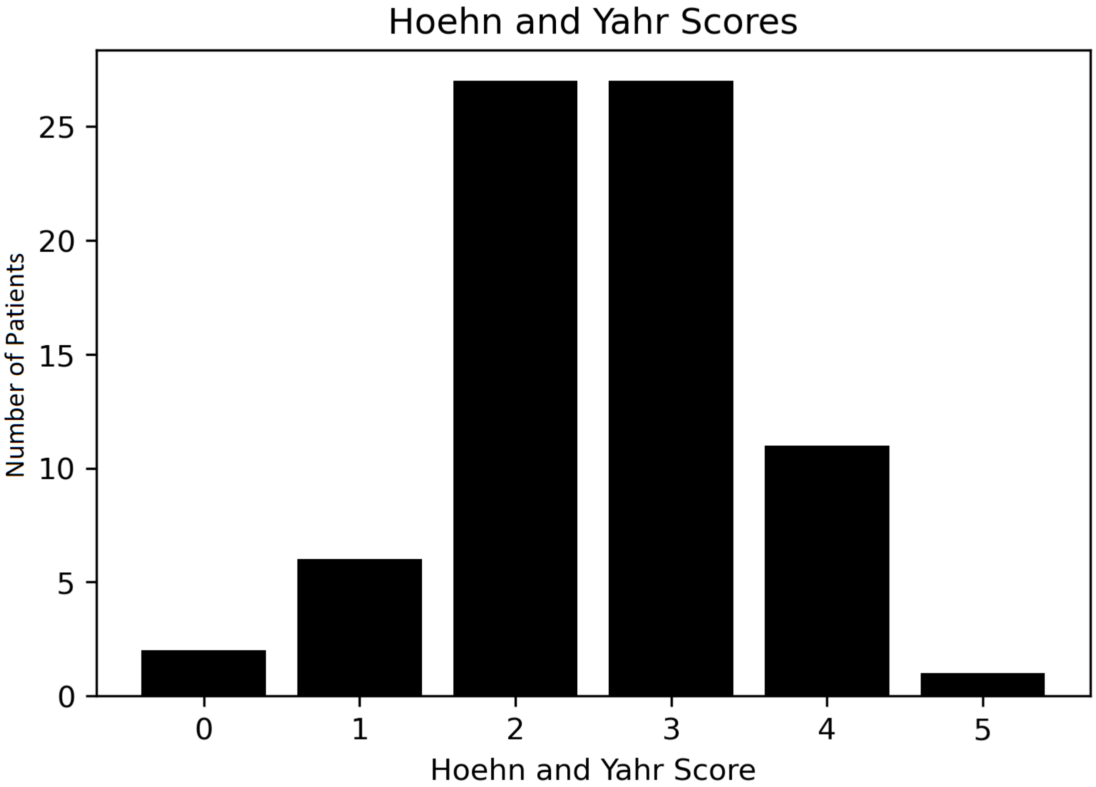

Figure 4.6: Hoehn and Yahr scale score distribution for patients in the dataset. The single patient in stage 5 met all inclusion criteria, despite requiring unilateral support.

Statistical Domain	Spectral Domain	Temporal Domain
ECDF	FFT mean coefficient	Absolute energy
ECDF Percentile	Fundamental frequency	Area under the curve
ECDF Percentile Count	Human range energy	Autocorrelation
Histogram	LPCC	Centroid
Interquartile range	MFCC	Entropy
Kurtosis	Max power spectrum	Mean absolute diff
Max	Maximum frequency	Mean diff
Mean	Median frequency	Median absolute diff
Mean absolute deviation	Power bandwidth	Median diff
Median	Spectral centroid	Negative turning points
Median absolute deviation	Spectral decrease	Peak to peak distance
Min	Spectral distance	Positive turning points
Root mean square	Spectral entropy	Signal distance
Skewness	Spectral kurtosis	Slope
Standard deviation	Spectral positive turning points	Sum absolute diff
Variance	Spectral roll-off	Total energy
	Spectral roll-on	Zero crossing rate
	Spectral skewness	Neighbourhood peaks
	Spectral slope	
	Spectral spread	
	Spectral variation	
	Wavelet absolute mean	
	Wavelet energy	
	Wavelet standard deviation	
	Wavelet entropy	
	Wavelet variance	

Table 4.1: Extracted features by domain

[9], in order to assess the effect of window size on the estimation task. During this feature extraction process MDS-UPDRS III scores were also computed and appended to the corresponding windows for both dataframes. An overview of the used features and their domains can be found in table 4.1.

4.4.2 Feature selection

Some of the selected machine learning models are sensible to feature sets that are redundant, poorly correlated with the target variable, or non-descriptive. While some of these models are less sensible to such problems given their capability to perform an intrinsic form of feature selection, all of them benefit from smaller feature sets to improve computation time. For this reason the first step towards feature selection was to use a variance filter to exclude features with low (<0.025%) or zero variance which lowered the feature space by up to 89% for the 2.5 second window. While this reduction may seem drastic, it is to be expected because of the way TSFEL works, computing the same feature several times for different frequencies for example which results in a large amount of feature columns with hardly any variability, and thus, descriptive power.

A further feature selection step was performed using four different feature selection methods that implement different strategies for feature ranking:

- f_regression which performs univariate linear regression tests returning F-statistic and p-values.
- mutual_info which measures the degree of dependency between the variables
- RelieF which is briefly described in the second chapter
- SURF which works similarly to RelieF but with automatic selection of the optimal number of neighbours

Each of these feature selection algorithms was used to rank and select the top 10/25/50 features to be used for the regression task using the linear regression algorithm, and with the support vector based model. The complete feature subset was also used for these models, in order to establish a baseline comparison with the remaining tree based models that are less affected by the number of feature due to their capability to perform intrinsic feature selection. A detailed list of the top ranking features for each of the best performing models accross both sliding window lengths can be found in the results section, in Tables 5.4 and 5.5.

4.4.3 Model Selection

The used machine learning models were selected attending to their prevalence in the literature relating to the objective assessment of PD using inertial data.

For each model, a set of parameters were selected and used in a grid search procedure to test all possible combinations. This procedure was then carried out for each sensor placement and the combined sensors, and for the different sliding window lengths used during feature extraction, in order to compare the effect of these variables for the estimation task. Leave One Subject Out (LOSO) cross validation was used during the grid search procedures in order to avoid overfitting and optimize the models for generizability. Finally, the optimal models for each combination of these variables were saved and used for the ensuing validation task. A list of the used models and corresponding parameter space can be found in Table 4.2.

4.4.4 Model scoring and validation

To validate the trained models, the original dataset was split into training and testing subsets. The training subset comprised 90% of the data and was used during the grid-

Model	GridSearch Parameters
Random Forest Regressor	{'criterion': ['mse','mae'], 'n_estimators': [250,500,1000], 'max_features':[0.333,0.666,1] }
XGBoost Regressor	{'learning_rate':[0.1], 'max_depth': [3,6,9], 'num_parallel_tree': [10,100,200], 'colsample_bynode': [0.333,0.666,1]}
SVM Regressor	{'kernel':['rbf','linear'], 'gamma':{'scale', 'auto'}, 'C':[0.1,1,10],'epsilon':[0.1,0.2,0.3]}
Linear Regression	NA

Table 4.2: Per model grid search parameter space

search procedure to train the models using LOSO cross validation. The remaining 10% of the data was then used as a validation set to test the model's performance on unseen data from patients whose data the model had already seen, providing information on the models' ability to estimate MDS-UPDRS III scores for patients that were already known to these models. These steps yield two different scores for each of the optimal models using the same Mean Absolute Error (MAE) evaluation metric: the average MAE for all LOSO splits during training, and one MAE score for each validation step. For the purpose of this study, this metric is defined as the mean absolute difference between real (y) and estimated ($\hat{y}$) MDS-UPDRS III scores over the amount of samples used for estimation:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| \,.$$

For the validation step using the held out data subset, the coefficient of determination (r^2) and Pearson's correlation (ρ), along with the corresponding p-value (p), for estimated and actual scores were also computed using the following formulae.

$$r^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}; \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$$

$$\rho = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 (y_i - \bar{y})^2}}; \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Chapter 5

Results

This chapter lays out the results from all of the steps taken towards UPDRS-III estimation, including data processing, feature extraction and selection, and finally model training and validation results. While some observations on these results may be made throughout, the discussion of their contribution for the continuous monitoring of PD and relevance for the defined research questions is left to the specific discussion chapter.

5.1 Optimal Configurations for UPDRS-III Estimation

The configuration with lowest prediction error on the left out 10% of data used data from both devices processed using a 2.5 second sliding window and a Random Forest model for prediction, achieving 4.26 MAE and strong correlation ($\rho = 0.93$) as illustrated in Figure 5.1. LOSO CV performance for this model was significantly lower, achieving a MAE of 11.50.

The best performing configuration when performing LOSO CV was a Support Vector based model, using data from both sensors but a 5 second feature extraction window, achieving a MAE of 9.99. While predictions using this model on the validation set were less accurate than some of the other options at 7.94 MAE, it maintained significant correlation on the left out set ($\rho = 0.63$) and achieved the best balance when considering both of the validation schemes.

Tables 5.1 and 5.2 summarize the optimal results achieved by each model along with the used data sources and sliding window length for the 10% left out and LOSO validation tasks respectively.

