Universiteit Leiden

Master Computer Science

Parkinson's Disease Pull Test Proxy Model Development using Daily Life Activities

Name:	Tanya Dimitrov
Student ID:	s2949024
Date:	19/07/2022
Specialisation:	Data Science
1st supervisor:	Arno Knobbe
2nd supervisor:	Vasileios Exadaktylos

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

A loss of postural control in Parkinson's Disease (PD) patients is typically accompanied with a loss of daily living independence and physical independence. The pull test assesses postural control in PD patients and is often used as a balance assessment. This poses limitations as pull test assessments are infrequent, are only administered in a clinic, and have patient performance variation. Accelerometer data has the potential to overcome such shortcomings, by providing data on patient balance that can be gathered from home. We assess the extent to which such data can be used to predict pull test assessment outcomes, using data from the MJFF Levodopa Response Study. In this study, 16 subjects have completed a pull test assessment and a set of tasks (walking and standing) while equipped with an accelerometer. Classification and regression models were trained to predict a patient's pull test recovery based on accelerometer derived features from representative postural control tasks in the study. Although select models outperformed the Majority Vote Classifier on average, no model performance was statistically significant and there was large patient classification variation. Overall, these results suggest that further investigation with at-home accelerometer tasks for postural control is required with a larger patient pool representative of PD patient physiology.

Contents

1. Introduction

Parkinson's Disease is a disorder marked by the impairment of motor control. Although the hallmark symptom of Parkinson's Disease is the tremor, that is a single symptom in the amalgam of motor afflictions affecting daily life in patients with Parkinson's Disease. Patients suffer from bradykinesia, muscle rigidity, postural control, postural imbalance, freezing of gait and gait dysfunction that increase in severity as the disease progresses.

As symptoms of Parkinson's Disease progress it becomes difficult for a patient to complete independent daily living tasks. In addition, disease progression raises concern for a patient's health and safety in daily living, as a patient's motor impairments may lead to serious injury. Patients with Parkinson's Disease suffer from a lack postural control, manifesting in a stooped posture and a lack of a timely postural response. The latter increases a patient's fall risk, subjecting them to a risk of serious injury. A patient's everyday environment can therefore become a booby-trapped and dangerous ground to navigate. External perturbations, such as a pedestrian collision, can pose a fall risk to a patient with Parkinson's Disease. Lack of postural response coupled with gait impairments in a patient with Parkinson's Disease may mean a walking stumble turning into a fall. For some patients, a lack of postural control may happen spontaneously, resulting in an inability to remain balanced while standing. Postural control and balance impairments are thus important symptoms of Parkinson's Disease to monitor in a patient for their health and safety in daily living.

Motor impairments, such as gait impairments and lack of postural response, are a criteria in Parkinson's Disease assessments. Two of the most utilized assessments in assessing symptoms and Parksinson's Disease severity are the Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Heohn & Yahr Scale (H&Y)
.

The MDS-UPDRS is a 4 part scoring system to assess a patient's non-motor experiences of daily living (Part I) , motor experiences of daily living (Part II), motor skills (Part III) and motor complications (Part IV). The scoring system is based on the completion of a set of tasks in clinic that are assessed and scored on a scale by a professional [1]. All tasks in this section are scored 0 (no problems), 1 (minimal problems), 2 (mild problems), 3 (moderate problems), or 4 (severe problems). The MDS-UPDRS is used to gauge a patient's progression in Parkinson's Disease with attention to daily living tasks.

The MDS-UPDRS postural stability test (referred to commonly as the pull test in literature) is one of the most commonly administered tasks in the assessment of Parkinson's Disease. To perform the

test, a professional will stand behind a Parkinson's disease patient and administer a rapid, forceful tug on the patient's shoulders in a backwards motion. The test administrator will then count the number of steps that were taken before a patient regained postural stability. Patients may also be unable to regain postural stability and may need to be caught by the test administrator to prevent falling. The test administrator will score the patient in the following manner: 0: No problems in regaining postural stability, 1-2 steps taken for recovery; 1: 3-5 steps taken to recover postural stability; 2: More than 5 steps taken to recover postural stability; 3: Does not recover postural stability after the test but is able to stand without falling; 4: is unable to remain standing, falls with gentle pull on shoulders. The pull test is often used as the primary tool in assessing postural stability in Parkinson's disease. The results of the pull test can reveal how capable a patient may be in daily and physically independent living.

Following the results of the MDS-UPDRS assessment, Parksinson's Disease patients are assigned a H&Y stage to categorize disease severity based on functional disability [2]. The H&Y scoring system was first introduced in 1967 to provide a 1 to 5 scoring scale for disease severity. The scale was later revised in 1983 to include two subsections, subsections 1.5 and 2.5, to better characterize Parkinson's Disease progression. Both scoring metrics are currently used in clinical practice. Patients with an H&Y stage of 1 and 2 exhibit unilateral/bilateral impairments with little functional disability. Patients in this category are able to continue independent living and completion of daily tasks. Stage 2.5 is considered to be mild bilateral disease and is based on a patient's performance on the pull test. A patient with an H&Y stage of 3 exhibits slowness of movement and balance impairment. Patients are still considered physically independent at this stage, although daily tasks may be affected and therefore may require assistance (such as eating and dressing). The progression from unassisted living (H&Y Stages 1, 2 and 2.5) to some assistance in daily tasks (Stage 3) are distinguished on the H&Y scale by assessing a patient's postural instability and postural reflexes. A patient with H&Y stage 4 requires assistance with daily activities and is no longer considered to be physically independent. A patient with H&Y stage 5 has limited or no mobility. H&Y stages are primarily assessed by evaluating a patient's overall mobility using oral accounts of the patient's lifestyle from both the patient and the caretaker, gait, postural control, results of the pull test and often on results from the MDS-UPDRS.

Although the pull test and assessments such as the H&Y scale and MDS-UPDRS are integral to understanding a patient's condition, they can be insufficient. There is often a natural variation in a patient's performance to tests depending on whether a patient is having a "good day". There are many factors that may contribute to a "good" or "bad" day, such as quality of sleep, stress, or existing health conditions. A patient's performance on a "good day" may not be an accurate representation of a patient's abilities in their daily living. In addition, the Hawthorne effect is often present in a patient's clinical assessment [3, 4]. Patients often perform better during tasks in a clinical setting both because they are focused on the task and because they are under supervision. Therefore, an assessment such as the pull test may represent a patient's performance during a singular point in time, and may not be representative of a patient's true condition.

The development of non-biased, portable, and non-invasive tools for the long-term monitoring of Parkinson's Disease would be useful for robust disease assessment. Take-home tools, such as a Inertial Measurement Units (IMUs) or insole sensors, provide researchers with patient data on daily living tasks in their home environment. This may help understand variation between a patient's performance on "good" or "bad" days to get a comprehensive assessment of a patient's condition. Finally, take-home tools would provide long-term monitoring and progression of a patient's Parkin-

son's Disease with measurable markers.

Accelerometers specifically can be utilized by researches as a take home tool to monitor Parkinson's Disease progression. By implementing an accelerometer, a patient's postural stability can be monitored over longer periods of time, providing robust information on a patient's fall risk and daily living capabilities. In Parkinson's Disease, postural stability is often assessed using the pull test, which can only be administered in a clinical setting by a medical professional. However, postural stability and balance is complex and balance is required in daily living tasks. In this study we will analyze daily living tasks, such as gait and standing, using an accelerometer to model a patient's performance on the pull test. We will first investigate whether gait is a good predictor of postural stability by using a patient's pull test as a proxy measure. Secondly, standing balance tasks and sway will be analyzed as predictors of postural stability by again using a patient's pull test as a proxy measure. Finally, we will investigate whether a combination of gait and standing balance features can be analyzed in order to model a patient's postural stability using their results on the pull test.

2. Related Work

2.1 Gait Analysis using Accelerometers

Current applications of gait analysis using accelerometer in PD include PD discrimination against healthy patients, PD diagnosis, and motor status discrimination. Many works exists in the development of accelerometry tools including machine learning integration to analyze gait in PD. However, all studies suffer from a lack of validation in a larger population. A review article published in 2020 evaluated publications concerning PD gait analysis in the last 5 years and found a lack of a uniform standardization protocol for gait analysis. They concluded that further research needs to be completed before the use of accelerometry in PD could be properly evaluated [5]. In a systematic review completed in 2021 of gait assessment using smartphone accelerometers, strong validity, good to excellent reliability and good discriminating properties were reported. However, only 19% of studies in the review evaluated methods that they had developed [6]. A review of machine learning models for the analysis of gait accelerometry reported high specificity and sensitivity across all models in their review. All models that were tested did, however, have a low number of participants and used a small feature space. Thus although some models showed potential in clinical application, overall no algorithm had been validated and tested on a larger patient population independently [7]. Gait analysis using accelerometers has produced excellent results in individual studies and has the potential to be a useful tool for PD monitoring and research. However, the lack of uniformity in the field and external validation is a limiting factor.

2.2 Postural Sway in PD

Postural sway is used in the assessment of human balance. To assess postural sway, participants stand still for a set amount of time (30 seconds - 2 minutes) while the motion of their center of gravity is measured. Variations in postural sway assessment exist such as the participant's stance position and whether the test should be performed with eyes open or closed. A greater motion in the center of gravity indicates greater instability. Conversely, a smaller center of gravity movement indicates higher stability. Postural sway has been used in the assessment of balance with fatigue, sporting, medication intake, and aging. Postural sway has also been researched as a tool for balance analysis in PD, although studies have reported contradictory findings.

A study concluded that PD patients exhibited faster and shakier body sway compared to controls independent of PD H&Y stage [8, 9]. Conversely, a subsequent study found that postural sway differed only in patients with advanced forms of PD [10]. Further research concluded that postural

sway in PD patients may be influenced by the patient treatment. The study [11], concluded that patients taking Levodopa did not exhibit improved postural control and Levodopa may worsen a patient's postural sway. However, if patients were both taking Levodopa and had undergone deep brain stimulation then the negative effects on postural sway caused by Levodopa were dampened. Further research into the effects of Levodopa indicated that postural sway may be affected in patients taking Levodopa who also experienced dyskinesias. In patients without dyskinesias, Levodopa improved patient postural sway [12].

The effect of PD on a patient's postural sway is inconclusive. Studies comparing postural sway in PD differ in patient demographic and in research design. All studies suffer from small patient sample testing. Studies indicate that Levodopa has an effect on patient performance, but more research needs to be completed before the interaction of Levodopa with different PD patient physiologies is understood. In addition, as patient PD severity increases, patients are prescribed additional medications. Patient are often taking several medications when enrolled in research studies. Research into the various different medications in addition to Levodopa have not yet been studied. Therefore, analyzing sway in PD comes with complexity.

2.3 Sway Analysis using Accelerometers

Postural sway is typically assessed using a force platform sensor. However, force platform sensors are costly and not portable, posing a hurdle for the measurement of postural sway in at-home trials. As an alternative, postural sway can be measured using an accelerometer, as it is an affordable and portable device that can be given to patients for trials. Additionally, smart phones can be utilized to measure postural sway using the built in accelerometer.

Accelerometer sway applications have been developed to be used in industry. One such metric is the ISway method which was validated with healthy and PD patients against a force-platform. This is an important study as it removed the need for costly equipment of a force plate, and can provide a means of measurement for online postural sway analysis, creating a more robust picture of postural sway for patients. The tool was tested on both healthy subjects and patients with Parkinson's Disease, and proved to have high sensitivity and reliability for several of the derived features in respect to identifying PD patients with postural instability and gait difficulty (PIGD). In [13], sway features collected using a smartphone were used to distinguish PD patients from healthy patients. The analysis was completed using a Random Forest Classifier and the study reported an excellent sensitivity and specificity.

