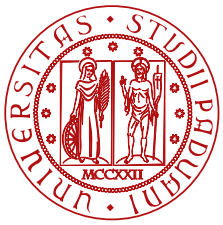


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITY OF PADUA

DEPARTMENT OF INFORMATION ENGINEERING

MASTER'S DEGREE IN BIOENGINEERING

Wearable-Based Assessment of Parkinson's Disease: insights from the Verily Study Watch Dataset

Supervisor
Prof. Mattia Veronese

Co-supervisor
Dr. Emma Roveroni

Candidate
Klaudio Curumi
Student ID:2148337

ACADEMIC YEAR 2025-2026

Graduation Date 20/04/2026

Abstract

Parkinson's Disease (PD) is a complex neurodegenerative disorder traditionally monitored through episodic and subjective clinical assessments. The advent of wearable technologies offers a paradigm shift, enabling the objective and continuous evaluation of motor symptoms directly in free-living conditions.

The primary objective of this thesis is to develop and validate a comprehensive methodological framework for the extraction, selection, and classification of digital motor biomarkers derived from raw inertial data acquired through a wrist-worn sensor. Specifically, the study aims to evaluate the concurrent validity of these features and to evaluate their ability to accurately classify PD patients from healthy controls, thereby contributing to the development of a robust continuous monitoring strategy.

To achieve this goal, real-world data from the Parkinson's Progression Markers Initiative (PPMI) wearable sub-study were analyzed. The implemented pipeline involves the extraction of multidimensional features across key motor domains, including tremor, bradykinesia, gait kinematics, and overall physical activity. Following a rigorous statistical screening designed to isolate the most reliable and discriminative metrics, the selected features were used to train and validate interpretable Machine Learning models. Finally, correlation analyses with standardized clinical scores were conducted to evaluate the clinical relevance of the selected biomarkers.

The results demonstrate the effectiveness of the proposed approach: the developed pipeline successfully identified a coherent digital motor signature capable of discriminating PD patients from HC with high accuracy. Beyond raw classification performance, this works translated the extracted metrics into clinical insights: by demonstrating associations between specific inertial features and established clinical scores, the study shows that these digital biomarkers reflect the real world severity of core PD motor domains. In conclusion, this work demonstrates that the integration of wrist-worn sensors with a rigorous computational pipeline may represent a reliable system for continuous, objective, and personalized monitoring of PD.

Sommario

La malattia di Parkinson (PD) è un disturbo neurodegenerativo complesso, tradizionalmente monitorato attraverso valutazioni cliniche episodiche e soggettive. L'avvento delle tecnologie indossabili offre un grande cambiamento, consentendo la valutazione oggettiva e continua dei sintomi motori direttamente in condizioni di vita quotidiana .

L'obiettivo primario di questa tesi è sviluppare e validare un quadro metodologico completo per l'estrazione, la selezione e la classificazione di biomarcatori motori digitali derivati da dati inerziali grezzi, acquisiti tramite un sensore da polso. Nello specifico, lo studio mira a valutare la validità concorrente di tali caratteristiche e la loro capacità di classificare accuratamente i pazienti affetti da Parkinson rispetto ai controlli sani, contribuendo così allo sviluppo di una strategia di monitoraggio continuo e robusta.

Per raggiungere questo obiettivo, sono stati analizzati dati reali provenienti dal sotto-studio sui dispositivi indossabili della Parkinson's Progression Markers Initiative (PPMI). La pipeline implementata prevede l'estrazione di caratteristiche multidimensionali attraverso i principali domini motori, tra cui tremore, bradicinesia, cinematica del cammino e attività fisica globale. In seguito a uno rigoroso screening statistico volto a isolare le metriche più affidabili e discriminanti, le caratteristiche selezionate sono state utilizzate per addestrare e validare modelli di Machine Learning interpretabili. Infine, sono state condotte analisi di correlazione con i punteggi clinici standardizzati per valutare la rilevanza clinica dei biomarcatori selezionati.