Model	**Device Placement**	**Win. Length**	**Ft. Selection**	**# Features**	**LOSO MAE**	**Test MAE**	**Test ρ (p)**
rf	combined	250	-	266	11.50	**4.26**	0.93(p<0.001)
xgboost	trunk	500	-	229	11.67	4.39	0.78(p<0.001)
svm	combined	500	SURF	25	9.99	7.95	0.63(p<0.001)
lin_reg	combined	500	reliefF	25	10.21	8.98	0.63(p<0.001)

Table 5.1: Optimal configurations used by each model to achieve optimal MAE on the left out 10% of data

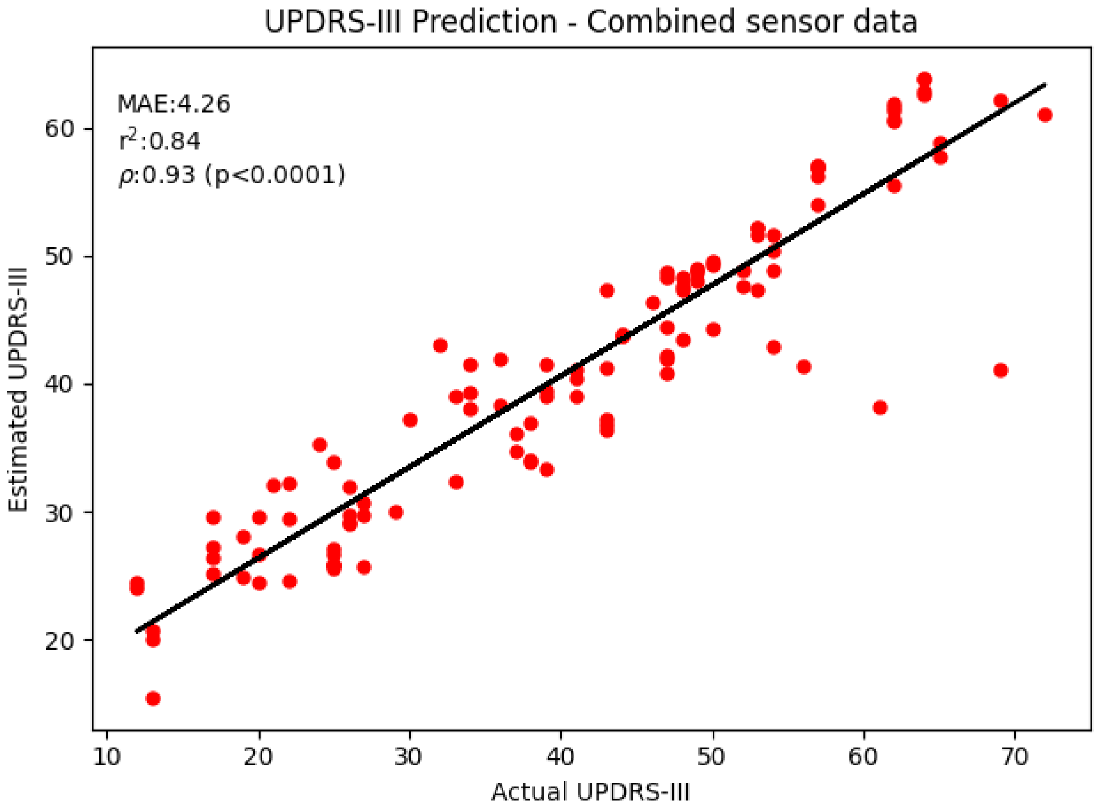

Figure 5.1: Overall optimal predictions on the 10% of left out data using a Random Forest model on data collected from both sensors and a 2.5s sliding window.

Model	Device Placement	Win. Length	Ft. Selection	# Features	LOSO MAE	Test MAE	Test ρ (p)
rf	combined	500	-	452	11.39	11.39	0.89(p<0.001)
xgboost	trunk	250	-	133	11.49	5.74	0.88(p<0.001)
svm	combined	500	SURF	25	**9.99**	7.95	0.63(p<0.001)
lin_reg	combined	500	reliefF	25	10.21	8.98	0.62(p<0.001)

Table 5.2: Optimal configurations used by each model to achieve optimal MAE during LOSO CV

5.1.1 Data Collection and Feature extraction variables

Both device placement and window length used during feature extraction had significant impact on the performance of all models. The following subsections contain a summary and, when relevant, illustrations of the effect of these variables on the prediction task.

Device Placement

For all of the selected models, the configurations that achieved the best results using either of the validation schemes used data collected from the lower back or both sensors combined.

Specifically, all of the non-tree based models performed better in both validation schemes using data from both sensors, with the exception of the SVM based model using a 2.5 second window, which compared to the other options using the same window length achieved lower, albeit negligible, validation MAE using data from the wrist.

As for the tree based models, optimal validation MAE was attained by models using both sensors with the 2.5 second sliding windows, and data from the lower back for the same models using the 5 second window.

Figures 5.2 and 5.3 illustrate the intra and inter model comparison for both of the validation schemes, using different window lengths.

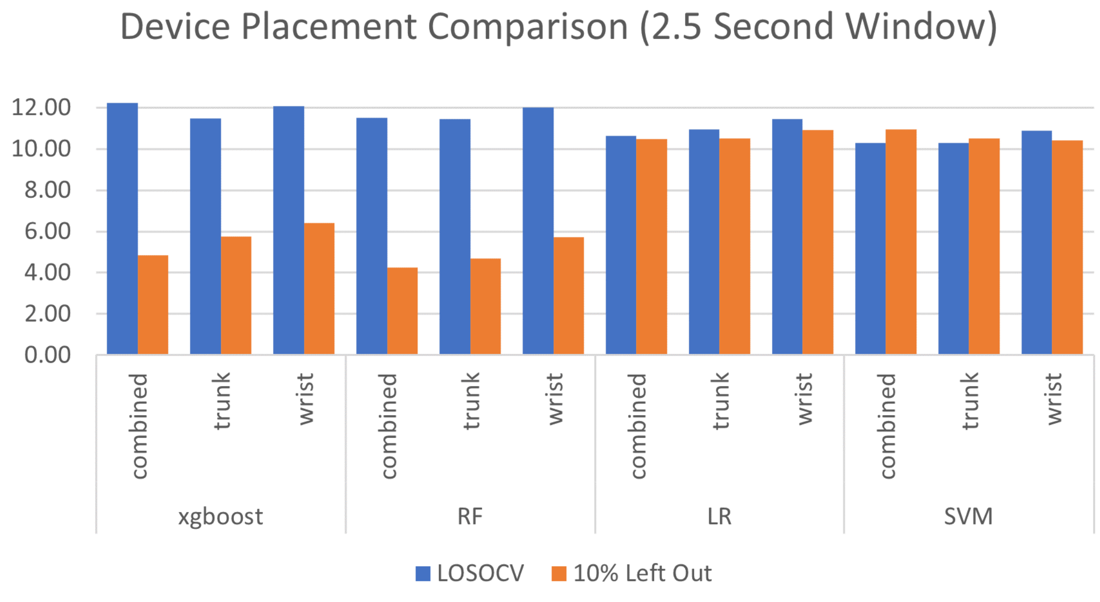

Figure 5.2: Effect of device placement on prediction outcomes using 2.5 second windows.

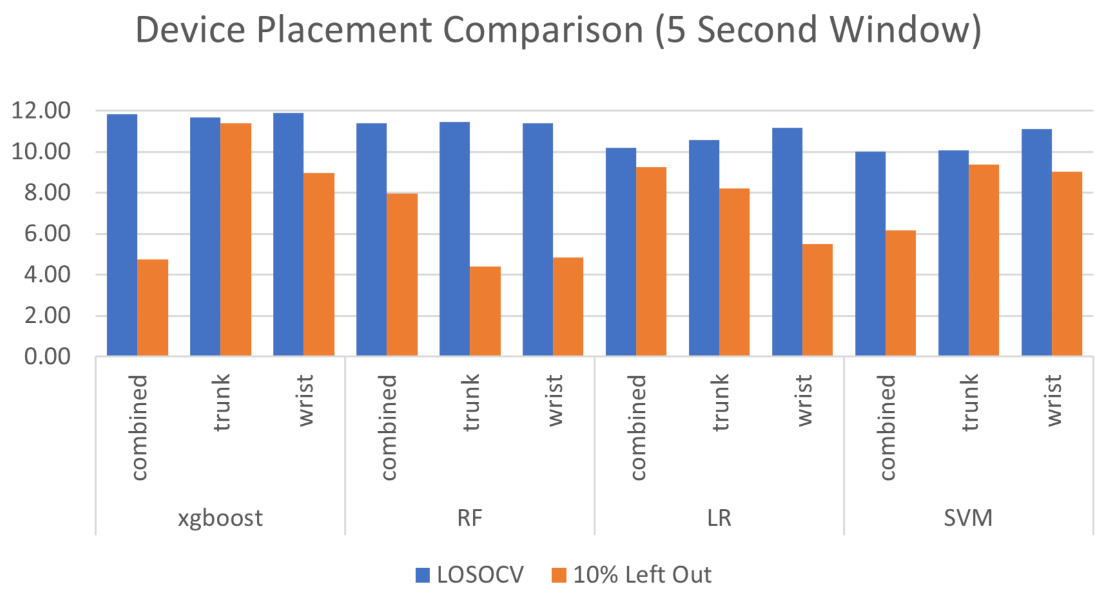

Figure 5.3: Effect of device placement on prediction outcomes using 5 second windows.

Sliding Window Length

Figures 5.4 5.5 and 5.6 compare the optimal performance of each model for the two tested sliding window lengths using data from both sensors combined and each separately, illustrating the MAE of predictions during LOSO CV and for the 10% of left out data.