In [14], a review of wearable sensors for standing balance assessment of patients with PD was completed using 26 studies. Findings revealed that 11% of the studies had high methodological quality. However, regardless of study methodological quality, 81% of all studies could distinguish between PD patients and healthy controls. The study concluded promising results for the analysis of postural sway with wearable sensors and the distinguishing of PD patients from a healthy population, although a larger study to validate findings would be required as PD and postural sway is still not understood entirely.

2.4 Machine Learning Prediction of Pull Test Performance

A study to predict a patient's pull test score using data from an ankle mounted accelerometer data can be found in [15]. The study was composed of 139 PD patients divided into three different subtypes: bradykinetic dominant (n=47), tremor dominant (n=31), bradykinetic-tremor equivalent (n=61). Patients were asked to complete a battery of tasks several times. The tasks included a 10 meter walk, a Stop and Go walking test, heel-to-toe tapping and circling of the foot. There were a total of 24 features that were derived from the task activities, selected from a larger set of features using the ReliefF ranking algorithm. To account for unbalanced classes in the pull test score distribution in the dataset, under-sampling and over-sampling was completed using the SMOTE algorithm. Data was fed into a one-versus-rest hyperparameter optimized SVM model to predict a patient's performance on the pull test. Model performance was evaluated using a 5-fold cross validation using a balanced accuracy scoring metric. The most predictive task for the equivalent subtype to be the circling of the foot task, achieving an average balanced accuracy of 0.79. The most predictive task for the bradykinetic-dominant subtype proved to be the circling of the foot task, achieving an average balanced accuracy of 0.75. The most predictive task for the tremor-dominant subtype was found to be the Stop and Go task, achieving an average balanced accuracy of 0.70.

3. Background

3.1 Gait Analysis

Gait is studied in literature to understand human locomotion under different conditions, such as how human dynamics may be affected by Parkinson's Disease [16]. To study gait, the gait cycle is broken down into individual components, summarized in Figure 3.1. Gait can be broken down into the stance and swing phase. Stance phase refers to the period of time in which the designated foot (left or right) is making contact with the ground. The swing phase refers to the time in which the designated foot is not making contact with the ground. Gait can be further analyzed by looking at a subject's double support and single support phase. Double support refers to the time of gait in which both feet are making contact with the ground. The single support phase is the portion of the gait cycle in which a subject only has one leg making contact with the ground. Gait can also be analyzed using steps and strides. Metrics such as time, length, frequency, speed, variability and asymmetry are used to analyze different components of the gait cycle to understand human locomotion.

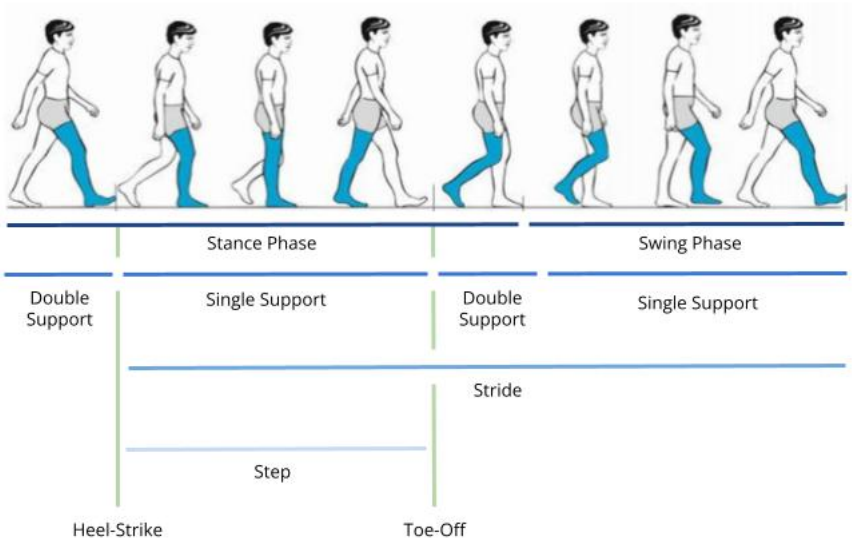


Figure 3.1: A visualization of different components of the gait cycle [16].

Heel-Strike and Toe-Off events are used to determine the different stages of gait. A Heel-Strike Event (HS) is the instance in which a foot makes initial contact with the ground. This is most often the heel of a foot making contact with the ground and entering a stance phase. A Toe-off (TO) event is last instance in which a foot is making contact with the ground. This is usually defined as the moment the toe is leaving the ground and the foot is entering the swing phase. HS and TO events are defined for both legs in the gait cycle and are used to calculate the different components of gait. A summary of the gait cycle in terms of HS and TO events can be seen in Table 3.1.

Gait Component	**Definition (HS and TO Events)**
Step	From HS of Side 1 to HS of Side 2.
Stride	From HS of Side 1 to the consecutive HS of Side 1.
Single Support	From the TO of Side 1 to the TO of Side 2.
Double Support	From the HS of Side 1 to the TO of Side 2.
Stance Phase	From the HS of Side 1 to the consecutive TO of Side 1.
Swing Phase	From TO of Side 1 to the consecutive HS of Side 1.

Table 3.1: The gait cycle in terms of HS and TO events. Side 1 and Side 2 correspond arbitrarily to the left and right legs.

3.2 Balance and Parkinson's Disease

In healthy subjects balance is defined as a combination of the vestibular system, proprioceptive system and visual system. In PD, balance is affected due to a faulty basalganglia and dopamine imbalances. As a result of the neurological differences in PD, rigidity, bradykinesia, impaired kinesthesia and sensory integration, gait impairments, freezing of gait and cognition impaired attention can be present and contribute to impaired balance [17]. Therefore, comprehensive and accurate balance assessment in PD is unintuitive.

There are several existing balance assessments that have been designed for the elderly population to assess mobility, fall risk and independent living capability. Many of these balance assessments have been marked as "recommended" or "suggested" by the International Parkinson and Movement Disorder Society in the assessment of PD patients [18]. Amongst the suggested assessments are the Berg Balance Scale (BERG), Tinetti Balance Scale (TBS), Balance Evaluation Systems Test (BESTest), and Timed Up and Go (TUG) Test. The tests vary on the number of assessment criteria, the average time of administration, the sensitivity and specificity, inter-rater scorer agreement and definitions of balance in the design of the assessment criteria. The assessments evaluate balance on a subset of different balance subcategories that can be summarized as static balance, dynamic balance and sensory orientation (not all assessments measure all the different categories) [19]. In addition to the broadly defined categories of balance assessment, some assessments will further develop subcategories for balance assessment, such as biomechanical constraints, stability limits/verticality, transitions/anticipatory balance, reactive balance, sensory orientation, and stability in gait [20]. BERG, TBS, BESTest and TUG all include a gait assessment task in their evaluation [21, 22, 23, 20, 24]. BERG, TBS, BESTest include an standing unsupported assessment task in their evaluation [21, 22, 23, 20]. The TBS and BESTest also include a variation of the pull test in their assessment [23, 20].

Despite the multitude of balance assessments, Parkinson's Disease patients are most often assessed

using the pull test. Results of the pull-test and the term "postural instability" have become interchangable in literature.

Balance, and balance in Parkinson's Disease, is an undefined field of study. There is still much to be understood about the mechanisms affecting balance control in Parkinson's Disease and how to properly assess balance in Parkinson's Disease.

4. Methods

4.1 MJFF Levodopa Response Study Dataset

The Michael J. Fox Foundation (MJFF) Levodopa Response Study is a public dataset that can be accessed at [25]. For this analysis, the Spaulding Rehabilitation Hospital (n=17) cohort will be used due to its detailed physician assessments of PD patients using MDS-UPDRS scoring criteria. The Spaulding Rehabilitation Hospital cohort is aged 46 to 80 with Heohn & Yahr scores ranging from stage 2 to stage 4. All subjects enrolled in the study were prescribed Levodopa, had symptoms of dyskinesia, had motor problems and were able to operate smart phones. Exclusion criteria for the study included suffering from other neurological conditions, having received Deep Brain Stimulation to treat Parkinson's Disease, and having extreme motor impairments that inhibit the subject's ability to walk for 30 seconds without a significant fall risk. A summary of patient statistics can be seen in Table 4.1.

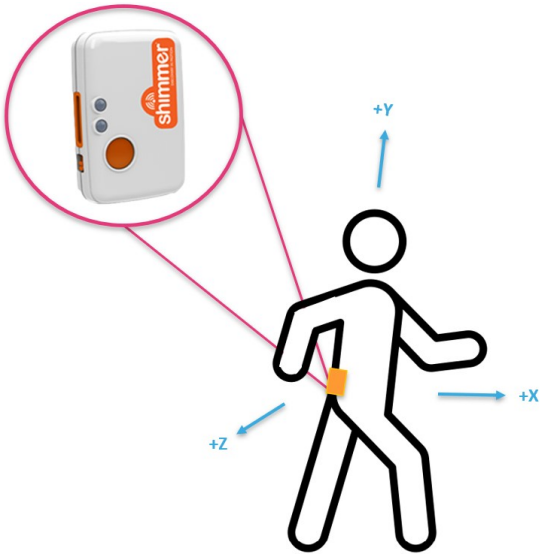


Figure 4.1: Diagram of Shimmer3 accelerometer attachment to patient at the lumbar. The blue lines denote accelerometer axis orientation [26].

Demographic	Descriptor	PD Patients
Gender (%)	Female	25
	Male	75
Age	Mean	63.52 (8.77)
	[Min, Max]	[46, 80]
Heohn & Yahr Score (%)	Stage 1	0
	Stage 2	68.75
	Stage 3	18.75
	Stage 4	12.50
MDS- UPDRS Total Score, ON State	Mean	77.76 (16.11)
	[Min,Max]	[56, 107]
MDS- UPDRS Part III Score, ON State	Mean	44.65 (11.802)
	[Min,Max]	[28, 66]
MDS- UPDRS Part III Score, OFF State	Mean	49.76 (12.62)
	[Min,Max]	[27, 76]

Table 4.1: Patient summary statistics of the Spaulding Rehabilitation Hospital cohort from the MJFF Levodopa Response Study.

Subjects were equipped with multiple sensors. Due to a high volume of literature supporting accelerometer analysis collected at the lumbar for gait and sway tasks, the Shimmer3 accelerometer attached to the lumbar will be used, see Figure 4.1. Subjects donned the sensors for 2 clinical visit days completing an MDS-UPDRS assessment and a battery of tasks under supervision of a professional at each visit. During clinical visit 1 subjects completed the MDS-UPDRS test and the battery of tasks while in an "On" state (Levodopa). During clinical visit 2 subjects were asked to come into the clinic in an "Off" state, having not taken their Levodopa medication since the night before. The subjects then performed the MDS-UPDRS assessment and one battery of tasks in an "Off" state. For a visualization of the relevant testing schedule, see Figure 4.2.