I risultati dimostrano l'efficacia dell'approccio proposto: la pipeline sviluppata ha identificato con successo una firma motoria digitale coerente, in grado di distinguere i pazienti PD dai controlli sani con un'elevata accuratezza. Al di là delle prestazioni di classificazione, questo lavoro ha tradotto le metriche estratte in approfondimenti clinici: dimostrando l'associazione tra specifiche caratteristiche inerziali e i punteggi clinici stabiliti, lo studio evidenzia come questi biomarcatori digitali riflettano la gravità reale dei principali domini motori della malattia. In conclusione, questo lavoro dimostra che l'integrazione di sensori da polso con una rigorosa pipeline computazionale può rappresentare un affidabile sistema per il monitoraggio continuo e oggettivo del Parkinson.

Contents

Abstract ii

Sommario iv

List of Figures x

List of Tables xi

List of Acronyms xii

1 Introduction 1

1.1 Parkinson's disease . . . 1
1.2 Digital biomarkers for Parkinson's disease . . . 2
1.2.1 Disease monitoring . . . 3
1.2.2 Therapy optimization . . . 3
1.2.3 Ecological validity . . . 3
1.3 Open questions about digital biomarkers in Parkinson's disease . . . 4
1.4 Aim of the thesis . . . 5
1.5 Outline of the thesis . . . 5

2 Parkinson's Progression Markers Initiative 7

2.1 The PPMI wearable sub-study . . . 7
2.1.1 Sensor specifications and capabilities . . . 7
2.1.2 Research potential . . . 8
2.2 Study cohort and dataset . . . 8
2.2.1 The Verily Study Watch . . . 10

3 Data completeness 12

3.1 Extraction of data completeness . . . 12
3.2 Analysis of hourly consistency . . . 13
3.3 Analysis of daily consistency . . . 15
3.4 Completeness group comparison of adherence profiles . . . 17

3.4.1 Mean comparison . . . 17
3.4.2 Variance comparison . . . 18
3.5 Temporal effects on data completeness . . . 19

4 Feature extraction **21**
4.1 Gait metrics . . . 22
4.1.1 Step count algorithm validation . . . 23
4.2 Tremor metrics . . . 25
4.3 Bradykinesia and hypokinesia metrics . . . 26
4.3.1 ParaDigMa framework . . . 27
4.3.2 Mahadevan et al. algorithm . . . 27
4.4 Global activity profile . . . 28

5 Feature selection **30**
5.1 Reliability analysis . . . 31
5.2 Redundancy and multicollinearity assessment . . . 32
5.3 Integrated reliability and redundancy analysis . . . 33
5.4 Temporal stability of extracted features . . . 37
5.5 Final set of features . . . 37

6 Discriminative statistical analysis **39**
6.1 Effect size estimation . . . 40
6.2 Discriminative power and diagnostic value . . . 42
6.2.1 Mann-Whitney U-test for feature comparison . . . 43
6.2.2 Receiver Operating Characteristic (ROC) analysis . . . 44
6.2.3 Integrated interpretation of statistical findings . . . 45
6.3 Intra-week temporal stability . . . 47
6.4 Multivariate group comparison (MANOVA) . . . 49

7 Machine learning for group classification **51**
7.1 Hybrid machine learning approach . . . 51
7.2 Machine learning algorithms for classification . . . 53
7.2.1 Random Forest . . . 53
7.2.2 Linear Support Vector Machine . . . 55
7.2.3 Support Vector Machine with Radial Basis Function . . . 56
7.2.4 K-Nearest Neighbors . . . 57
7.2.5 Logistic Regression . . . 57
7.2.6 Extreme Gradient Boosting . . . 58
7.3 Comparative analysis of machine learning models . . . 60
7.3.1 Accuracy and model complexity . . . 63

7.3.2 Sensitivity-specificity trade off 63
7.3.3 Receiver Operating Characteristic curve analysis 65
7.3.4 Model selection 66