While the fluctuations were relatively low during LOSO CV, most models performed better using a 5 second window length, with the exception of the xgboost model. MAE using the left out 10% of validation data fluctuated more considerably, but was also lowest using 5 second windows for all models except RF.

5.1.2 Model Parameters

As for model parameters, Table 5.3 summarizes which values yielded best performance during LOSO CV for each model, using a Grid Search procedure that exhaustively tested all parameter combinations for each model, independently of the used device placements and sliding window lengths.

The exhaustive nature of the grid search procedure makes this method of parameter optimization computationally expensive. For this reason, and considering that the procedure was used for several models, the used parameter space for each model was not as comprehensive as those used in some other works with a smaller scope and narrower focus. However, the present results should still serve as a good starting point for model tuning in future research.

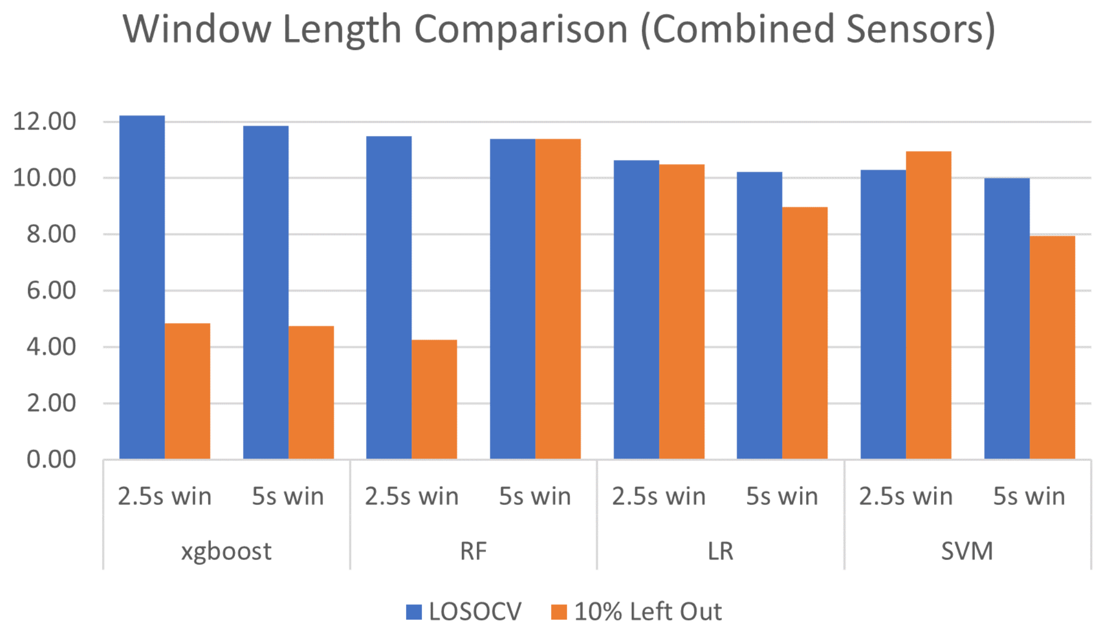

Figure 5.4: Effect of window length on prediction error for both validation schemes using data from both sensors.

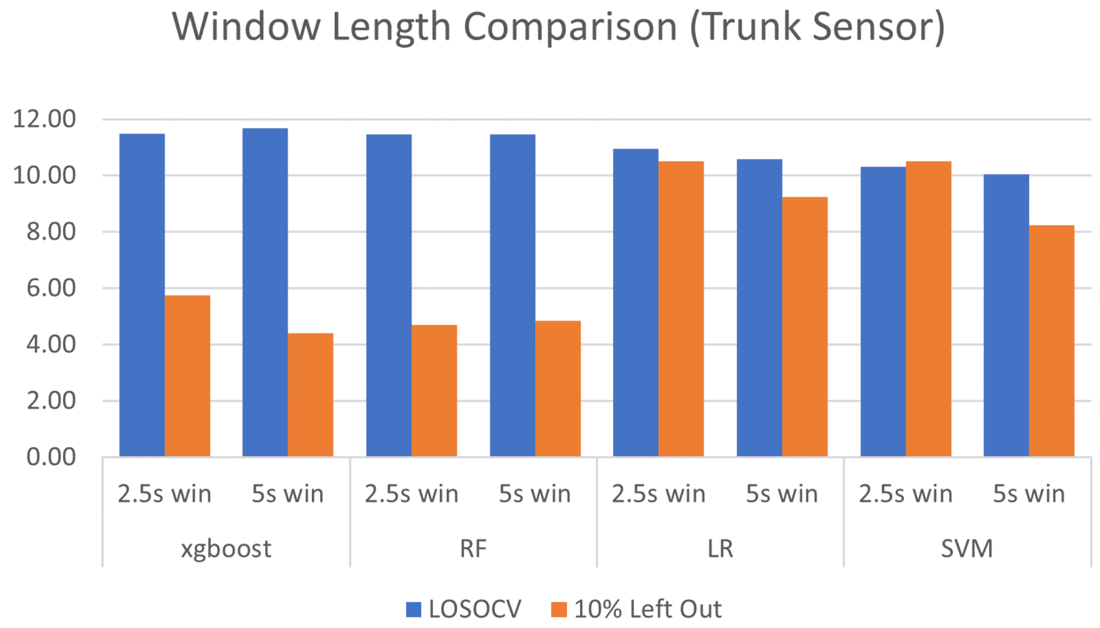

Figure 5.5: Effect of window length on prediction error for both validation schemes using data from the lower back sensor.

Model	Device Placement	Optimal Parameters
rf	combined	{'criterion': 'mae', 'max_features': 0.333, 'n_estimators': 250}
xgboost	trunk	{'colsample_bynode': 1, 'eta': 0.1, 'importance_type': 'total_gain', 'max_depth': 3, 'num_parallel_tree': 100, 'tree_method': 'gpu_hist'}
svm	combined	{'C': 10, 'epsilon': 0.3, 'gamma': 'auto', 'kernel': 'rbf'}
lin_reg	combined	-

Table 5.3: Optimal parameters used by models with the best LOSO CV performance.

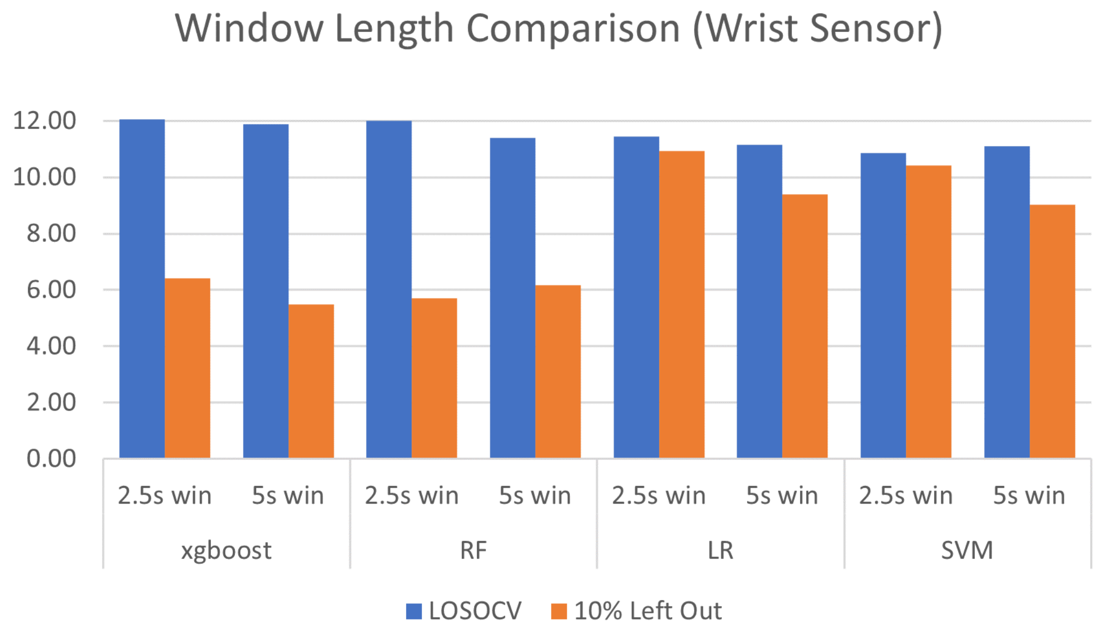

Figure 5.6: Effect of window length on prediction error for both validation schemes using data from the wrist sensor.

5.1.3 Feature Selection

For the models that benefited from it, several feature selection methods were tested, along with different numbers of features to select. The best performing linear regression and SVM based models used the SURF and relieF feature selection methods respectively, both selecting 25 as the optimal number of features. An analysis on the contribution of these features along with those used by the remaining models can be found in the following section.