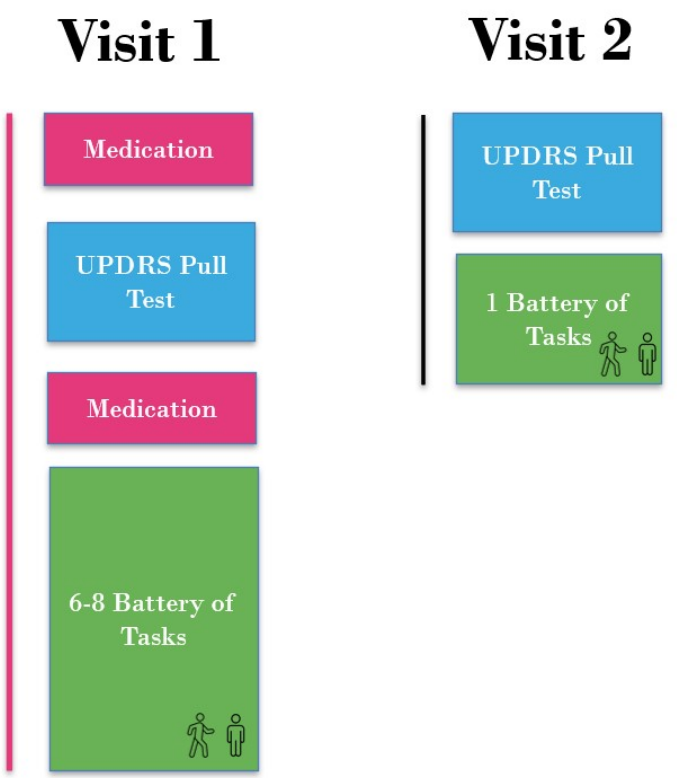


Figure 4.2: A visualization of the relevant assessment schedule from the MJFF Levodopa Response Study. The pink line on the left of Visit 1 represents that the subjects were on medication during testing. The black line on the left of Visit 2 represents that the subjects were off medication during testing.

In each battery of tasks, subjects completed a 30 second quiescent standing task and a 30 second walking straight task under supervision of a clinician. These are the relevant tasks that will be used in the analysis. During each MDS-UPDRS assessment, the pull test was completed once by a physician. The pull test requires the physician to stand behind the standing patient and administer a quick, forceful pull in a backwards direction on the patient's shoulders. The physician will then count the number of steps the patient took to recover their postural stability and score the patient from 0 - 4 based on the criteria outlined in the MDS-UPDRS assessment. A summary of the scoring criteria can be found in Table **??**. Each subject performed a pull test during Visit 1 (ON state, with medication) and during Visit 2 (OFF state, no medication).

Score	Category	Number of Steps	Recovery	Notes
0	Normal	1-2	Unaided	
1	Slight	3-5	Unaided	
2	Mild	> 5	Unaided	
3	Moderate	> 5	Aided	Stands safely, absence of postural response.
4	Severe	> 5	Aided	Loses balance spontaneously.

Table 4.2: MDS-UPDRS assessment criteria for the pull test.

4.2 Preprocessing

4.2.1 Processing completed by MJFF Levodopa Response Study

The dataset was minimally processed prior to being published. Accelerometry data was resampled to 50 Hz using linear interpolation. Timestamps were aligned by "shaking" the accelerometer at designated times during data collection for data alignment [27].

4.2.2 Data Segmentation

Accelerometry data in the study was published as a single accelerometry file per day (containing activity both in and outside of the clinic). The 30 second walking and standing tasks needed to be extracted from the full day accelerometry using the provided unix timestamps of task start and end times. The standing task was extracted and the whole ~ 30 second segment was used for analysis. The walking tasks were extracted and underwent further segmentation. The first 5 seconds and last 5 seconds of the recording were removed to have a clean gait segment without acceleration effects during gait initiation and deceleration effects during termination of gait.

4.2.3 Missing Data

There were instances of missing data for subjects. If a task segment was composed of $\geq 50\%$ NaN values then it was discarded from the analysis. One whole subject was removed from the analysis due to NaN files. The total number of patients in the cohort was therefore reduced to 16 patients.

4.2.4 Axis Transformation

To address the possibility of constant accelerometer tilt present during data collection, the algorithm developed by [28] was implemented. The tilt angle, θ, is determined by calculating how the gravitational component is distributed amongst the original axes, see Figure 4.3. Once θ is determined, trigonometric functions map the accelerometer data to a new x, y, z axis and correct for the accelerometer tilt angle.

4.2.5 Signal Processing

All task segments were detrended using the SciPy detrend function [29]. Filtering was completed for each data segment using a low pass filter with a threshold unique to each derived feature.

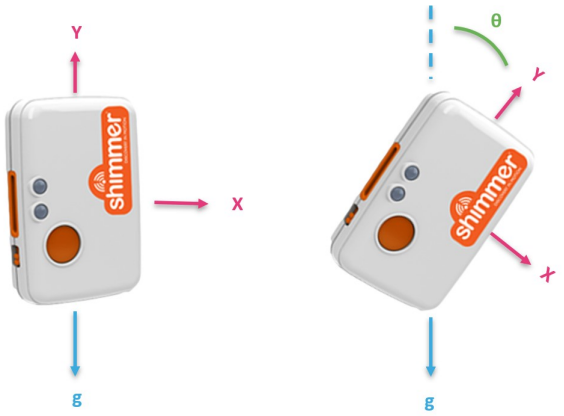


Figure 4.3: Axis transformation was completed to account for constant accelerometer tilt, θ, that may be present during accelerometer recordings. Accelerometer tilt was determined by looking at the signal recordings in respect to gravity (g).

4.3 Feature Engineering

Several features were derived from the walking and standing tasks. Features were selected from published literature. Features for gait unique analysis, sway unique analysis and features present in both analysis are summarized in the following sections.

4.3.1 Detection of Heel-Strike and TO-Events

Heel-Strike(HS) and Toe-Off(TO) Events in walking segments need to be extracted to calculate gait cycle features. Due to the lack of a gold standard comparison, HS and TO events were extracted solely from accelerometer data. To extract HS and TO events, the algorithm in [30] was utilized, see Figure 4.4. The method proposed in [30] was evaluated on 18 healthy subjects and reported a mean absolute difference of 0.019 seconds for detecting HS events, 0.032 seconds for extracting TO events and 0.018 seconds for step duration when compared to a validated instrumented walkway. The algorithm utilizes a series of signal processing techniques and continuous wavelet transforms (CWT) to determine HS and TO Events.

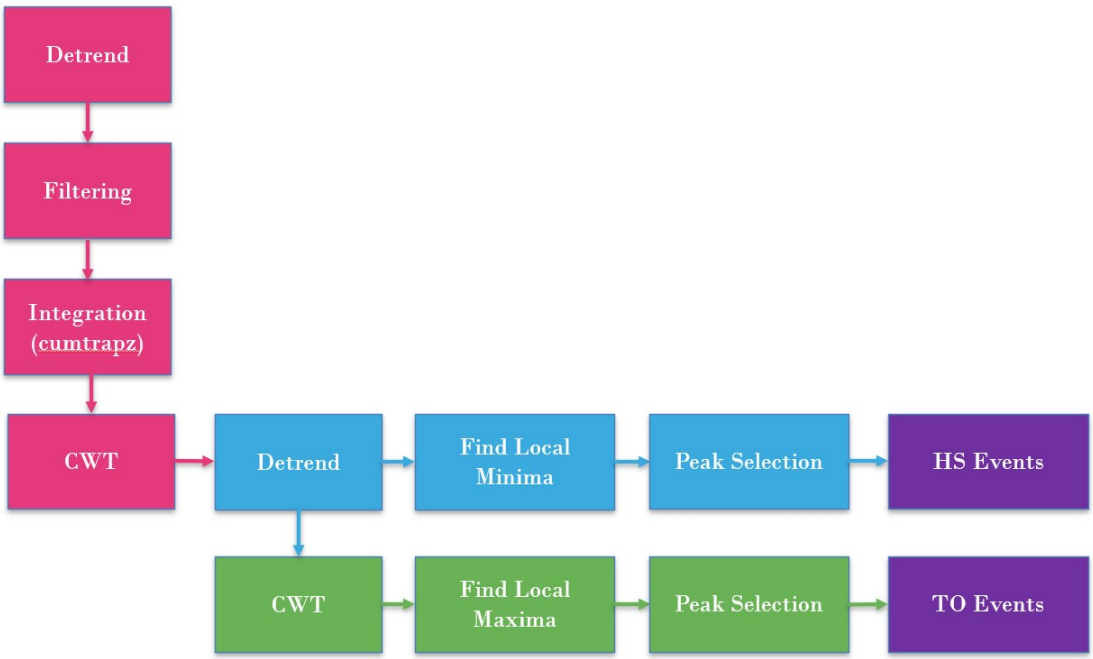


Figure 4.4: Algorithm for the detection of HS and TO events from accelerometry data.

The algorithm has been further validated in studies with PD patients. The study [31], validated the algorithm using accelerometer data with an instrumented walkway with PD patients and reported a mean absolute difference of 0.012 seconds in step time after adjustments to the algorithm. The authors found that extra HS events were detected with both the healthy and PD patient group. To correct for this they implemented an extra algorithm requirement that requires HS events to be within 0.25-2.25s of one another, as proposed in research by [32], in order to account for spurious events. The algorithm was also validated in [33] with PD patients against a Vicon imaging system. The study reported (95% CI) that HS events were within -0.09s to 0.1s of the Vicon detection and TO events were within -0.12s and 0.12s of the Vicon detection. The algorithm was modified by the authors of [33] to increase detection rates. The algorithm analyzed the anterior-posterior direction accelerometry signal instead of the vertical direction and the threshold for peak detection was lowered from 0.4 to 0.3.

The algorithm is implemented on the MJFF Levodopa Response Study in the following manner. The raw accelerometer signal is first preprocessed. The signal then undergoes detrending using the SciPy detrend toolbox. The signal is then filtered with a 4th-order low-pass bandpass filter at 10 Hz. The signal is then integrated using the SciPy integrate function. The signal undergoes a continuous wavelet transform with a Gaussian first order wavelet. A peak detection function is implemented and all peak troughs exceeding the threshold (0.25 x the mean detected peak heights) are identified as HS events. The signal undergoes a second continuous wavelet transform using a second-order Gaussian transform. A peak detection function is implemented and all peaks exceeding the threshold (0.25 x the mean detected peak heights) are identified as TO events. Further peak selection processing was required to account for patient gait morphologies present in the dataset. An example of gait segments from two subjects (subjects A and B) are presented in Figure 4.5. Figure 4.5 has the

the filtered walking signal in the top panel, the first order gaussian CWT in the second panel from which HS events are detected and the second order Gaussian CWT in the third panel from which TO events are detected. Subject A has a uniform walking pace and all three signals are clean. Subject B has a different gait physiology than Subject A, one that is less uniform. Due to differences in gait, the continuous wavelet transformed signals have a double-peak morphology in subject B. To account for the morphology that may be present in patients, the threshold for peak detection was set to 0.25. In instances of double peak detection in the first CWT wavelet (multiple HS detections) the first and deepest trough was chosen to be the estimate for the HS event. Where double peaks in the second CWT were detected (multiple TO detections) the second peak was used as the estimate for the TO event.

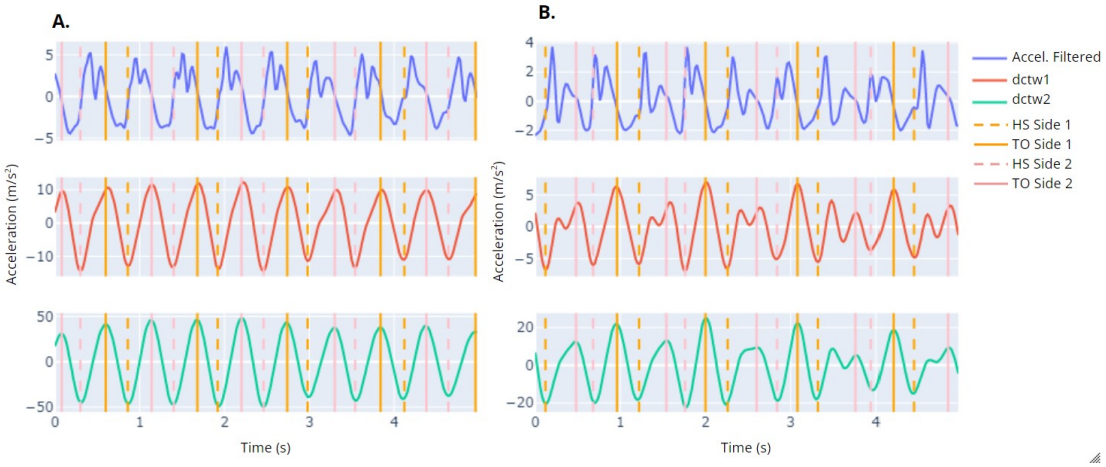


Figure 4.5: Accelerometry data (blue), first-order Gaussian CWT data (red) and second-order Gaussian CWT data (green) from two subjects, A and B. Dashed lines in the figure represent HS events. HS events are detected from red signals. Solid lines in the figure represent TO events. TO events are detected from the green signals. The solid and dashed lines are alternating in color to represent the different sides (left and right leg) of the patient during walking.