8 Concurrent validity and clinical associations of digital biomarkers 67
8.1 Correlation with standardized clinical scales 68
8.1.1 Step cadence as a proxy for functional disability (MDS-UPDRS II) . . 68
8.1.2 Predictors of motor severity (MDS- UPDRS III ON) 71
8.2 Integrated discussion 72

9 Conclusions 74
9.1 Key findings 74
9.2 Study limitations and future directions 76
9.3 Final remarks 77

Bibliography 79

List of Figures

1.1 Analytical pipeline of the thesis 6

2.1 The Verily Study Watch 10

3.1 Evolution of data completeness in two participants 13
3.2 Comparison of data completeness between the 07:00–11:00 and 11:00–22:00 time windows 14
3.3 Best and worst case of data completeness profiles 16
3.4 Comparison of data completeness between the groups 18
3.5 Boxplots illustrating the distribution of completeness ratios for PD and HC groups across the 11:00-22:00 interval. 19

5.1 Visualization of the global mean of the features during the week 32
5.2 Heatmap of the correlation matrix of the features 34
5.3 Scatterplot of ICC versus maximum absolute correlation 36

6.1 Effect size calculated between the PD and HC groups 41
6.2 Visualization of the four highest AUC. 46
6.3 Results of the multivariate analysis of variance (MANOVA) 50

7.1 Variable importance ranking with the random forest approach. 54
7.2 Visualization of the performance of the algorithm of random forest 54
7.3 Feature importance obtained from the SVM algorithm. 56
7.4 Visualization of the Odds Ratio of the logistic regression algorithm. 58
7.5 Feature importance provided by the XGBoost algorithm. 59
7.6 Heatmap of the variable importances of the algorithms. 61
7.7 Radar chart comparing the standardized motor features of patient 3119 against the average profile of HC and PD. 65

8.1 Spearman correlation of the final set of features 69
8.2 Visualization of the correlation between step cadence and mean bout duration with the MDS-UPDRS Part 2 score. 70

8.3 Visualization of the correlation between the step count and the MDS-UPDRS Score ON. 72

List of Tables

2.1 Demographic profile and clinical score comparison (mean (SD)) between the two groups 9

4.1 mapping of the extracted multidimensional digital phenotype 22
4.2 Visualization of the comparison of the OxWearables step-count algorithm output and the ground truth values from the UK Biobank dataset. 24
4.3 Visualization of the comparison of the OxWearables step-count algorithm output and the ground truth in the controlled real-world experiment. 25

6.1 Visualization of the results of the statistical test of group comparison 43
6.2 Results of the RM-ANOVA analysis 48

7.1 Final performance comparison of the machine learning algorithms for the classification of PD and HC 62

List of Acronyms

AUC Area Under the Curve

CI Confidence Interval

CV Cross-Validation

FDR False Discovery Rate

FoG Freezing of Gait

H&Y Hoehn and Yahr Scale

HC Healthy Control

ICC Intraclass Correlation Coefficient

IMU Inertial Measurement Unit

k-NN k-Nearest Neighbors

MDS-UPDRS Movement Disorder Society - Unified Parkinson's Disease Rating Scale

MoCA Montreal Cognitive Assessment

ML Machine Learning

PD Parkinson's Disease

PPMI Parkinson's Progression Markers Initiative

RBF Radial Basis Function

RF Random Forest

RM-ANOVA Repeated Measures ANOVA

ROC Receiver Operating Characteristic

SVM Support Vector Machine

XGBOOST Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Parkinson's disease

Parkinson's Disease (PD) is a chronic, progressive neurodegenerative disorder that primarily affects the motor system but increasingly manifests as a multisystem pathology [1]. The disease is characterized by an extensive prodromal phase and a highly heterogeneous clinical presentation, representing one of the most complex challenges in modern neurology. Its biological hallmark is the selective loss of dopaminergic neurons and the accumulation of pathological alpha-synuclein aggregates, leading to a steady decline in both physical and cognitive domains.