5.2 Feature Contribution

5.2.1 Top Ranking Features

Finally, the top 20 contributing features used by each model in their optimal configuration for each window length are listed in Tables 5.4 and 5.5. While features from the tree based models are listed according to their importance during the estimation task on the 10% of left out data, features for the remaining models that were selected by feature selection algorithms are sorted according to their ranking during the feature selection task. The listed feature names were prefixed with the device placement that originated the feature and a number in the range of 0 to 3 representing the x, y and z axis or VM. The used sensor placement for each model is also included in the column headers.

svm (combined)	**lin_reg (combined)**	**rf (trunk)**	**xgboost (trunk)**
trunk_0_Histogram_0	trunk_0_Histogram_0	trunk_1_Histogram_4	trunk_1_Histogram_4
trunk_0_Histogram_1	trunk_0_Histogram_1	trunk_3_ECDF Percentile_0	trunk_3_ECDF Percentile_0
trunk_0_LPCC_4	trunk_0_LPCC_4	trunk_2_Wavelet energy_8	trunk_2_Wavelet standard deviation_8
trunk_0_LPCC_5	trunk_0_LPCC_5	trunk_3_Area under the curve	trunk_2_Histogram_5
trunk_0_LPCC_8	trunk_0_LPCC_8	trunk_2_Wavelet standard deviation_8	trunk_3_Area under the curve
trunk_0_LPCC_9	trunk_0_LPCC_9	trunk_3_Mean	trunk_3_Histogram_9
trunk_0_Spectral centroid	trunk_0_Spectral centroid	trunk_3_Histogram_9	trunk_1_Histogram_6
trunk_0_Spectral entropy	trunk_0_Spectral entropy	trunk_2_Wavelet standard deviation_7	trunk_2_Power bandwidth
trunk_0_Spectral slope	trunk_0_Spectral slope	trunk_1_Histogram_5	trunk_1_Histogram_5
trunk_1_Histogram_4	trunk_1_Histogram_4	trunk_1_Histogram_6	trunk_2_Wavelet energy_8
		trunk_1_Human range energy	trunk_2_Interquartile range
		trunk_2_Histogram_5	trunk_0_Spectral entropy
		trunk_2_Wavelet energy_7	trunk_2_ECDF Percentile_0
		trunk_1_Histogram_3	trunk_2_Histogram_4
		trunk_2_Histogram_4	trunk_3_Mean
		trunk_1_Median frequency	trunk_1_Human range energy
		trunk_2_ECDF Percentile_0	trunk_2_Wavelet standard deviation_7
		trunk_3_Spectral entropy	trunk_2_Root mean square
		trunk_2_Min	trunk_2_Neighbourhood peaks
		trunk_2_Wavelet energy_6	trunk_1_LPCC_10

Table 5.4: Top 20 features for models trained using 2.5 second windows

svm (combined)	lin_reg (combined)	rf (combined)	xgboost (trunk)
trunk_1_Histogram_4	trunk_3_Mean	trunk_1_Histogram_4	trunk_1_Histogram_4
trunk_3_Mean	trunk_3_Area under the curve	trunk_1_Median frequency	trunk_2_Wavelet energy_7
trunk_3_Area under the curve	trunk_1_Histogram_4	trunk_3_ECDF Percentile_0	trunk_2_Wavelet standard deviation_7
trunk_2_Mean absolute deviation	wrist_2_Median frequency	trunk_3_Histogram_9	trunk_3_ECDF Percentile_0
trunk_0_Spectral entropy	trunk_3_ECDF Percentile_0	trunk_3_Mean	trunk_1_ECDF Percentile_0
wrist_2_Median frequency	wrist_2_Fundamental frequency	trunk_2_Wavelet standard deviation_6	trunk_2_Wavelet variance_7
trunk_2_Wavelet energy_8	trunk_0_Spectral entropy	trunk_2_Wavelet energy_6	trunk_2_Wavelet energy_6
trunk_2_Wavelet standard deviation_8	trunk_1_Median	trunk_3_Area under the curve	trunk_3_Area under the curve
trunk_2_Interquartile range	trunk_2_Neighbourhood peaks	trunk_2_Wavelet energy_7	trunk_1_Median
trunk_2_Standard deviation	wrist_2_LPCC_9	trunk_1_ECDF Percentile_0	trunk_3_Mean
trunk_1_Median frequency	wrist_2_LPCC_4	trunk_2_Wavelet standard deviation_7	trunk_3_Histogram_9
trunk_3_ECDF Percentile_0	trunk_2_Wavelet energy_8	trunk_1_Median	trunk_1_Histogram_5
trunk_1_Median	trunk_2_Wavelet standard deviation_8	trunk_2_Wavelet energy_8	trunk_2_Area under the curve
trunk_3_Neighbourhood peaks	trunk_2_Mean absolute deviation	wrist_0_Interquartile range	trunk_2_Wavelet standard deviation_6
trunk_2_Wavelet energy_7	trunk_1_ECDF Percentile_0	trunk_2_Wavelet variance_7	trunk_2_Max power spectrum
trunk_2_Wavelet standard deviation_7	trunk_2_Power bandwidth	trunk_1_Spectral centroid	trunk_0_Spectral entropy
trunk_2_Median absolute deviation	trunk_3_Median	trunk_3_Spectral entropy	trunk_2_ECDF Percentile_0
trunk_2_Wavelet standard deviation_6	wrist_2_Zero crossing rate	trunk_1_Histogram_5	trunk_3_Neighbourhood peaks
trunk_2_Wavelet energy_6	wrist_0_Wavelet absolute mean_8	trunk_2_Wavelet variance_6	trunk_1_Mean absolute deviation
trunk_1_ECDF Percentile_0	wrist_0_Wavelet absolute mean_7	trunk_2_Max power spectrum	trunk_1_FFT mean coefficient_5

Table 5.5: Top 20 features for models trained using 5 second windows

Model	Device Placement	Window Length	# Trunk	# Wrist
svm	combined	500	19	1
lin_reg	combined	500	13	7
rf	combined	500	19	1
svm	combined	250	10	0
lin_reg	combined	250	10	0

Table 5.6: Comparison of feature counts extracted from either sensor for models trained on the combined feature space

5.2.2 Features by Device Placement

Among the 8 top performing models across the two tested window lengths, no model used data exclusively from the wrist, and only 3 models used data exclusively from the trunk. As for the remaining models, the majority of top ranking features were extracted from devices mounted on the lower back. Table 5.6 summarizes the number of features extracted from each sensor placement for the models that used both sensors. In some cases, no wrist features were ranked among the top 20, which suggests that although these were used for the estimation task, their contribution is minimal, which is line with the minimal performance gain in these models when compared against their counterparts using data exclusively from the lower back.

5.2.3 Features by Accelerometry Axis

Features from the antero-posetrior plane of movement (z axis) were the most prevalent among the top 20 extracted from the trunk sensor, consisting of 50 out of the 140 features considered for this analysis. The vertical plane of movement (x axis) produced the least amount of features among those considered here, with only 22 ranking among the top contributing features. Finally, the feature counts from the remaining y axis and VM were 32 and 27 respectively.

A similar analysis for features extracted from the wrist can be found in the wrist specific subsection, given the relatively low representation of such features among the best performing models.

5.2.4 Features by Domain

As previously stated, the feature space used for the estimation tasks consists of features spanning statistical, temporal and spectral domains. Spectral domain features were the most prevalent among these, making up almost half of the 140 considered features, with temporal domain features coming in second by a small margin, and temporal features last consisting of a quarter of this total. A comprehensive summary of these numbers comparing the total and per model prevalence of features from these domains can be found in Table 5.7

Model	Placement	Window Length	Statistical	Spectral	Temporal
svm	combined	500	9	9	2
lin_reg	combined	500	7	10	3
rf	combined	500	8	11	1
xgboost	trunk	500	9	8	3
svm	combined	250	3	7	0
lin_reg	combined	250	3	7	0
rf	trunk	250	11	8	1
xgboost	trunk	250	11	7	2
Totals			61	67	12

Table 5.7: Feature Domains used by optimal models for each window length

5.2.5 Wrist Based Features

Given their lower prevalence among the top performing models, features extracted from the wrist were largely disregarded in the previous subsections. This subsection contains an analysis of the features used by the top performing model using exclusively wrist feature.

Among those trained exclusively using wrist data, the best performing model during LOSO CV (MAE = 10.97) was an SVM trained using the 2,5 second window length for feature extraction. This model used a feature space of 25 features selected through the f statistic based feature ranking method. The selected features are listed in table 5.8 according to the previously described nomenclature.

Out of the 25 selected features, none were extracted from the temporal domain, while 8 and 17 were extracted from the statistical and spectral domains respectively, which were similarly predominant in models trained with data from the sensor located in the lower back. Features from the z axis were the most predominant among these, making up more than half of the subset, however it is difficult to draw conclusions from these results attending to the limitations discussed in the specific subsection about device orientation.

Features
wrist_0_ECDF Percentile_1
wrist_0_Interquartile range
wrist_0_Max
wrist_0_Mean absolute deviation
wrist_1_Median absolute diff
wrist_1_Spectral spread
wrist_2_Histogram_2
wrist_2_Histogram_4
wrist_2_Human range energy
wrist_2_LPCC_0
wrist_2_LPCC_10
wrist_2_LPCC_3
wrist_2_LPCC_4
wrist_2_LPCC_5
wrist_2_LPCC_6
wrist_2_LPCC_7
wrist_2_LPCC_8
wrist_2_LPCC_9
wrist_2_MFCC_0
wrist_2_MFCC_3
wrist_2_Spectral centroid
wrist_2_Spectral slope
wrist_2_Wavelet absolute mean_8
wrist_3_Histogram_8
wrist_3_Human range energy

Table 5.8: List of features used by the best performing model using exlusively wrist data

Chapter 6

Discussion

This chapter focuses on understanding the results beyond the reported values, and leveraging this knowledge to answer or discuss the posed research questions. Additionally, a summary of the limitations encountered throughout this project and possible work directions towards the objective monitoring of progression in PD is also included.

6.1 Research Questions

6.1.1 Comparison with state of the art deep learning models

In this study, data collected from PwPD during gait tasks was collected with the aim of automatically estimating motor symptom severity through the most clinically used scale for this purpose: part 3 of the MDS-UPDRS.