4.3.2 Gait Unique Features

Gait parameters were derived from previous literature [34, 35]. Gait cycle measures were determined using the HS and TO events and are summarized by the mean, variability and asymmetry of the different components of gait. Asymmetry refers to the difference in averages between both the left and right legs during gait. The frequency domain features were computed on both the Fast-Fourier Transformed signal and the Power Spectral Density estimation of the signal. A summary of all the features can be found in Table 4.3.

Measurement	Description
Gait Cycle Measures	
Average Step Time	Defined as the time between two consecutive HS events.
Average Stride Time	Defined as the time between two consecutive HS events of the same side.
Average Stance Phase Time	The time spent with foot contact of a side. The time between a HS to TO of one one side.
Average Swing Phase Time	The time that the foot is in the air . The time between a a TO event and HS event of a side.
Average Double Support Time	The time spent with both feet making contact with the ground. Defined as the time between a HS of one side to the TO of the other side.
Variability Step Time	The variability of all step times from both sides.
Variability Stride Time	The variability of all stride times from both sides.
Variability Stance Time	The variability of all stance times from both sides.
Variability Swing Time	The variability of all swing times from both sides.
Variability Double Support Time	The variability of all double support times from both sides
Asymmetry Step Time	The asymmetry in step time between sides 1 and sides 2 defined as the absolute difference of the means of the step times of both sides.
Asymmetry Stride Time	The asymmetry in stride time between sides 1 and sides 2 defined as the absolute difference of the means of the strode times of both sides.
Asymmetry Stance Time	The asymmetry in stance time between sides 1 and sides 2 defined as the absolute difference of the means of the stance times of both sides.
Asymmetry Swing Time	The asymmetry in swing time between sides 1 and sides 2 defined as the absolute difference of the means of the swing times of both sides.
Asymmetry Double Support Time	The asymmetry in double support time between sides 1 and sides 2 defined as the absolute difference of the means of the double support times of both sides.
Frequency Domain Measures	
Maximum Frequency	The frequency with the highest power in the PSD.
Centroid Frequency	The frequency at which the spectral mass is concentrated.
Power in Frequency Band Low (0.3-5 Hz)	The cumulative power from the PSD in the frequency band 0.3-5 Hz.
Power in Frequency Band High (5-8 Hz)	The cumulative power from the PSD in the frequency band 5-8 Hz.

Table 4.3: Gait Unique features derived from walking task data.

4.3.3 Sway Unique Features

Sway features for the analysis of standing tasks were derived from the following publications. The ISway method was tested on both healthy subjects and patients with Parkinson's Disease, and proved to have high sensitivity and reliability for several of the derived features in respect to identifying PD patients with PIGD [36]. The features derived in [37] were tested on both young and elderly healthy individuals. The research indicated that the accelerometer derived features were effective in separating the two subject groups and provided higher sensitivity than the traditional method of

using a force plate.

All time domain features except the 95% Confidence Circle Area and Sway Area were derived on the standing task accelerometer signal. The remaining two features, 95% Confidence Circle Area and Sway Area, were calculated on a mapped X-Z coordinate signal that can be seen in Figure 4.6. This is often referred to as a "spaghetti plot" and represents the movement in the anterior-posterior direction as well as medial-lateral direction during the standing tasks. It is often used in the analysis of postural sway and instability. A larger sway area indicates greater movement during standing and more instability. The frequency domain features computed on the Power Spectral Density transformation of the signal. A summary of sway unique features can be found in Table 4.4.

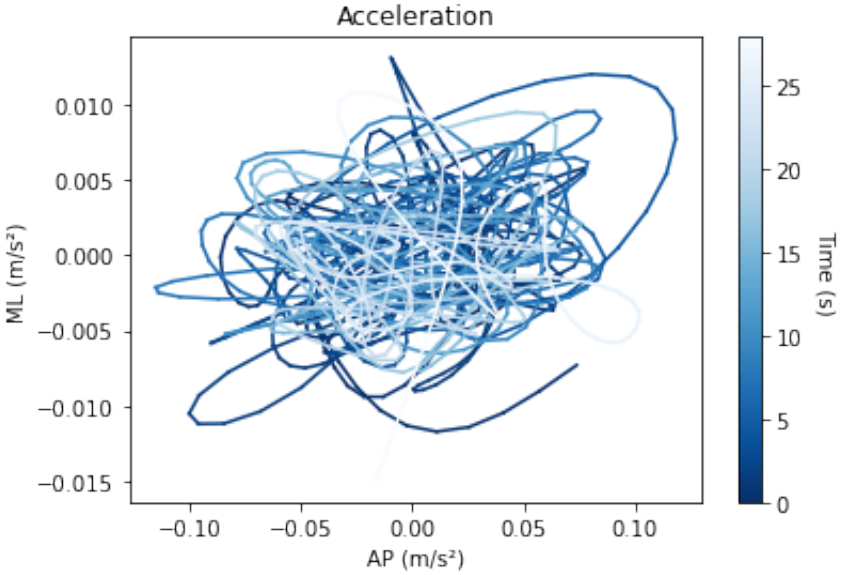


Figure 4.6: A standing task segment from a PD patient. The axis are the anterior-posterior Acceleration direction and the medio-lateral direction. The segment maps a patient's acceleration during standing task in both directions across time.

Measurement	Description
Time Domain Measures	
Average Jerk	Rate of change of acceleration in the X-Z plane.
Average Acceleration Magnitude	Average acceleration computed for each axis and the resultant vector.
Root Mean Square	Root mean square of every axis and the resultant vector.
Range	The range of acceleration signal, calculated for each axis and the resultant vector.
95% Confidence Circle Area	The modeled circle over the COP diagram in which 95% of the distances from the mean COP.
Sway Area	Approximates the area of the sway path on the ML-AP path.
Frequency Domain Measures	
Total Power	The total power of the power spectrum.
Median Power Frequency - 50	The median power frequency where 50% of the power is found.
Median Power Frequency - 95	The median power frequency where 95% of the power is found.
Centroidal Frequency	The frequency at which the spectral mass is concentrated.
Mean Frequency	Synonymous with the rotational frequency (Hz).

Table 4.4: Sway specific feature selection for standing task analysis.

4.3.4 Features Used In Both Gait and Sway

The features found in Table 4.5 were implemented in both the walking and standing task signals. The time domain measures were calculated on the filtered task segments. The Frequency Domain measures were calculated on the filtered Fast-Fourier Transformed task segments.

Measurement	Description
Time Domain Measures	
Standard Deviation	
Minimum	
Maximum	
Median	
Range	
Median Absolute Deviation	The median absolute difference of data points and the signal median; signal dispersion.
Inter Quartile Range	
Mean Absolute Deviation	The mean absolute difference of data points and the signal median; signal dispersion.
Inter Quartile Range	
Negative Count	The number of data points in the signal <0.
Positive Count	The number of data points in the signal >0.
Peak Count	The number of peaks in the signal as found using the SciPy toolbox.
Skewness	Signal skewness found using the stats toolbox.
Kurtosis	Signal kurtosis found using the stats toolbox.
Energy	Defined as the sum of the square of the data divided by 100.
Signal Magnitude Area	The sum the absolute signals across all three axes.
Frequency Domain Measures	
Mean	
Standard deviation	
Average Absolute Deviation	The mean absolute difference of data points and the signal mean; signal dispersion.
Minimum	
Maximum	
Range	
Median	
Median Absolute Deviation	The median absolute difference of data points and the signal median; signal dispersion.
Inter Quartile Range	
Count of Values above Mean	Count of the number of datat points above the signal mean .
Peak Count	The number of peaks in the signal as found using the SciPy toolbox.
Skewness	Signal skewness found using the stats toolbox.
Kurtosis	Signal kurtosis found using the stats toolbox.
Energy	Defined as the sum of the square of the data divided by 100.
Average Resultant Acceleration	

Table 4.5: Time and frequency domain features calculated for both walking and standing task segments. The frequency domain features were calculated on the Fast-Fourier Transformed signal.

4.4 Modeling

An overview of the modeling pipeline can be seen in Figure 4.7. A detailed explanation of each part of the pipeline will be discussed in the following sections.

4.4.1 Modeling Tasks

A subject's pull test score will be predicted using features derived from walking and standing task segments. Pull test scores in this assessment follow the MDS-UPDRS rating scale range from 0 to 4. The rating scale is a fine scale assessment that categorize patients as "Normal" (0), "Slight" (1), "Mild" (2), "Moderate" (3) and "Severe" (4). A patient can no longer recover independently from the pull test when they receive a score of 3. Each task that a patient completes (walking and standing) will be assigned a true pull test score. This corresponds to the pull test score that was completed during that clinical visit. Therefore, several tasks completed in one day will have the same pull task score assigned.

The prediction of the pull test score can be completed with several different modeling tasks. The pull test rating scale is an ordinal scale, increasing in severity with an increase in score assignment. Therefore, the task could be modeled as a regression task. However, the pull test scores can also represent a multi-class classification task with each score represented as one category. Finally, the modeling task can be represented in an application-focused task. The inability to recover from the pull test means the patient poses as a potential fall risk. Therefore, it is important to detect the moment of the transition from daily living tasks and low fall risk to impaired daily task and posing a fall risk. This can be completed by grouping the pull test scores 0-2 into a single category representing "Unaided Recovery" from the pull test. Pull test scores 3 and 4 can be grouped in a second category representing patients requiring "Aided Recovery". These patients pose a fall risk. Therefore the modeling task can also be posed as a binary classification task.

The classes in the dataset are unbalanced. When looking at the modeling task as a regression task the scores have the following distribution: Score 0: Score 1: Score 2: Score 3 is 51:15:12:29. When viewing the modeling task as a binary classification tasks the "Unaided Recovery": "Aided Recovery" is 78:30.

4.4.2 Models

For the regression task, the following models will be tested: Linear Regression, Lasso Regression, Decision Tree, Random Forest Regressor, XGBoost Regressor. For the multi-class classification task and the binary classification task the following models will be tested: Majority Vote, Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tree, SVM and Random Forest. All models were implemented using Python using scikit-learn toolboxes and the XGBoost toolbox [38, 39].

4.4.3 Feature Selection

Feature selection was completed by analyzing collinear features in the training set during each fold in the cross validation. Collinear features hold similar information to one another. By removing one of the collinear features the feature space is reduced, providing for better model training and reduce the risk of overfitting on the dataset. The reduced feature space was used in both the training and testing set.

Collinearity was detected using Variable Inflation Factors (VIF). In VIF analysis, each feature is regressed against the remaining features to determine the strength of feature correlation. In the regression analysis, the R^2 variable is extracted between both features, and the VIF is calculated using the following equation:

$$VIF = \frac{1}{1 - R^2} \tag{4.1}$$

Two independent features with 0 correlation will produce an R^2 value of 0, indicating a VIF score of 1. Highly correlated features will cause higher R^2 values , approaching the number 1. The VIF is unbounded in its upper limit. High VIF values mean features are very correlated. As a rule of thumb, a VIF score of 5-10 indicates high multicollinearity between the feature and remaining features. For this analysis, a VIF threshold of 7.5 was used. All features that had a VIF score above 7.5 were removed from the training and test set for the fold.