Epidemiologically, PD is currently the fastest growing neurological condition globally. The global prevalence has risen from approximately 2.5 million patients in 1990 to over 10 million, establishing PD as the second most common neurodegenerative disease after Alzheimer's disease [2]. This Parkinson's pandemic is attributed to an aging global population, increased life expectancy and potentially higher exposure to industrial environmental factors such as pesticides or chemicals, known to be harmful to PD-related neurons and brain circuits.

The management of the disease is primarily symptomatic, aimed at restoring dopaminergic tone and improving the patient's quality of life. The pharmacological cornerstone is Levodopa (L-Dopa), a precursor to dopamine that crosses the blood-brain barrier that is highly effective in alleviating cardinal motor symptoms.

The clinical manifestations of PD are conventionally divided into motor and non-motor symptoms.

Motor symptoms Motor impairment is classically defined by the following key features:

- **Bradykinesia**: slowness of voluntary movement, reduced movement amplitude, and progressive decrement during repetitive actions;
- **Resting tremor**: typically occurring at a frequency of 4–6 Hz and most evident when the affected limb is at rest;

- **Muscle rigidity**: increased resistance to passive movement;
- **Postural instability**: usually emerging in later disease stages, contributing to balance impairment and increased fall risk.

In addition to these cardinal signs, patients often exhibit gait disturbances such as reduced stride length, shuffling gaits and episodes of freezing of gait (FoG), which severely affect mobility and independence [3]. These symptoms typically appear when 50-70% of dopaminergic neurons have already been lost. Their severity and progression are assessed using standardized clinical rating scales such as the Unified Parkinson's Disease Rating Scale (UPDRS) and its revised version (MDS-UPDRS).

Non-motor symptoms Non-motor manifestations, which are often more disabling than motor symptoms, include autonomic dysfunction, sleep disturbances, sensory deficits, mood disorders, and cognitive decline. Notably, non-motor symptoms may precede motor signs by several years and significantly impact patients' quality of life.

Diagnosis remains a clinical process, based on medical history and neurological examination, and guided by the Movement Disorder Society (MDS) criteria, which focus on identifying cardinal signs while excluding red flags indicative of atypical parkinsonism.

Neuroimaging techniques, such as dopamine transporter, single-photon emission computed tomography (DaT-SPECT), may be used as supportive tools to differentiate PD from other movement disorders, but they are not diagnostic on their own [4].

However, traditional clinical assessment and rating scales provide only a limited snapshot of the state of the disease. Several studies highlighted the importance of moving toward a more specialized approach. This includes the potential of objective monitoring and technology-based assessment to capture the patient's status in daily life, addressing the limitations of subjective clinical scores and improving the management of motor fluctuations and the overall quality of life [5].

1.2 Digital biomarkers for Parkinson's disease

In the evolving landscape of neurodegenerative disease research, digital biomarkers represent a transformative approach to quantifying disease states in an objective and continuous manner. They are defined as objective, quantifiable physiological and behavioural metrics collected and measured by digital devices such as portable, wearables or sensors. Digital biomarkers offer a shift from episodic to continuous monitoring, providing a high-resolution map of a patient's health status over prolonged periods [6]. This paradigm shift is particularly relevant in PD,

a condition characterized by progressive motor impairment, marked inter- and intra-individual variability, and strong dependence on contextual and medication-related factors.

1.2.1 Disease monitoring

The relevance of digital biomarkers in PD is particularly evident in disease monitoring and progression tracking. For example, traditional scales, although widely adopted, often lack the sensitivity to detect minor changes in motor performance during clinical trials, offering only a snapshot of the patient's condition. As a result, early or gradual deterioration may remain undetected for extended periods.

Continuous digital monitoring can identify sub-clinical tremors or minor gait instabilities that are undetectable through conventional assessments and that may precede clinically evident deterioration [7], enhancing measurement sensitivity. Disease progression in PD is generally slow, and even minimal annual changes in motor behaviour can be clinically meaningful. A very small decline in walking speed over several months is clinically significant but impossible for a human eye to detect during a routine check-up. While the clinical score may remain unchanged due to limited resolution, digital biomarkers can detect these consistent trends that act as an early warning system, signaling the need for physical therapy intervention.