Recently, research on this topic has emerged using deep learning approaches for the estimation task. Specifically, the works of Hssayeni *et al.* [8]. and Rehman *et al.*[9] were used as references for comparison with the tested models. However, there are some methodological differences that are relevant for this discussion. These differences are summarised in Table 6.1.

As displayed in the table above, the tested models failed to achieve lower MAE scores than either of the studies considered for this comparison, with the difference in MAE to the best performing of these models being 4.44 points. Considering a possible maximum

Study	**[9]**	**[8]**	**Present Work**
Sample Size	119 PwPD	24 PwPD + 8 HC	74 PwPD
Monitored Activities	Lab Gait	Lab Simulated ADL	Lab Gait
Device Location(s)	Lower Back	Wrist, Ankle	Lower Back, Wrist
Data Type	Accelerometry (VM)	Angular Velocity (All axis)	Accelerometry (All axis + VM)
Type of Study	Longitudinal	Cross Sectional	Cross Sectional
Validation	Longitudinal	LOSOCV + 20% Left Out	LOSOCV + 10% Left Out
Results	MAE = 6.29	MAE = 5.95	MAE = 9.99

Table 6.1: Comparison of methodologies and results between this, and other studies for the automatic estimation of MDS-UPDRS III scores

of 132 points for Part III of the MDS-UPDRS, or 72 points for the population in this study, this difference may seem low, falling within the accumulated intra-rater variability commonly found in regular evaluations. However, having established that lower scores are possible, and attending to the reasonable results achieved by some of the tested configurations, it is necessary to understand where this difference may come from in order to direct further research.

While these studies share some commonalities, like the estimation of the same target variable and the usage of inertial data, there are several differences in methodology that can explain the difference in results. The longitudinal nature of [9] for example, is a fundamentally different approach for the estimation of the same scale. It poses the question of whether it is possible for a model trained on data from a set of patients to predict motor symptom severity for the same patients in a future point in time. This is significantly different from using data of novel patients that were not included in the training process, ie. LOSO CV, making it difficult to directly compare all approaches. In spite of this difference, both approaches are equally important, as the deployment of such a system in the 'real world' would eventually require the estimation of scores in both scenarios, and thus, the understanding of a model's capabilities for either. To bridge this gap, it would be interesting to know the LOSO CV MAE of the models used by Rehman *et al.* [9], and as mentioned in the section dedicated to future work, to establish a standard for data collection for this purpose that could enable more longitudinal research on the subject.

The use of angular velocity data and measurement of different activities by Hssayen *et al.* [8] are another set of differences that muddle comparisons. While their study suggests that data collected from such activities can provide relatively accurate estimations, the detection of these tasks in free living conditions, would be more complex than gait detection. This, along with the battery of research that has studied gait as a bio-marker for disease progression in PD and the marginal performance difference between the works of [9] and [8] could arguably suggest that gait is a more fitting activity to monitor, facilitating the comparison and validation of research in this field towards the deployment of this technology in free-living conditions.

Finally, and having discussed some of the shortcomings of the tested models when compared to deep learning approaches, there are some arguments in favor of their usage. The analysis of feature contribution for example, can give information on the relative descriptive power of different sensor locations. The work of [8] for instance, uses data from sensors mounted on the wrist and ankle, but due to the used models, it is harder to quantify how much each of these contributed in the estimation task. The analysis of the top contributing features, their domains, and the axis from which they were can also be used in future work to guide further research on the objective monitoring of PD. Finally, the relatively lower computational requirements and simplicity in implementing and training these models allows for a more exhaustive analysis on the optimal parameters and variables for the optimization of the estimation task. Considering these factors, while the tested models failed to match the performance of current state of the art deep learning approaches, further research is required before completely discarding their usage for the estimation of MDS-UPDRS III scores.

6.1.2 Effect of variables on the estimation task

Sliding window length

While fluctuations in LOSO MAE across the different sliding window lengths were relatively low, the best results for all sensor placements were achieved using a 5 second window. This window length was used in the works of Hssayeni *et al.* [8] and Rehman *et al.* [9], which is a strong suggestion towards the benefits of using window lengths that capture a larger amount of movement. As for the held out validation data, the MAE was similarly lower using 5 second windows for the SVM and Linear regression models, but not uniformly so for the tree based models. These latter models presented significantly lower MAE in their optimal configurations, which may have been the product of models overfitting on data from patients that were used in the training set, and exacerbated by the decreased window lengths, as results for the same models using a larger window are significantly closer to those achieved by the non-tree based models. In any case, these results suggest that a larger window favors MDS-UPDRS III estimation, which is in line with the window length used in recent research research on this topic.

Sensor Placement

The detection or measurement of tremor and bradykinesia using wrist worn devices has been a common approach for monitoring motor impairment in PD [63, 64]. For this purpose, measurements from the lower back have also been the object of several studies focusing on the gait of PwPD, and supporting gait characteristics as biomarkers for it's progression [16, 50] . As such, it is reasonable to expect that features extracted from data collected at these placements during gait could contribute for the estimation of motor symptom severity.

The results in this study suggest that the usage of sensors mounted on the lower back is more fitting for the estimation of MDS-UPDRS III scores, which is in agreement with the data collection setup used in the work of Rehman *et al.* [9].

Contrastingly, features extracted from data collected at the wrist seemed to improve the tested models' performance only slightly or not at all. Furthermore, the use of data exclusively from such sensor placement produced the worst results among all tested configurations for most of the models. In spite of the different data types and activities measured by Hssayeni *et al.* [8], their work using data collected from the wrists and ankles yielded better results than the methods used in this study. A plausible explanation for this may be that the degree of contribution of features extracted from each sensor placement was not measured separately due to the black box nature of the used algorithms, making it impossible to assess the contribution of either for the estimation task, meaning that the performance of the model could be mostly based on data collected from the ankle mounted sensors. However, given the extensive amount of research supporting the use of data collected from wrist worn devices for the objective monitoring of different symptoms in PD, it is unlikely that this is the case.

The findings of [ref] may partially explain the lower descriptive power of data collected from the wrist. In their study, the authors concluded that features computed from wrist data were more important for classification of HC vs. PwPD at H&Y Stage I, and decayed in importance for pairwise classification of each stage and the next, becoming significantly less relevant for discriminating stages II and III, which are the most dominant among patients in the dataset used for the current work. Furthermore the limitations described in the following section may have also negatively affected the results. As mentioned before, in spite of the results further research addressing these limitations is needed before discarding the use of such data for the estimation of MDS-UPDRS III scores.

Model selection

One of the goals of this study was to understand whether traditional, feature engineered models could estimate MDS-UPDRS III scores with comparable accuracy to deep learning based approaches described in current, state of the art research. As mentioned, among the tested models and variables, even the optimal configuration failed to match the performance of such models. However, the performance of the best model using among those tested using LOSO CV, is only 4.44 points worse than the best performing of the deep learning alternatives, which attending to the many differences between these studies and some of the limitations discussed in the following section, suggests that traditional feature engineered models may be viable for this task.

As mentioned in the Results section, the best performing model during LOSO CV was an SVM based model trained using 5 second windows, and 25 features from both used sensors selected through a relief based method. Considering only the optimal configurations for all models, fluctuations in LOSO MAE were relatively low, with the worst among these being a boosted trees model using 2.5 second windows and data collected exclusively from the sensor mounted on the lower back, achieving 11.49 LOSO MAE. Unexpectedly, the linear regression model held up to these results, outperforming both tree based models in LOSO CV. Attending to the relatively small feature subset that was considered optimal for the models using feature selection, this suggests that the larger feature space may be negatively affecting the performance of tree based models. This may also explain the significantly lower MAE achieved by these models when testing on the left out 10% of data, which is probably a product of these models overfitting on patient data. One possibility to confirm this hypothesis, would be to widen the parameter space to favor smaller values for the parameters controlling the number of features selected at each node for these models.

Ultimately, the relatively low variability among these models' results along with their reasonable performance for the estimation of MDS-UPDRS III scores are favorable arguments for further research using these models, addressing the limitations of this study, and expanding the search for optimal configurations of different variables towards this task.

6.1.3 Feature analysis

This subsection addresses some characteristics of the used feature spaces that could inform future research. In the previous chapter, the top 20 features used by each of the best performing models across both considered window lengths were reported, and briefly discussed. This analysis covered 140 features: the top 20 for 6 of the 8 models, and the 10 selected during feature selection from the remaining 2. The following discussion is mainly focused on these features, with references to the remaining feature space when relevant.

Number of Features

As previously discussed, the best performing models during LOSO CV used relatively small feature spaces, with only one of the top ten tested configurations using more than 25 features for the estimation task. Tree based models were expected to benefit from a larger feature set, but performed worse on average, when compared to their counterparts using smaller numbers of features. A wider search for optimal parameters and testing on different feature subsets of different sizes should be prioritized in future work, as it could drastically improve performance of these models, and in this process, possibly address the overfitting found during testing on the left out 10% of data.

Features by Axis

As the analysis made in the results section revealed, features extracted from data collected at the lower back were overwhelmingly more prevalent among the top features used by each model. Features extracted from the vertical axis of this sensor were the least prevalent compared to the remaining axis and VM, which attending to the existing literature on the study of gait characteristics in PwPD using similar data was unexpected . While the reasons for this discrepancy are difficult to assess, gait and it's characteristics have been repeatedly validated as biomarkers for the progression of PD, and as such, it would be interesting to test the usage of these characteristics as features or other approaches to exploit this relationship between vertical acceleration and gait[4, 65].