4.4.4 Model Evaluation

To prevent data leakage due to underlying patient physiology, the model will be evaluated using a Leave-One-Subject-Out Cross-Validation [40, 41, 42]. This is a representative estimate of model performance in the real world as each fold represents the model performance on a new incoming patient. The models will be evaluated on 16 folds representing the 16 patients in the Spaulding Rehabilitation Hospital cohort.

In regression models, the R^2 value and the Mean Absolute Error (Test MAE) will be reported. In classification models, due to the unbalanced classes models will be evaluated using the balanced accuracy score. The balanced accuracy is the average of the sensitivity and specificity score, see Equation 5.2. The sensitivity represents the portion of correctly classified positive class. The specificity represents the portion of correctly classified negative class. In multi-class classification, the specificity and sensitivity are the averaged specificity and sensitivity across all classes [43].

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \tag{4.2}$$

4.4.5 Post-Hoc Model Evaluation

To further evaluate model performance, classification models will be compared to the Majority Vote Classifier. The Majority Vote Classifier is the "Dummy Classifier" that classifies the test subject with the most frequent class present in the training set. A t-test will be performed to test classifier model predictors against the "Dummy Classifier" to test for model significance. Significant models are models that have a p-value < 0.05.

In addition, models will be compared against one another by using the Nemenyi test. The Nemenyi test evaluates whether models are statistically different from one another based on their average rank performance on test datasets. In every fold of the cross validation, a model was ranked 1 if the model achieved the highest balanced accuracy for that fold and a model was ranked 8 if it achieved the lowest balanced accuracy score. Models with tied balanced accuracy score are given the average rank performance (i.e. if 2 models were 2nd highest in balanced accuracy scores they would both receive the rank 2.5). The Nemenyi test then calculates the critical difference of the models based on the number of models and the number of datasets the models were tested on. All models within

the critical distance rank of one another are displayed with a bar and are not statistically different. All models not connected to a bar with other models are statistically different models [44].

4.4.6 Feature Importance

One way of interpretting model performance is by assessing the feature importance of each model. This is important in analysis as it gives insight into the most predictive features. In a medical data application, this gives context to the modelling task. Although not all models are interpretable, a metric such as the Shapley Additive Explanations (SHAP) value can increase model transparency. A SHAP value will calculate the contribution of each feature in the model to the prediction task for each instance. A SHAP value will be calculated by creating 2^F models, where F is the number of features in the dataset, and computing each features marginal contribution. For more information, see [45].

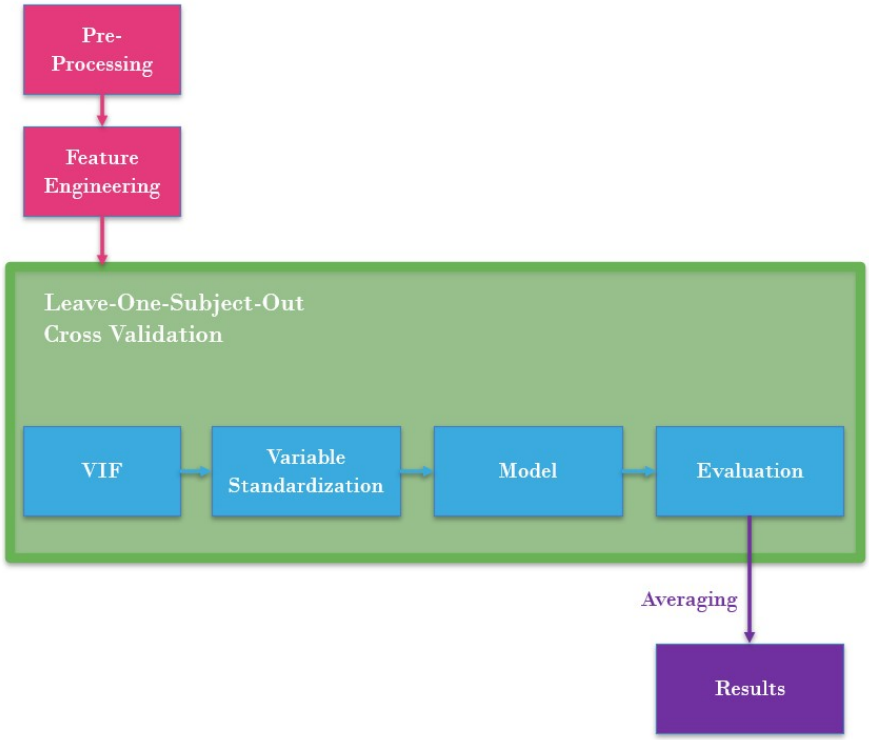


Figure 4.7: A pipeline for the model analysis. The data first undergoes preprocessing and feature engineering. Then 16-Fold Leave-One-Subject-Out Cross Validation is performed. During each fold of the cross validation VIF feature selection and variable standardization (when applicable) is performed. The model is then evaluated with the predefined metrics. Results of the model performance are averaged across all 16 folds.

5. Results

VIF feature selection was completed at each fold in the LOSO CV. Feature selection was completed to reduce the feature space and prevent overfitting in the dataset. Table 5.1 presents the list of features that were selected in one of more folds of the CV to be used in training the models. In the standing task models, the majority of the features that were found to be uncorrelated were the frequency-domain features. In the combined task models, the final feature set was composed of more gait features than sway features.

Final Features	**Axes**		
Standing Task Models		*Walking Task Models*	
Average Resultant FFT		Stance Time Variability	
Negative count	y, z	Stance Time Asymmetry	
Postive count	x, y, z	Swing Time Asymmetry	
Peak count	x, y, z	Median	x
Skewness	x, y, z	Median Absolute Deviation	x, y
Kurtosis	x, y, z	Inter Quartile Range	y
Energy	x, y, z	Positive Peak Count	z
Signal Magnitude Area		Skewness	x, z
Mean FFT	x, y, z	Kurtosis	x, z
Standard Deviation FFT	x, y, z	Maximum Frequency	z
Average Absolute Difference FFT	x, y, z	FFT Energy	z
Min FFT	x, y, z	*Combined Task Models*	
Max FFT	x, y, z	Median Absolute Devaition Gait	x, y
Range FFT	x, y, z	Negative Count Gait	y
Median FFT	x, y, z	Positive Count Gait	z
Median Absolute Deviation FFT	x, y, z	Skewness Gait	x, z
Inter Quartile Range FFT	x, y, z	Kurtosis Gait	x, z
Above Mean Count FFT	x, y, z	Max Frequency Gait	z
Peak Count FFT	x, y, z	Kurtosis FFT Gait	x
Skewness FFT	x, y, z	Negative Count Standing	y
Kurtosis FFT	x, y, z	Above Mean Count FFT Standing	y
Energy FFT	x, y, z	Energy FFT Standing	x

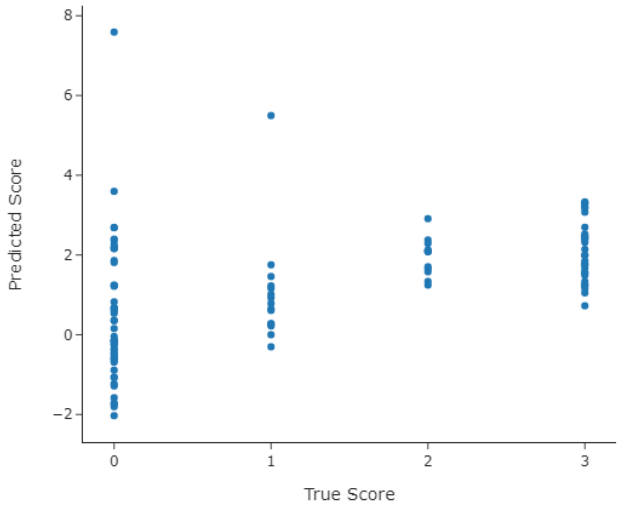
Table 5.1: Features selected in one or more folds after VIF. The selected features were used to train the models for walking, standing and combined tasks.

Pull test prediction was modeled using three modeling tasks: regression, multi-class classification and binary classification. The results of the regression task can be found in Table 5.2.

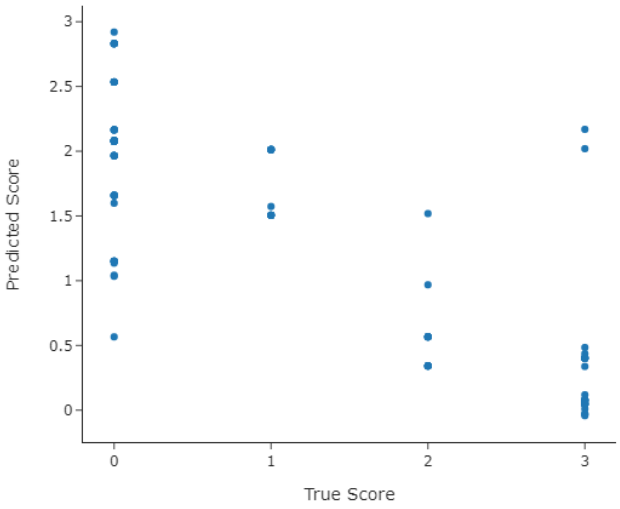
		Walking	**Standing**	**Combined**
Linear Regression	Train R^2	0.77 ±0.05	1.00 ±0.00	0.71 ±0.05
	Test R^2	-9.61 ±34.94	-1.49 ±0.03	-0.16 ±0.01
	Test MAE	1.01 ±0.90	2.85 ±2.31	0.98 ±0.84
Lasso Regression	Train R^2	-0.85 ±0.10	-0.83 ±0.10	-0.81 ±0.10
	Test R^2	-5.21 ±14.61	-0.23 ±1.25	0.42 ±0.58
	Test MAE	1.16 ±1.27	1.13 ±1.27	1.13 ±1.27
Decision Tree	Train R^2	0.88 ±0.05	0.80 ±0.11	0.91 ±0.02
	Test R^2	-5.16 ±11.71	-1.34 ±0.60	-0.09 ±0.28
	Test MAE	1.42 ±0.99	2.15 ±0.98	0.98 ±0.92
Random Forest	Train R^2	0.77 ±0.04	0.77 ±0.04	0.80 ±0.04
	Test R^2	-2.52 ±6.80	-1.49 ±0.02	-0.16 ±0.01
	Test MAE	1.29 ±0.87	1.70 ±0.78	1.20 ±0.84
XGBoost	Train R^2	1.00 ±0.00	1.00 ±0.00	1.00 ±0.0
	Test R^2	-2.67 ±5.50	-1.49 ±0.02	-0.16 ±0.01
	Test MAE	1.15 ±0.79	1.86 ±0.79	1.03 ±0.60

Table 5.2: Train R^2, test R^2, and test MAE of the regression models averaged (± standard deviation) across all folds. The regression task independent variable ranged from 0-3.