The most predominant features extracted from data collected at the lower back, were those extracted from the antero-posterior plane of movement, followed by the medio-lateral plane and finally the VM of the signal. While literature on gait assessment in PD has extensively used this device placement, there is a lack of information on the relative contribution of features or characteristics extracted from each axis of movement, making it difficult to draw conclusions on the topic. There is however evidence of some correlation between the severity of bradykinesia and features extracted from these axis of movement, collected in similar manner, for the analysis of Prakinsonian gait [66]. This, coupled with the previously mentioned results, make a compelling argument for the usage of all axis of acceleration in future work using similar data collection setups.

As for features extracted from wrist worn devices, it is difficult to properly analyze their contribution given the limitations discussed in both the previous and following sections. However, attending to the literature successfully using such data for the objective monitoring of several symptoms in PD, future work should address these limitations before opting for using exclusively data collected from the lower back.

Feature Domains

Most of the top ranking features across the optimal models were from the statistical and spectral domains, with those from the temporal domain making up only 12% of the total number considered for this analysis. Between the first two, while features from the spectral domain were slightly more prevalent, those from the statistical domain were overwhelmingly present in the top three of features listed according to their importance or ranking. Previous work on the use of signal features from motion data has extensively documented the possibilities for objective monitoring through the use of features from the spectral domain, with one of the most notable examples being tremor assessment. As for features from the statistical domain, although not as widely documented, several studies have included them for this purpose, although sometimes categorizing them differently [63, 53]. Finally, although features from the temporal domain were significantly under-represented among the top performing models, their use has also been documented in the literature. As such, and given the relatively low computational cost of having a slightly larger feature space, these features should still be investigated in future works, possibly even along others, like the phase plot features from the complexity domain described in the work of Dunne-Willows *et al.* [67].

Top Ranking Features

The top 3 most prevalent features among these, occurring more than 10 times each across this total, and consistently ranking among the top 5, could inform future research towards this estimation task. The following paragraphs provide an overview of these top ranking features, and, when possible, a brief discussion on the possible reason for their contribution.

With 27 occurrences, the most prevalent feature among the best performing models was a simple amplitude histogram of each signal window. This feature is computed by establishing a number of bins (10 by default) and using the value of each bin as a separate feature. Specifically, the fifth bin of the vertical acceleration was consistently ranked high across models, which may be related to the previously discussed value of vertical acceleration for gait assessment in PD.

The second and third most prevalent features were tied with 13 occurrences; however, features based on the Empirical Cumulative Distribution Function (ECDF) of the signal ranked higher, on average, than the wavelet based features. Specifically, the lowest ECDF percentiles extracted from the VM of trunk data were both higher ranking and more prevalent than wavelet energy. In spite of the difference in objective and methodology, ECDF derived features have repeatedly been used in activity classification studies [68], supporting their correlation with mobility features . This, along with the findings of Hammerla *et al.* [69] and the present results, support the use of ECDF based features in future work for the estimation of motor severity scores in PD.

As mentioned, the wavelet energy feature was equally as prevalent as the ECDF percentile. This specific feature computes the energy in each scale of the Continuous Wavelet Transform, although other CWT derived features were also present among the top ranking subset. The extensive literature on the usage of WT derived features from motion data collected in different setups [70] along with the high prevalence of these features among the most prevalent in this study are also good indicators for their usage in future work.

6.2 Limitations

As demonstrated, the tested methods for disease stage prediction were not as successful as current, state of the art deep learning models. While this may be due to some limitations intrinsic to the tested models, some external factors may have negatively affected the results.

6.2.1 Data Limitations

As previously stated, the dataset used for this project is the product of an ongoing collaboration between LASIGE at FCUL and the CNS. Given the large scope of this collaboration, and thus the volume of data produced, it is to be expected that not all data is adequate for the specific purposes of the present project. While some data selection methods were employed in order to address this issue, several factors may have negatively affected performance.

Device Alignment

While a protocol was established for data collection at the CNS, the large volume of data correct sensor alignment can not always be guaranteed. This misalignment is most obvious in data segments where a component of gravitational acceleration is added to one of the other axes due to sensor tilt, causing the anteroposterior and lateral accelaration data to have non zero mean as expected, and the gravity vector to have an average value different from the expected -1g [4]. However, this isn't the only case where misalignment may negatively effect the results. Musculoskeletal anomalies involving the hands of PwPD, can cause contractions and involuntary movements that may lead to unintentional rotations of a subject's wrists or palms along the vertical axis [71], resulting in skewed data that is imperceptible when visualizing accelerometry data.

The collection of orientation data using gyroscopes along the used accelerometer data, could enable the normalization of the accelerometery referentials, alleviating or completely eliminating this issue. While many modern IMU like the AX6 are capable of recording both of these data types simultaneously, the collection of an additional data stream poses other challenges for the continuous monitoring of PD progression. One major issue is the power consumption of gyroscopes, which can be an order of magnitude higher than accelerometers [64], limiting the recording time of these devices, or adding a burden for patients who would have to recharge their devices more frequently.

While some strategies could be employed to address this limitation, like using a significantly lower sampling rate for the gyroscope data, further work is required in order to find the optimal balance between higher battery usage and possible contributions of this data to improve the estimation tasks.

Walk initiation and ending

Gait initiation requires postural adjustements that are otherwise absent during the gait cycle. For PwPD, this phenomenon is even more prominent, with some patients exhibiting the previously mentioned FoG episodes [72]. Similarly, the final steps towards ending the gait cycle may exhibit differences in gait characteristics when compared to those of a regular, prolonged walk. Both of these phenomena may have negatively affected the results reported in this study. Due to the defined data collection protocol, and lack of a ground truth in the form of, for example, a video recording of each patients' walk, it is nearly impossible to exclude these periods to focus only on steady gait, without discarding a significant amount of data. While discarding the first and final windows of each gait instance for each patient was considered, this would result in the loss of more than 10% of all windows with 2.5 second length, and a significantly higher percentage for 5 second windows.

In the future, longer periods of gait can be favored over intermittent short walks to leave a bigger margin for the exclusion of these stages. Another option towards minimizing this problem could be the definition of data collection protocols that somehow account for or register the initiation and ending stages of a walk. This data could then be leveraged to feed other approaches for the objective monitoring of PD, like FoG detection as demonstrated in [ref], allowing for a more holistic monitoring of motor symptoms in PD.

MDS-UPDRS III representation and personalization

The dataset used in this study consisted of data collected from 74 patients. While this number of patients is significant for preliminary results, a larger sample size could improve the estimation task and further validate the present findings. Beyond the volume of data used to train the models, a wider range of MDS-UPDRS III scores could also improve the results, by including a wider variety of walking patterns that in a smaller sample size could be considered outliers, and negatively affect performance. Furthermore, the inclusion of a healthy cohort in the dataset could provide a baseline for the models to recognize healthy gait, exacerbating the difference between data from healthy and affected subjects and once again improving model performance.

6.2.2 Computational resources

The amount of variables considered in this study resulted in a fairly computationally expensive pipeline. While the data processing and feature extraction tasks were relatively simple to compute and save for further usage, model training and tuning, along with the chosen validation schemes required significant computational resources. In spite of the significant speed up enabled by the usage of a high performance cluster, more computational power and time would enable a wider search over the several variables considered, which could drastically improve model performance.

6.3 Future Work

Future research on the estimation of MDS-UPDRS III scores from gait data should continue to explore and assess the validity of different combinations of models, parameters, features and other variables towards the optimization of the estimation task. The previous section laid out most of the limitations of this study, which have a lot in common with research on the objective monitoring of PD.

Specifically, future work should strive to include more data, possibly including a HC, and spanning a wider range of disease stages and motor symptom severity. Furthermore, a validation of the used methods is required, and should be done by addressing some of the discussed limitations by including angular data from the used sensors and larger periods of gait to assess. The expansion of the search space when it comes to feature subsets and parameter optimization is also a recommended avenue for further work that could significantly improve performance. Finally, a similar study including gait data collected in free ling conditions should also be prioritized, as data collected under these conditions presents particular challenges and difference which have to be considered towards the deployment of any objective monitoring system for use in the ’real world.

Chapter 7

Conclusion

This study compared the performance of several feature engineered machine learning models trained using gait data against state of the art, deep learning approaches for the estimation of the MDS-UPDRS III, the most clinically used scale for motor assessment of PwPD. Furthermore, an analysis on the contribution and effect of different models, features and sensor locations of the collected data was also performed, providing a solid foundation for further research on this topic. Future work for the estimation of this scale using motion data should address the discussed limitations. Besides validating the current results, including more data with longer walks and spanning a wider range of motor symptom severity and possibly a HC, along with a more comprehensive search space for the tested models' parameters and used feature subsets could significantly improve model performance. Furthermore, the study of data collected in free living settings for disease staging through MDS-UPDRS III prediction should also be prioratized. The validation of the feature engineered and deep learning models for the objective monitoring of the disease using such data is extremely important, as this is a fundamental requirement towards the use of these approaches for objectively monitoring PD in the 'real world'.

Bibliography

[1] Ole-Bjørn Tysnes and Anette Storstein. Epidemiology of Parkinson's disease. *Journal of Neural Transmission*, 124(8):901–905, August 2017.

[2] Marios Politis, Kit Wu, Sophie Molloy, Peter G. Bain, K. Ray Chaudhuri, and Paola Piccini. Parkinson's disease symptoms: The patient's perspective. *Movement Disorders*, 25(11):1646–1651, 2010.

[3] K Ray Chaudhuri, Daniel G Healy, and Anthony HV Schapira. Non-motor symptoms of Parkinson's disease: diagnosis and management. *The Lancet Neurology*, 5(3):235–245, March 2006.

[4] Silvia Del Din, Alan Godfrey, Brook Galna, Sue Lord, and Lynn Rochester. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. *Journal of NeuroEngineering and Rehabilitation*, 13(1):46–46, December 2016.

[5] Spencer L. James, Degu Abate, Kalkidan Hassen Abate, and et. al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159):1789–1858, November 2018.

[6] Michelle Braybrook, Sam O'Connor, Philip Churchward, Thushara Perera, Parisa Farzanehfar, and Malcolm Horne. An Ambulatory Tremor Score for Parkinson's Disease. *Journal of Parkinson's Disease*, 6(4):723–731, 2016.

[7] Robert I. Griffiths, Katya Kotschet, Sian Arfon, Zheng Ming Xu, William Johnson, John Drago, Andrew Evans, Peter Kempster, Sanjay Raghav, and Malcolm K. Horne. Automated assessment of bradykinesia and dyskinesia in Parkinson's disease. *Journal of Parkinson's Disease*, 2(1):47–55, 2012.

[8] Murtadha D. Hssayeni, Joohi Jimenez-Shahed, Michelle A. Burack, and Behnaz Ghoraani. Ensemble deep model for continuous estimation of Unified Parkinson's Disease Rating Scale III. *Biomedical Engineering Online*, 20(1):32, March 2021.

[9] Rana zia ur Rehman, Lynn Rochester, Alison Yarnall, and Silvia Din. Predicting the Progression of Parkinson's Disease MDS-UPDRS-III Motor Severity Score from Gait Data using Deep Learning. volume 2021, pages 249–252, November 2021.

[10] Elizabeth Thompson, Peter Agada, W. Geoffrey Wright, Hendrik Reimann, and John Jeka. Spatiotemporal gait changes with use of an arm swing cueing device in people with Parkinson's disease. *Gait & Posture*, 58:46–51, October 2017.

[11] Megh Patel, Gottumukala Sai Rama Krishna, Abhijit Das, and Uttama Lahiri. A Technology for Prediction and Prevention of Freezing of Gait (FOG) in Individuals with Parkinson Disease. pages 395–403. Springer, Cham, 2017.

[12] J. Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, 79(4):368–376, April 2008.

[13] Christopher L Pulliam, Dustin A Heldman, Elizabeth B Brokaw, Thomas O Mera, Zoltan K Mari, and Michelle A Burack. Continuous Assessment of Levodopa Response in Parkinson's Disease Using Wearable Motion Sensors. *IEEE transactions on bio-medical engineering*, 65(1):159–164, 2018.

[14] Christopher G. Goetz, Barbara C. Tilley, Stephanie R. Shaftman, Glenn T. Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B. Stern, Richard Dodel, Bruno Dubois, Robert Holloway, Joseph Jankovic, Jaime Kulisevsky, Anthony E. Lang, Andrew Lees, Sue Leurgans, Peter A. LeWitt, David Nyenhuis, C. Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A. Teresi, Jacobus J. van Hilten, Nancy LaPelle, and Movement Disorder Society UPDRS Revision Task Force. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders: Official Journal of the Movement Disorder Society*, 23(15):2129–2170, November 2008.

[15] Roongroj Bhidayasiri and Pablo Martinez-Martin. Chapter Six - Clinical Assessments in Parkinson's Disease: Scales and Monitoring. In Kailash P. Bhatia, K. Ray Chaudhuri, and Maria Stamelou, editors, *International Review of Neurobiology*, volume 132 of *Parkinson's Disease*, pages 129–182. Academic Press, January 2017.

[16] Silvia Del Din, Alan Godfrey, Claudia Mazzà, Sue Lord, and Lynn Rochester. Free-living monitoring of Parkinson's disease: Lessons from the field. *Movement Disorders: Official Journal of the Movement Disorder Society*, 31(9):1293–1313, 2016.

[17] Antoine Regnault, Babak Boroojerdi, Juliette Meunier, Massimo Bani, Thomas Morel, and Stefan Cano. Does the MDS-UPDRS provide the precision to assess

progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. *Journal of Neurology*, 266(8):1927–1936, 2019.

[18] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken, 4th edition edition, April 2020.

[19] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009.

[21] Koray Açıcı, Çağatay Berke Erdaş, Tunç Aşuroğlu, Münire Kılınç Toprak, Hamit Erdem, and Hasan Oğul. A Random Forest Method to Detect Parkinson's Disease via Gait Analysis. In Giacomo Boracchi, Lazaros Iliadis, Chrisina Jayne, and Aristidis Likas, editors, *Engineering Applications of Neural Networks*, Communications in Computer and Information Science, pages 609–619, Cham, 2017. Springer International Publishing.

[22] Zhonelue Chen, Gen Li, Chao Gao, Yuyan Tan, Jun Liu, Jin Zhao, Yun Ling, Xiaoliu Yu, Kang Ren, and Shengdi Chen. Prediction of Freezing of Gait in Parkinson's Disease Using a Random Forest Model Based on an Orthogonal Experimental Design: A Pilot Study. *Frontiers in Human Neuroscience*, 15:636414, 2021.

[23] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[24] Decho Surangsrirat, Chusak Thanawattano, Ronachai Pongthornseri, Songphon Dumnin, Chanawat Anan, and Roongroj Bhidayasiri. Support vector machine classification of Parkinson's disease and essential tremor subjects based on temporal fluctuation. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2016:6389–6392, August 2016.

[25] Shyamal Patel, Konrad Lorincz, Richard Hughes, Nancy Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Matt Welsh, and Paolo Bonato. Monitoring Motor Fluctuations in Patients with Parkinson's Disease Using Wearable Sensors. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 13(6):864–873, November 2009.

[26] Zoubin Ghahramani. Unsupervised Learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen,*

Germany, August 4 - 16, 2003, Revised Lectures, Lecture Notes in Computer Science, pages 72–112. Springer, Berlin, Heidelberg, 2004.

[27] Hüseyin Gürüler. A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. *Neural Computing and Applications*, 28(7):1657–1666, July 2017.

[28] Jesse Mu, Kallol R. Chaudhuri, Concha Bielza, Jesus de Pedro-Cuesta, Pedro Larrañaga, and Pablo Martinez-Martin. Parkinson's Disease Subtypes Identified from Cluster Analysis of Motor and Non-motor Symptoms. *Frontiers in Aging Neuroscience*, 9:301, September 2017.

[29] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4:237–285, May 1996.

[30] Yejin Kim, Jessika Suescun, Mya C. Schiess, and Xiaoqian Jiang. Computational medication regimen for parkinson's disease using reinforcement learning. *Scientific Reports*, 11(1):9313, Apr 2021.

[31] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, USA, 1998.

[32] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature Selection: A Data Perspective. *ACM Computing Surveys*, 50(6):94:1–94:45, December 2017.

[33] Kenji Kira and Larry A. Rendell. The feature selection problem: traditional methods and a new algorithm. In *Proceedings of the tenth national conference on Artificial intelligence*, AAAI'92, pages 129–134, San Jose, California, July 1992. AAAI Press.

[34] Kenji Kira and Larry A. Rendell. A Practical Approach to Feature Selection. In Derek Sleeman and Peter Edwards, editors, *Machine Learning Proceedings 1992*, pages 249–256. Morgan Kaufmann, San Francisco (CA), January 1992.

[35] Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, September 2018.

[36] Tzu-Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 2015.

[37] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. PDVocal. In *The 25th Annual International Conference on Mobile Computing*

and Networking - MobiCom '19, pages 1–16, New York, New York, USA, 2019. ACM Press.

[38] Andong Zhan, Srihari Mohan, Christopher Tarolli, Ruth B. Schneider, Jamie L. Adams, Saloni Sharma, Molly J. Elson, Kelsey L. Spear, Alistair M. Glidden, Max A. Little, Andreas Terzis, E. Ray Dorsey, and Suchi Saria. Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity: The Mobile Parkinson Disease Score. *JAMA neurology*, 75(7):876–880, July 2018.

[39] Diogo Branco, Tiago Guerreiro, Ricardo Pereira, César Mendes, André Rodrigues, Raquel Bouça-Machado, Kyle Montague, and Joaquim Ferreira. DataPark: A Data-Driven Platform for Parkinson's Disease Monitoring. In *WISH Symposium - Workgrounp on Interactive Systems in Healthcare, co-located with CHI'19*, Glasgow, UK, 2019.

[40] Mathias Sundgren, Mattias Andréasson, Per Svenningsson, Rose-Marie Noori, and Anders Johansson. Does Information from the Parkinson KinetiGraph™ (PKG) Influence the Neurologist's Treatment Decisions?—An Observational Study in Routine Clinical Care of People with Parkinson's Disease. *Journal of Personalized Medicine*, 11(6):519, June 2021. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[41] Alberto J. Espay, Paolo Bonato, Fatta B. Nahab, Walter Maetzler, John M. Dean, Jochen Klucken, Bjoern M. Eskofier, Aristide Merola, Fay Horak, Anthony E. Lang, Ralf Reilmann, Joe Giuffrida, Alice Nieuwboer, Malcolm Horne, Max A. Little, Irene Litvan, Tanya Simuni, E. Ray Dorsey, Michelle A. Burack, Ken Kubota, Anita Kamondi, Catarina Godinho, Jean-Francois Daneault, Georgia Mitsi, Lothar Krinke, Jeffery M. Hausdorff, Bastiaan R. Bloem, Spyros Papapetropoulos, and Movement Disorders Society Task Force on Technology. Technology in Parkinson's disease: Challenges and opportunities. *Movement Disorders: Official Journal of the Movement Disorder Society*, 31(9):1272–1282, 2016.