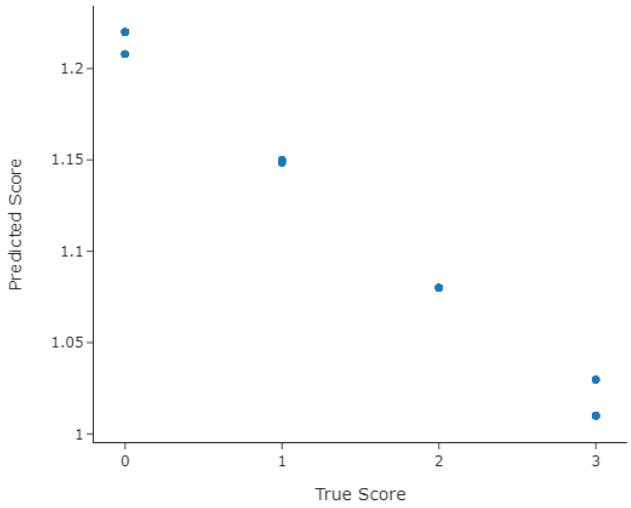
Pull test prediction as an ordinal regression problem could not be modeled effectively using the derived features and dataset. All models reported high train R^2 values and negative test R^2. Models are overfitting on the training set and are not generalizing to new data. The negative test R^2 indicates the model is performing worse than the average pull test score prediction. An investigation into pull test predictions in the test sets across all patients for select models can be found in Figure 5.1. Three models, walking task Linear Regression, sway task XGBoost and combined task Lasso Regression were chosen to provide context for different types of error in the regression models. The walking task Linear Regression model was the worst performing model in the regression modeling task both in average test R^2 and standard deviation. The true vs. predicted values reveal that the model is creating a negative association between true values and predicted values (i.e. low test values predict high values and vice versa). In addition, the model is predicting within a range of [-2, 8] while the true values range from [0, 3]. In the sway task XGBoost true vs.predicted plot, the range of predictions is in the correct range but the negative association between true and predicted values pervades. The negative association between true and predicted values is also present in the combined task Lasso Regression model. In addition, the model is predicting the same 6 values across all patients. Finally, the predicted score range is between [1, 1.4] whereas the true prediction range is [0, 3].


Predicted Score
8
6
4
2
0
−2
0
1
2
3
True Score

(a)


Predicted Score
3
2.5
2
1.5
1
0.5
0
0
1
2
3
True Score

(b)



(c)

Figure 5.1: True vs Predicted values on the all test sets in a model for (a) walking task Linear Regression model, (b) standing task XGBoost model and (c) combined task Lasso Regression model.

Results of the multi-class classification (4 classes) task can be seen in Table 5.3. All models in the standing task except for K-Nearest Neighbors were not able to perform equal to or better than chance (0.25). All walking task models and combined task models performed better than chance, but did not outperform the Majority Vote classifier model. High standard deviation scores across folds for all models indicate that there is high patient variability in model performance. Overall, low average balanced accuracy scores and high standard deviations indicate that a 4-class classification model may be too complex for the current sample size and feature space.

	Walking	**Standing**	**Combined**
Majority Vote	0.50 ±0.50	0.50 ±0.50	0.50 ±0.50
Logistic Regression	**0.46 ±0.36**	0.20 ±0.37	0.43 ±0.44
Naive Bayes	0.38 ±0.40	0.19 ±0.37	0.45 ±0.42
K-Nearest Neighbors	0.43 ±0.38	**0.48 ±0.47**	0.27 ±0.35
Decision Tree	0.34 ±0.33	0.14 ±0.30	**0.45 ±0.44**
SVM	0.46 ±0.40	0.23 ±0.40	0.43 ±0.46
Random Forest	0.33 ±0.34	0.07 ±0.21	0.39 ±0.43
XGBoost	0.34 ±0.35	0.19 ±0.35	0.27 ±0.33

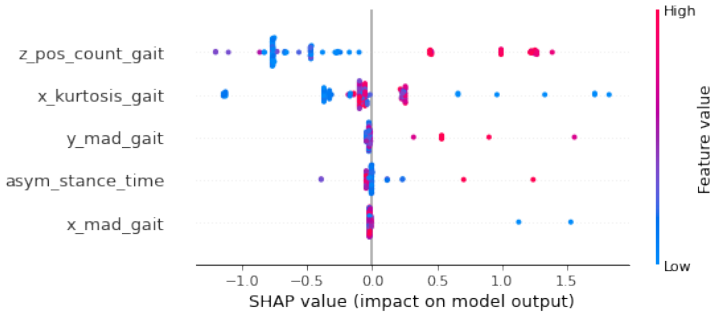
Table 5.3: Multi-class (4 classes) classification performance on the test set assessed with averaged balanced accuracy scores (± standard deviation) across all folds. Items in bold represent the top performing model on average within a task category. The Majority Vote classifier serves as the baseline model for model comparison.

Pull test recovery ("Aided Recovery" or "Unaided Recovery") prediction using binary classification can be found in Table 5.4. Once the complexity of the task was reduced (reducing 4 classes to 2 classes), model balanced accuracy performance increased compared to multi-class classification. This is an unsurprising finding as the prediction task was simplified. Overall, all reported standard deviations are high regardless of the model and task. This indicates high patient performance variability. All gait models outperformed all respective sway models in prediction of pull test recovery. All combined task models outperformed all gait models and sway models respectively. Therefore, combined task models were the most predictive for pull test recovery. The highest performing model for the gait task was the XGBoost. The highest performing model for the sway task was K-Nearest Neighors. The highest performing model for the combined task and all tasks was the XGBoost combined task model with the highest balanced accuracy score and lowest standard deviation. Logistic Regression, Decision Tree and XGBoost combined task models were able to outperform the Majority Vote classification model. The binary classification models will be analyzed in subsequent results due to their high results.

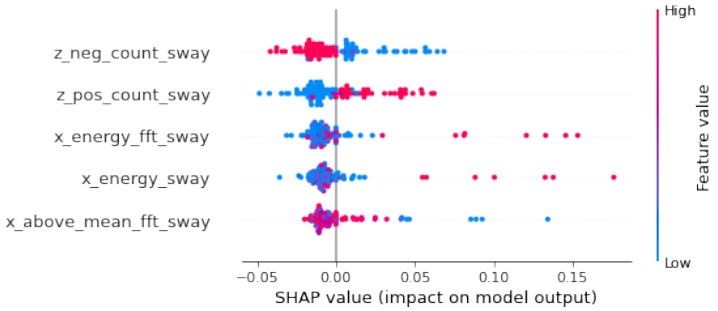
	Walking	**Standing**	**Combined**
Majority Vote	0.75 ±0.43	0.75 ±0.43	0.75 ±0.43
Logistic Regression	0.68 ±0.37	0.51 ±0.49	0.78 ±0.38
Naive Bayes	0.64 ±0.35	0.38 ±0.46	0.69 ±0.39
K-Nearest Neighbors	0.71 ±0.38	**0.69 ±0.43**	0.74 ±0.38
Decision Tree	0.67 ±0.37	0.40 ±0.45	0.80 ±0.39
SVM	0.70 ±0.39	0.56 ±0.46	0.70 ±0.42
Random Forest	0.65 ±0.39	0.52 ±0.47	0.67 ±0.45
XGBoost	**0.71 ±0.36**	0.33 ±0.44	**0.84 ±0.34**

Table 5.4: Binary classification performance on the test set assessed with averaged balanced accuracy scores (± standard deviation) across all folds. Items in bold represent the top performing model on average within a task category. The Majority Vote classifier serves as the baseline model for model comparison.

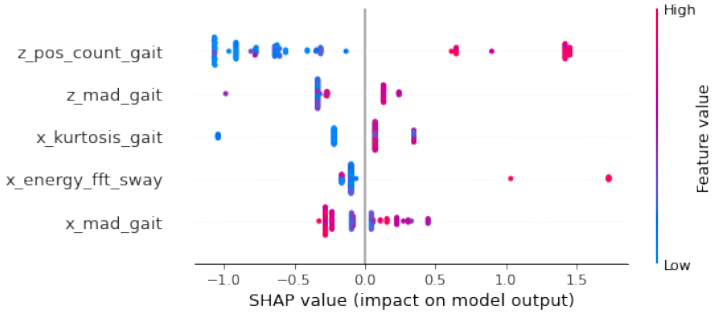
The use of feature importance can increase model interpretablity. To assess feature importance the SHAP values for each of the highest performing models in the walking, standing and combined tasks were calculated. The 5 most predictive features of each model are displayed using a SHAP summary plot in Figure 5.2. Features are ranked in terms of importance in model performance. Most impactful features are found at the top of the plot, and decrease in impactfulness with each subsequent row. Each point in the plots is a SHAP value for an instance of that feature (1 sample). High SHAP values indicate high model impact. The most impactful features for the walking models were the positive acceleration in the z-axis, kurtosis in the time domain in the x-axis, mean absolute deviation in the y-axis, asymmetrical stance time, and mean absolute deviation in the x-axis. The most impactful features for the standing task assessments are the negative acceleration in the z-axis, positive acceleration in the z-axis, energy in the frequency domain in the x-axis, energy of the time domain signal in the x-axis, and the values above the mean in the frequency domain in the x-axis. For the combined task model the most impactful features were the z-axis positive acceleration in gait, the z-axis mean absolute deviation in gait, the x-axis kurtosis in the time domain for gait, the x-axis energy in the frequency domain in sway and the x-axis mean absolute deviation for gait. Overall, the count of positive and negative acceleration in the z-axis (medial-lateral) direction were the most impactful features. In the combined task model 4 of the top 5 most impactful features were gait tasks, with only one sway task.



(a)



(b)



(c)

Figure 5.2: a) Walking task feature impactfulness assessed with the XGBoost model. b) Standing task feature impactfulness assessed with a K-Nearest Neighbor. c) Combined task feature impactfulness assessed with the XGBoost model. Features in the top row are the most impactful and decrease in impactfulness with every subsequent row. Each point in the SHAP summary plot represents one instance of a feature value and is colored pink for high values and blue for low values. The x-axis represents the shap value, where values away from 0 are more impactful than points close to 0.

Post-hoc model evaluation was completed to compare the statistical significance of model results. A paired t-test was completed to assess whether model results across all folds were statistically different than the performance of the Majority Vote Classifier. No models reported a statistically significant difference in performance. In addition, model performance was evaluated using the Nemenyi test to assess whether models were significant from one another in their performance across folds. Results for model significance can be seen in Figure 5.3. The results of the evaluation indicate that no models performed statistically different from one another in their averaged ranked performance on the individual folds of the dataset across all tasks.

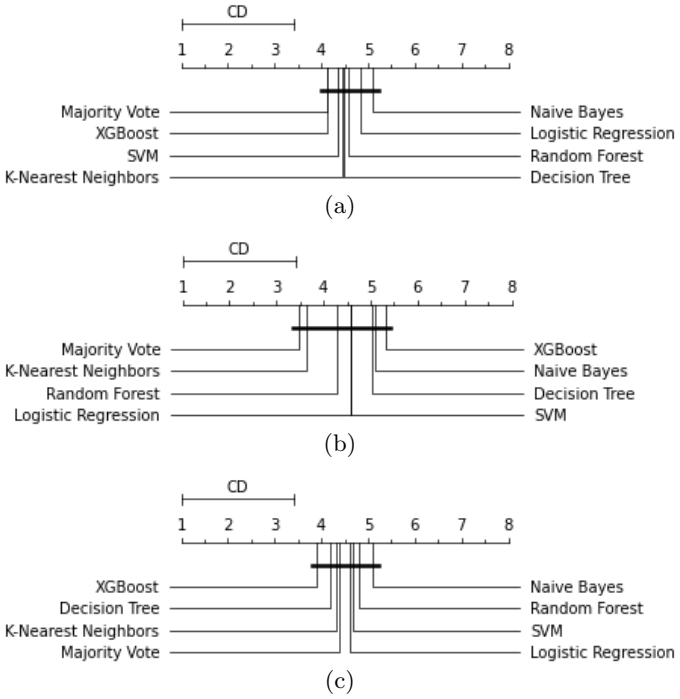


Figure 5.3: Nemenyi Test to assess model performance in prediction using (a) walking tasks,(b) standing tasks and (c) combined tasks. The critical distance for all three task model performance is CD = 2.41. Models connected by horizontal bars did not perform significantly different in their CV evaluation.