[42] Hyoseon Jeon, Woongwoo Lee, Hyeyoung Park, Hong Ji Lee, Sang Kyong Kim, Han Byul Kim, Beomseok Jeon, and Kwang Suk Park. Automatic Classification of Tremor Severity in Parkinson's Disease Using a Wearable Device. *Sensors*, 17(9):2067, September 2017. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[43] O. Martinez-Manzanera, E. Roosma, M. Beudel, R. W. K. Borgemeester, T. van Laar, and N. M. Maurits. A Method for Automatic and Objective Scoring of Bradykinesia Using Orientation Sensors and Classification Algorithms. *IEEE Trans-*

actions on Biomedical Engineering, 63(5):1016–1024, May 2016. Conference Name: IEEE Transactions on Biomedical Engineering.

[44] Federico Parisi, Gianluigi Ferrari, Matteo Giuberti, Laura Contin, Veronica Cimolin, Corrado Azzaro, Giovanni Albani, and Alessandro Mauro. 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. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1777–1793, November 2015. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[45] Federico Parisi, Gianluigi Ferrari, Matteo Giuberti, Laura Contin, Veronica Cimolin, Corrado Azzaro, Giovanni Albani, and Alessandro Mauro. Inertial BSN-Based Characterization and Automatic UPDRS Evaluation of the Gait Task of Parkinsonians. *IEEE Transactions on Affective Computing*, 7(3):258–271, July 2016. Conference Name: IEEE Transactions on Affective Computing.

[46] Johannes C. M. Schlachetzki, Jens Barth, Franz Marxreiter, Julia Gossler, Zacharias Kohl, Samuel Reinfelder, Heiko Gassner, Kamiar Aminian, Bjoern M. Eskofier, Jürgen Winkler, and Jochen Klucken. Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLOS ONE*, 12(10):e0183989, October 2017.

[47] Evanthia E. Tripoliti, Alexandros T. Tzallas, Markos G. Tsipouras, George Rigas, Panagiota Bougia, Michael Leontiou, Spiros Konitsiotis, Maria Chondrogiorgi, Sofia Tsouli, and Dimitrios I. Fotiadis. Automatic detection of freezing of gait events in patients with Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 110(1):12–26, April 2013.

[48] Yuqian Zhang, Weiwu Yan, Yifei Yao, Jamirah Bint Ahmed, Yuyan Tan, and Dongyun Gu. Prediction of freezing of gait in patients with parkinson's disease by identifying impaired gait patterns. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(3):591–600, 2020.

[49] A. Godfrey, S. Del Din, G. Barry, J.C. Mathers, and L. Rochester. Instrumenting gait with an accelerometer: A system and algorithm examination. *Medical Engineering & Physics*, 37(4):400–407, April 2015.

[50] Silvia Del Din, Alan Godfrey, and Lynn 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. *IEEE Journal of Biomedical and Health Informatics*, 20(3):838–847, May 2016.

[51] Rana Zia Ur Rehman, Silvia Del Din, Yu Guan, Alison J. Yarnall, Jian Qing Shi, and Lynn Rochester. Selecting Clinically Relevant Gait Characteristics for Classification

of Early Parkinson's Disease: A Comprehensive Machine Learning Approach. *Scientific Reports*, 9(1):1–12, November 2019.

[52] André Branquinho, Helena R. Gonçalves, Joana F. Pinto, Ana M. Rodrigues, and Cristina P. Santos. Wearable gait Analysis LAB as a biomarker of Parkinson's disease motor stages and Quality of life: a preliminary study. In *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 234–239, April 2021.

[53] Rana Zia Ur Rehman, Christopher Buckley, Maria Encarna Micó-Amigo, Cameron Kirk, Michael Dunne-Willows, Claudia Mazzà, Jian Qing Shi, Lisa Alcock, Lynn Rochester, and Silvia Del Din. Accelerometry-Based Digital Gait Characteristics for Classification of Parkinson's Disease: What Counts? *IEEE Open Journal of Engineering in Medicine and Biology*, 1:65–73, 2020. Conference Name: IEEE Open Journal of Engineering in Medicine and Biology.

[54] M. Encarna Micó-Amigo, Idsart Kingma, Sebastian Heinzel, Sietse M. Rispens, Tanja Heger, Susanne Nussbaum, Rob C. van Lummel, Daniela Berg, Walter Maetzler, and Jaap H. van Dieën. Potential Markers of Progression in Idiopathic Parkinson's Disease Derived From Assessment of Circular Gait With a Single Body-Fixed-Sensor: A 5 Year Longitudinal Study. *Frontiers in Human Neuroscience*, 13:59, 2019.

[55] Anat Mirelman, Mor Frank, Michal Melamed, Lena Granovsky, Alice Nieuwboer, Lynn Rochester, Silvia Din, Laura Avanzino, Elisa Pelosin, Bas Bloem, Ugo Della Croce, Andrea Cereatti, Paolo Bonato, Richard Camicioli, Theresa Ellis, Jamie Hamilton, Chris Hass, Quincy Almeida, Maidan Inbal, and Jeffrey Hausdorff. Detecting Sensitive Mobility Features for Parkinson's Disease Stages Via Machine Learning. *Movement disorders : official journal of the Movement Disorder Society*, May 2021.

[56] Clare L Clarke, Judith Taylor, Linda J Crighton, James A Goodbrand, Marion E T McMurdo, and Miles D Witham. Validation of the AX3 triaxial accelerometer in older functionally impaired people. *Aging Clin. Exp. Res.*, 29(3):451–457, June 2017.

[57] Catarina Godinho, Josefa Domingos, Guilherme Cunha, Ana T Santos, Ricardo M Fernandes, Daisy Abreu, Nilza Gonçalves, Helen Matthews, Tom Isaacs, Joy Duffen, Ahmed Al-Jawad, Frank Larsen, Artur Serrano, Peter Weber, Andrea Thoms, Stefan Sollinger, Holm Graessner, Walter Maetzler, and Joaquim J Ferreira. A systematic review of the characteristics and validity of monitoring technologies to assess parkinson's disease. *J. Neuroeng. Rehabil.*, 13(1):24, March 2016.

[58] Marília Barandas, Duarte Folgado, Letícia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. Tsfel: Time series feature extraction library. *SoftwareX*, 11:100456, 2020.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[61] Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, 85:168–188, 2018.

[62] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H. Granat, Tom White, Vincent T. van Hees, Michael I. Trenell, Christoper G. Owen, Stephen J. Preece, Rob Gillions, Simon Sheard, Tim Peakman, Soren Brage, and Nicholas J. Wareham. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*, 12(2):e0169649–e0169649, February 2017.

[63] Luca Lonini, Andrew Dai, Nicholas Shawen, Tanya Simuni, Cynthia Poon, Leo Shimanovich, Margaret Daeschler, Roozbeh Ghaffari, John A. Rogers, and Arun Jayaraman. Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *npj Digital Medicine*, 1(1):1–8, November 2018. Number: 1 Publisher: Nature Publishing Group.

[64] Nikhil Mahadevan, Charmaine Demanuele, Hao Zhang, Dmitri Volfson, Bryan Ho, Michael Kelley Erb, and Shyamal Patel. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *npj Digital Medicine*, 3(1):1–12, January 2020. Number: 1 Publisher: Nature Publishing Group.

[65] Silvia Del Din, Aodhán Hickey, Cassim Ladha, Sam Stuart, Alan K. Bourke, Patrick Esser, Lynn Rochester, and Alan Godfrey. Instrumented gait assessment with a single wearable: an introductory tutorial. *F1000Research*, 5:2323, September 2016.

[66] Marie Demonceau, Anne-Françoise Donneau, Jean-Louis Croisier, Eva Skawiniak, Mohamed Boutaayamou, Didier Maquet, and Gaëtan Garraux. Contribution of a

trunk accelerometer system to the characterization of gait in patients with mild-to-moderate parkinson's disease. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1803–1808, 2015.

[67] Michael Dunne-Willows, Paul Watson, Jian Shi, Lynn Rochester, and Silvia Del Din. A novel parameterisation of phase plots for monitoring of parkinson's disease. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5890–5893, 2019.

[68] Megha Vij, Vinayak Naik, and Venkata M. V. Gunturi. Use of ecdf-based features and ensemble of classifiers to accurately detect mobility activities of people using accelerometers. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, pages 39–46, 2017.

[69] Nils Y. Hammerla, Reuben Kirkham, Peter Andras, and Thomas Ploetz. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ISWC '13, page 65–68, New York, NY, USA, 2013. Association for Computing Machinery.

[70] Avishai Wagner, Naama Fixler, and Yehezkel S. Resheff. A wavelet-based approach to monitoring Parkinson's disease symptoms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5980–5984, March 2017. ISSN: 2379-190X.

[71] Subhashie Wijemanne and Joseph Jankovic. Hand, foot, and spine deformities in parkinsonian disorders. *Journal of Neural Transmission*, 126(3):253–264, Mar 2019.

[72] Tiziana Lencioni, Mario Meloni, Thomas Bowman, Alberto Marzegan, Antonio Caronni, Ilaria Carpinella, Anna Castagna, Valerio Gower, Maurizio Ferrarin, and Elisa Pelosin. Events detection of anticipatory postural adjustments through a wearable accelerometer sensor is comparable to that measured by the force platform in subjects with parkinson's disease. *Sensors*, 22(7), 2022.