6. Discussion

The three model types (regression, multi-class classification and binary classification) showed variable performance on the pull test score prediction task. Overall, models did not generalize to new patients and had high patient performance variability. In regression analysis, models were overfitting to the dataset. In classification models, patients were likely to be classified correctly for all tasks (1.00 accuracy) or misclassified for all tasks (0.00 accuracy), contributing to the high model standard deviations. Of the three model types, the simplest task, the binary classification model for pull test recovery prediction, had the most promising results. However, it is not possible to confidently answer the research questions outlined at the beginning of the analysis (*Can gait, sway or a combination of both task features predict pull test results?*) due to a small dataset and a lack of statistical significance.

The MJFF Levodopa Response Study is limited in the number of patients with a full MDS-UPDRS evaluation and clean data recordings (n=16). This restricts the ability to design a representative patient model for both regression and classification tasks. All regression models showed signs of overfitting on the training dataset. However, hyperparameter optimization to reduce model overfitting was not possible due to the small dataset size. In classification model development, the small dataset size was exacerbated by the unbalanced pull test scores. There were only 12 tasks in the dataset (2 patients) who received a pull test score of a 2. In multi-class classification, the model had two few instances of the 2 class to train a classifier. In binary classification, the transition from "Unaided Recovery" to "Aided Recovery" occurs at the transition from pull test score 2 to pull test score 3. A model with too few patients with a pull test score 2 may not fully capture the patient groups.

Overall, the LOSO CV produced volatile results. One interpretation of this result is that the dataset did not effectively capture patient variation that may be present in PD. As a result, incoming data from a patient with an unseen physiology, say a unique gait pattern, is not able to classify correctly as it has not been seen in model training. In this analysis, removing one patient from the dataset has major impact on model training and performance. More data is needed to train models on different PD physiologies representative of the population.

In addition to limitations to model creation and model training, the feature engineering process had drawbacks. The MJFF Levodopa Response Study aims to capture free-living task activities in PD patients with accelerometers. Although the study provides timestamps for select task activities, it lacks a second form of validation for noisy accelerometer signal, such as video data or pressure insole recordings. Therefore, the gait cycle features that were derived from accelerometry signal

in the dataset could not be validated against a second source of data collection, a"ground truth" measurement. PD patients have a slew of gait related motor impairments that can cause the already noisy accelerometry signal to become increasingly difficult to interpret. For example, PD patients may exhibit shuffling of the feet as a motor impairment. Analyzing data from a segment of gait with feet shuffling is a complicated task as the foot does not follow the same heel to toe arc that allows for easy identification of HS and TO events. To mitigate the lack of a ground truth, only methods previously validated against gold standards and tested with PD patients were implemented.

Although the results of this analysis did not produce a statistically significant model for the prediction of pull test recovery, two result findings are worth further investigation: The combined task model for pull test recovery prediction and trends in gait and sway trends in respect to pull test balance. Despite the lack of statistical significance, the combination task XGBoost pull test recovery prediction model on average outperformed all comparative classifiers including the Majority Vote Classifier. This shows potential for the development of a model that can analyze walking and standing segments for the prediction of pull test recovery. The development of such a tool would bring benefit to a patient and the clinical team. An accelerometer is a non-invasive, portable device that can be given to a patient for continuous wear (many patients unknowingly own accelerometers in their daily devices embedded in phones and watches). Implementing the pull test score predictor model on a continuous data stream would allow for frequent estimation of a patient's postural stability and daily living independence. When model estimations yield results that a patient's postural stability may be compromised, a patient and their caretaker could be notified for early intervention. A postural instability detector may help save a patient from a fall or injury.

Another finding of the analysis is the trend in task prediction capability. Gait derived features from walking tasks regularly outperformed sway derived features from standing tasks in the prediction of pull test recovery. This gives insight that gait features may be more predictive features of balance than sway features. Furthermore, gait and balance may be more connected than initially presumed. However, the most optimal results from the pull test recovery prediction task were a combination of gait and sway features. If the pull test is representative of balance, then the best performing model uses features from both dynamic balance (gait) and static balance (sway). However, there is currently not enough evidence to indicate that the best estimate of balance in PD is the pull test. The pull test may only be representative of one form of balance, perturbed balance. Further research needs to be completed beyond this study to understand how balance in PD should be defined. Subsequently, research can be completed to understand the relationship between gait, sway and balance in PD.

The relationship between gait, sway and balance using accelerometry is an underexplored field both in PD patients and in healthy subject. To better understand the research concept, it may be useful to reduce the problem complexity and first study the relationship of variables with healthy controls. Therefore, future work is the development of a study to investigate gait, sway and balance in healthy controls.

7. Future Work: A Study on Balance in Healthy Controls

Using the MJFF Levodopa Response Study, walking and standing tasks where used to estimate pull test results. However, due to the small number of participants and the physiological variation of PD patients, a conclusion could not be drawn. To investigate the concept of gait, balance and sway further, a simplification of the problem is necessary. The development of a study using accelerometers in healthy adults may be beneficial in furthering balance research.

A study investigating gait, sway and balance will help investigate the following questions further: *Is it possible to estimate a subject's balance from a gait segment using an accelerometer? Can we categorize balance severity? Is there a relationship between dynamic balance and static balance have a connection?*

An exploratory study in gait, sway and balance in healthy individuals was developed for the Centre for Human Drug Research (CHDR) in Leiden, Netherlands. CHDR is an organization conducting first in human clinical drug trials. In addition to conducting human trials, CHDR investigates tools for hybrid and remote patient monitoring for clinical trials. A portable tool to understand gait, sway and balance poses benefit for hybrid patient monitoring in drug studies. A synopsis of the study protocol is presented.

Study Synopsis

Title : An Exploration of Static Balance and Dynamic Gait Balance in Healthy Controls with Simulated Impaired Balance Measured Using Accelerometers and BodySway as a Gold Standard.

Background Rationale:
Human balance is an intricate system composed of three major components: vision, vestibular control, and proprioception. The three mechanisms of balance comprise an integral portion of human locomotion. Balance disruption via one of the three sub-mechanisms, or multiple mechanisms, can manifest as changes in gait. Alterations in gait due to imbalance can result in falls and serious injury. Medication side-effects may result in balance impairments. Instruments such as the BodySway are used as a proxy for measurements of postural control and balance. The BodySway test provides a measure of total body displacement of a subject during standing sway with eyes closed. Although the BodySway test is a useful tool for monitoring postural control measures and the effects of medication on balance, it is not a portable tool and does not give a measure of dynamic balance. This study is important as it will provide a tool for measuring both static and dynamic balance in patients. The study will investigate balance measures using accelerometers as they are affordable and portable sensors. Accelerometers may be sent home with subjects for at home continuous monitoring of metrics of interests. A surrogate method for the BodySway task will be developed using accelerometer measurements in both static and dynamic gait balance. Further, changes in gait due to balance impairment will be studied and quantified. Finally, the relationship between dynamic and static balance will be investigated. An at home method for the measurement of static and dynamic gait balance provides researchers with more data points and variation in data type for measurements of balance. This method can be used in the investigation of drugs and balance impairment over a continuous time period in at home setting.

Objectives:
Primary Objectives:

- Develop a surrogate measure for BodySway using accelerometers.

Secondary Objectives:

- Quantify the ability of the surrogate measure to distinguish between "balanced" and "artificially imbalanced" healthy participants during a short walk.
- Quantify the magnitude of balance impairment in dynamic gait balance.
- Investigate the relationship between static balance and dynamic gait balance.

Design:

The study will be a cross-sectional study.

The study will be comprised of 20 healthy subjects with no known gait or balance disorders. Subjects will be instructed to complete balance and gait tasks. Subjects will complete the balance and gait tasks while having six accelerometers attached to points on the body: the lumbar, the sternum, the left ankle, right ankle, left wrist and right wrist. Subjects will complete the balance

and exercise tasks with no balance impairment. Subjects will then complete the balance and gait tasks with different forms of balance impairment using tools such as foam mats for proprioception impairment, vision impairment goggles and spinning of the vestibular system. Subjects will repeat the experimental procedure on two separate days.

Investigational Technologies:
Accelerometers:
Six accelerometers will be attached to a subject on the lumbar, sternum, left ankle, right ankle, left wrist and right wrist.

Subjects/Groups:
20 healthy controls with no known balance or gait impairments.

Main inclusion criteria:

- Healthy condition
- Can walk 20+ meters consecutively
- Age 18 years or older

Main exclusion criteria

- Inability to give consent
- Gait or balance impairments
- Self-report of drug use that may affect gait or balance
- Current or past injury that may affect balance or gait
- Other indications of impaired gait or balance as observed by the researcher

Endpoints:
To develop a method for the calculation of a surrogate BodySway measure the following features will be used:

- BodySway scores
- Accelerometer features from static balance:
  - Jerk
  - Distance
  - RMS
  - Path
  - Mean Velocity
  - Area
  - Total Power
  - F50 Frequency range
  - F95 Frequency range
  - Centroid frequency
  - Frequency dispersion

Gait features derived from accelerometers during healthy and imbalanced gait:

- Gait statistics:
  - Average stride time

  - Average step time
  - Average stance time
  - Average single support time
  - Average double support time
  - Average swing time
  - Variability between gait cycles:
  - Stride time variability
  - Step time variability
  - Stance time variability
  - Single Support time variability
  - Double Support time variability
  - Swing time variability
  - Measures of asymmetry between left and right foot:
  - Stride time asymmetry
  - Step time asymmetry
  - Stance time asymmetry
  - Swing time asymmetry
  - Single support time asymmetry
- Other accelerometer derived data during gait:
  - Pitch
  - Roll

Sample Size Justification:
Since this study is exploratory, sample size is not based on power calculations.

Statistical methodology:
The BodySway Score replication using accelerometer data will be calculated to minimize error. To classify subject gait as healthy or imbalanced a machine learning model will be developed. Random tree classifiers and gradient boosted trees will be implemented. The models will be developed in a manner such as to minimize error without overfitting. Hyperparameter optimization will be used for model tuning. Cross Validation will be used to minimize the effects of overfitting. In addition, other classifier methods will also be investigated such as logistic regression and Knearest Neighbor. Statistical methods, such as the sensitivity, specificity, positive predictive value(PPV) and negative predictive value (NPV) will be calculated. Receiver Operating Characteristics (ROC) curves will be implemented to find an optimal model that minimizes model specificity and sensitivity. For the calculation of the magnitude of imbalance that may be present in gait, the previous methods will be repeated. First, a classification task with a scale of 0-3 will be created to classify different levels of imbalance. The same statistical methodology will follow. We will create multiple machine learning models to investigate the relationship between dynamic walking gait and static gait. Regression models will be tested such as Random Forest regression and boosted trees. The models will be developed in a manner such as to minimize error without overfitting. Hyperparameter optimization will be used for model tuning. Cross Validation will be used to minimize the effects of overfitting. In addition, a neural network prediction task will be developed with gait sequence as input and the score of the BodySway task as an output.

Task	**SCR (-10 days)**	**Visit 1**	**Visit 2**
Informed consent	X		
Inclusion and exclusion criteria	X		
Medical history	X		
Collect Physical Measurements	X		
Instructions		X	X
Accelerometer Setup		X	X
Walk 20m, Normal Gait		X	X
BodySway Task, Eyes Open		X	X
BodySway Task, Eyes Closed		X	X
BodySway Task, Eyes Closed and Foam Mat		X	X
BodySway Task, Vision Impairment 1		X	X
BodySway Task, Vision Impairment 2		X	X
BodySway Task, Vision Impairment 3		X	X
Walk 20m, Impaired Vision 1		X	X
Walk 20m, Impaired Vision 2		X	X
Walk 20m, Impaired Vision 3		X	X
5 Turns in a Rotating Chair		X	X
BodySway Task, Vestibular Balance Impairment		X	X
5 Turns in a Rotating Chair		X	X
Walk 20m, Vestibular Balance Impairment		X	X
Accelerometer Removal		X	X
Discharge	X	X	X

Table 7.1: Table 1: Visit and Assessment Schedule

8. Conclusion

The pull test is an integral assessment for understanding a patient's physical independence, daily living capabilities and fall risk. However, the pull test is only performed in clinic, offering infrequent measures of a patients postural instability. Modeling the pull test using accelerometer data from at-home daily tasks, such as walking and standing, would provide patients and physicians with a frequent assessment of postural instability and an early warning for health and safety risk. The pull test is an ordinal scale and can therefore be modeled as a regression task and a multi-class classification task. The pull test can also be simplified into two categories: "Unaided Recovery" and "Aided Recovery". Although select models outperformed the Majority Vote Classifier on average, no model performance was statistically significantly different than remaining models. All models exhibited large patient performance variation. Overall, these results suggest that further investigation with at-home accelerometer tasks for postural control is required with a larger patient pool better representative of PD patient physiology. In addition, investigation of gait, sway and balance in healthy controls would be a way to reduce the complexity of the problem statement and develop a foundation to understand balance.

Acknowledgements

I would like to thank my supervisors Arno Knobbe and Vasileios Exadaktylos for their guidance, continued support and patience throughout the entire thesis adventure.

I would also like to thank my honorary supervisor, Eleni Kapousidou, for her assistance in the thesis process. She was available to discuss all questions and findings, big or small.

Ahnjili Zhuparris was an integral mentor in the thesis completion. She selflessly offered her time and assistance. For every answer she provided to a question, she also provided three brilliant ideas for future analysis.

I am grateful for the assistance from Jacobus J. Van Hilten and Roel Weijer. Their assistance throughout this project helped bridge the gap between data science and medicine. Similarly, I would like to thank Ingrid Koopmans for answering any and all questions concerning PD assessment. The opportunity to meet patients with PD helped me understand the scope of the problem statement.

Special thanks to Bart de Zoete who offered thesis guidance and support when he also had his own thesis to write.

Finally, I would like to thank CHDR for allowing me the opportunity to work on a unique project in a dynamic and exciting environment.

Bibliography

[1] Christopher G. Goetz, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Glenn T. Stebbins, Matthew B. Stern, Barbara C. Tilley, Richard Dodel, Bruno Dubois, Robert Holloway, Joseph Jankovic, Jaime Kulisevsky, Anthony E. Lang, Andrew J. Lees, Sue Leurgans, Peter A. LeWitt, David Nyenhuis, C. Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A. Teresi, Jacobus J. van Hilten, and Nancy LaPelle. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Process, format, and clinimetric testing plan. *Movement Disorders*, 22(1):41–47, 2007.

[2] Roongroj Bhidayasiri and Daniel Tarsy. Parkinson's disease: Hoehn and yahr scale. *Current Clinical Neurology*, pages 4–5, 2012.

[3] Verónica Robles-García, Yoanna Corral-Bergantiños, Nelson Espinosa, María Amalia Jácome, Carlos García-Sancho, Javier Cudeiro, and Pablo Arias. Spatiotemporal gait patterns during overt and covert evaluation in patients with parkinson's disease and healthy subjects: Is there a hawthorne effect? *Journal of Applied Biomechanics*, 31(3):189–194, 2015.

[4] Rosie Morris, Aodhán Hickey, Silvia Del Din, Alan Godfrey, Sue Lord, and Lynn Rochester. A model of free-living gait: A factor analysis in parkinson's disease. *Gait amp; Posture*, 52:68–71, 2017.

[5] Ihana Thaís Guerra de Oliveira Gondim, Caroline de Cássia Batista de Souza, Marco Aurélio Benedetti Rodrigues, Izaura Muniz Azevedo, Maria das Graças Wanderley de Sales Coriolano, and Otávio Gomes Lins. Portable accelerometers for the evaluation of spatio-temporal gait parameters in people with parkinson's disease: an integrative review. *Archives of Gerontology and Geriatrics*, 90:104097, 2020.

[6] Libak Abou, Joseph Peters, Ellyce Wong, Rebecca Akers, Mauricette Sènan Dossou, Jacob J. Sosnoff, and Laura A. Rice. Gait and balance assessments using smartphone applications in parkinson's disease: A systematic review. *Journal of Medical Systems*, 45(9), 2021.

[7] Lazzaro di Biase, Alessandro Di Santo, Maria Letizia Caminiti, Alfredo De Liso, Syed Ahmar Shah, Lorenzo Ricci, and Vincenzo Di Lazzaro. Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, 20(12):3529, 2020.

[8] Adriana Menezes Degani, Vinicius Saura Cardoso, Alessandra Tanuri Magalhães, Ana Larissa Sousa Assunção, Erica de Carvalho Soares, and Alessander Danna-dos Santos. Postural behavior in medicated parkinson disease patients: A preliminary study searching for indicators to track progress. *Journal of Central Nervous System Disease*, 12:117957352092264, 2020.

[9] Martina Mancini, Patricia Carlson-Kuhta, Cris Zampieri, John G. Nutt, Lorenzo Chiari, and Fay B. Horak. Postural sway as a marker of progression in parkinson's disease: A pilot longitudinal study. *Gait amp; Posture*, 36(3):471–476, 2012.

[10] Anna Frenklach, Stephanie Louie, Mandy Miller Koop, and Helen Bronte-Stewart. Excessive postural sway and the risk of falls at different stages of parkinson's disease. *Movement Disorders*, 24(3):377–385, 2008.

[11] L Rocchi. Effects of deep brain stimulation and levodopa on postural sway in parkinson's disease. *Journal of Neurology, Neurosurgery amp; Psychiatry*, 73(3):267–274, 2002.

[12] Fredy J. Revilla, Travis R. Larsh, Ashutosh Mani, Andrew P. Duker, Cyndy Cox, Paul Succop, Maureen Gartner, Claudia Jarrin Tejada, and Amit Bhattacharya. Effect of dopaminergic medication on postural sway in advanced parkinson's disease. *Frontiers in Neurology*, 4, 2013.

[13] Siddharth Arora, Vinayak Venkataraman, Sean Donohue, Kevin M. Biglan, Earl R. Dorsey, and Max A. Little. High accuracy discrimination of parkinson's disease participants from healthy controls using smartphones. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[14] Ryan P. Hubble, Geraldine A. Naughton, Peter A. Silburn, and Michael H. Cole. Wearable sensor use for assessing standing balance and walking stability in people with parkinson's disease: A systematic review. *PLOS ONE*, 10(4):e0123705, 2015.

[15] Cristian F. Pasluosta, Jens Barth, Heiko Gassner, Jochen Klucken, and Bjoern M. Eskofier. Pull test estimation in parkinson's disease patients using wearable sensor technology. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

[16] Walter Pirker and Regina Katzenschlager. Gait disorders in adults and the elderly. *Wiener klinische Wochenschrift*, 129(3-4):81–95, 2016.

[17] Jeong-Ho Park, Yeo-Jeong Kang, and Fay Bahling Horak. What is wrong with balance in parkinson's disease? *Journal of Movement Disorders*, 8(3):109–114, 2015.

[18] Bastiaan R. Bloem, Johan Marinus, Quincy Almeida, Lee Dibble, Alice Nieuwboer, Bart Post, Evzen Ruzicka, Christopher Goetz, Glenn Stebbins, and Pablo et al. Martinez-Martin. Measurement instruments to assess posture, gait, and balance in parkinson's disease: Critique and recommendations. *Movement Disorders*, 31(9):1342–1355, 2016.

[19] Karolina Krzysztoń, Jakub Stolarski, and Jan Kochanowski. Evaluation of balance disorders in parkinson's disease using simple diagnostic tests—not so simple to choose. *Frontiers in Neurology*, 9, 2018.

[20] Fay B Horak, Diane M Wrisley, and James Frank. The balance evaluation systems test (bestest) to differentiate balance deficits. *Physical Therapy*, 89(5):484–498, 2009.

[21] Katherine Berg. Measuring balance in the elderly: preliminary development of an instrument. *Physiotherapy Canada*, 41(6):304–311, 1989.

[22] Katherine Berg. Measuring balance in the elderly: preliminary development of an instrument. *Physiotherapy Canada*, 41(6):304–311, 1989.

[23] Mary E. Tinetti, T. Franklin Williams, and Raymond Mayewski. Fall risk index for elderly patients based on number of chronic disabilities. *The American Journal of Medicine*, 80(3):429–434, 1986.

[24] Breelan M. Kear, Thomas P. Guck, and Amy L. McGaha. Timed up and go (tug) test. *Journal of Primary Care amp; Community Health*, 8(1):9–13, 2016.

[25] info@sagebase.org Sage Bionetworks. Mjff levodopa response study, 2019.

[26] Shimmer. Shimmer wearable sensor technology, 2022.

[27] Gloria Vergara-Diaz, Jean-Francois Daneault, Federico Parisi, Chen Admati, Christina Alfonso, Matilde Bertoli, Edoardo Bonizzoni, Gabriela Ferreira Carvalho, Gianluca Costante, and Eric Eduardo et al. Fabara. Limb and trunk accelerometer data collected with wearable sensors from subjects with parkinson's disease. *Scientific Data*, 8(1), 2021.

[28] R. Moe-Nilssen. A new method for evaluating motor control in gait under real-life environmental conditions. part 1: The instrument. *Clinical Biomechanics*, 13(4-5):320–327, 1998.

[29] The SciPy community. "scipy.signal.detrend", 2022.

[30] John McCamley, Marco Donati, Eleni Grimpampi, and Claudia Mazzà. An enhanced estimate of initial contact and final contact instants of time using lower trunk inertial sensor data. *Gait amp; Posture*, 36(2):316–318, 2012.

[31] 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 amp; Physics*, 37(4):400–407, 2015.

[32] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C.J. Bula, and P. Robert. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. *IEEE Transactions on Biomedical Engineering*, 50(6):711–723, 2003.

[33] Minh H. Pham, Morad Elshehabi, Linda Haertner, Silvia Del Din, Karin Srulijes, Tanja Heger, Matthis Synofzik, Markus A. Hobert, Gert S. Faber, and Clint et al. Hansen. Validation of a step detection algorithm during straight walking and turning in patients with parkinson's disease and older adults using an inertial measurement unit at the lower back. *Frontiers in Neurology*, 8, 2017.

[34] Ervin Sejdic, Kristin A. Lowry, Jennica Bellanca, Mark S. Redfern, and Jennifer S. Brach. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(3):603–612, 2014.

[35] 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, 2016.

[36] Martina Mancini, Arash Salarian, Patricia Carlson-Kuhta, Cris Zampieri, Laurie King, Lorenzo Chiari, and Fay B Horak. Isway: a sensitive, valid and reliable measure of postural control. *Journal of NeuroEngineering and Rehabilitation*, 9(1):59, 2012.

[37] Rigoberto Martinez-Mendez, Masaki Sekine, and Toshiyo Tamura. Postural sway parameters using a triaxial accelerometer: comparing elderly and young healthy adults. *Computer Methods in Biomechanics and Biomedical Engineering*, 15(9):899–910, 2012.

[38] Scikit-Learn Developers. Supervised learning, 2022.

[39] XGBoost Developers. Xgboost documentation, 2021.

[40] Leonardo Medeiros, Hyggo Almeida, Leandro Dias, Mirko Perkusich, and Robert Fischer. A gait analysis approach to track parkinson's disease evolution using principal component analysis. *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, 2016.

[41] Yunfeng Wu and Sridhar Krishnan. Statistical analysis of gait rhythm in patients with parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(2):150–158, 2010.

[42] Ganggang Xu and Jianhua Z. Huang. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics*, 40(6), 2012.

[43] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview, 2020.

[44] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[45] SHAP. an introduction to explainable ai with shapley values — shap latest documentation 2022, 2019.