1.2.2 Therapy optimization

Beyond disease monitoring, digital biomarkers play a crucial role in therapy optimization and personalized treatment management. One of the most studied applications is the management of the motor "ON–OFF" phenomenon: the ON state is the clinical state during which the medication is effectively controlling symptoms, the OFF state is the clinical state that occurs when the plasma levels of Levodopa fall below the therapeutic threshold leaving the patient with severe symptoms. This phenomenon is notoriously difficult to manage because it is highly variable and often unpredictable, but by identifying the exact moment when a patient enters an OFF state, through digital biomarkers, clinicians can move toward a precision medicine model, tailoring treatment to the individual's unique daily fluctuations, optimizing the timing and dosage of dopaminergic therapy to stabilize the patient's health state.

1.2.3 Ecological validity

One of the most significant advantages of digital biomarkers is their ability to operate in daily life conditions: monitoring patients while they perform activities of daily living provides a more ecological and realistic assessment of the disease than laboratory settings. Notably, gait parameters measured in the laboratory often correlate poorly with gait measured in the wild [8]. This discrepancy can be attributed to several factors:

• The white coat effect: patients tend to compensate for their motor deficits when they are aware of being observed by a clinician.

• Environmental complexity: laboratory settings are typically unobstructed and flat.

• Cognitive load: real-life mobility is a dual task activity, requiring the patient to walk while talking, carrying objects or processing environmental stimuli.

Consequently, digital biomarkers offer a more accurate representation of functional impairment and disease burden, reinforcing their potential role as objective measures of motor function and disease severity.

1.3 Open questions about digital biomarkers in Parkinson's disease

Despite the high potential of digital biomarkers, several open questions remain that hinder their implementation as reliable decision support systems in clinical settings [9].

One of the primary challenges in free-living monitoring is the **management of missing data**. While fragmented recordings are often caused by technical factors such as connectivity failures or battery drainage, the role of human compliance is equally determinant. In longitudinal studies, maintaining high adherence is challenging; participants may experience behavioral fatigue, leading to periods where patients may forget to re-wear the device or lose motivation to actively participate over extended periods. This decline in active participation can introduce significant gaps in the dataset. Ensuring data density is therefore essential, as data scarcity can lead to an underestimation of symptom severity and compromise the statistical validity of the comparison.

Another major challenge concerns the **overinformation problem inherent to continuous wearable monitoring**. The high frequency sampling of inertial sensors yields a massive volume of data, allowing for the extraction of a large number of potential metrics. Consequently, the primary difficulty lies in extracting the right parameters from an overwhelming sea of data. This data overload increases the risk of overfitting and limits comparability across studies, making rigorous feature selection protocols essential to distinguish true pathological signatures from mathematical noise.

A further open question relates to the **discriminative power of these metrics at the single patient level**: the ability of biomarkers to accurately profile an individual's pathological deviation from a normative baseline. Sensors can capture features that often yield statistically significant differences between groups, but a significant p-value does not necessarily imply a strong discriminative power. There is often a critical gap between identifying a group-level effect, quantified by the effect size, and achieving high classification accuracy at the individual level. A feature may demonstrate a large difference between the average PD patient and the average control, yet still suffer from substantial distributional overlap due to high inter-subject variability. Consequently, a biomarker might be scientifically valid for characterizing a population but

clinically insufficient for supporting a patient level decision. The translation of group-level differences into meaningful individual-level assessments remains a central unresolved issue in the validation of digital biomarkers.

1.4 Aim of the thesis

The primary aim of this thesis is to address the critical challenges of digital biomarkers by leveraging the comprehensive dataset provided by the Parkinson's Progression Markers Initiative (PPMI).

While PPMI is a landmark observational study collecting a vast array of clinical, imaging, and biological data to identify markers of Parkinson's progression, this thesis focuses specifically on the PPMI wearable sub-study. This initiative integrates the longitudinal clinical phenotyping of the main cohort with high-resolution, continuous sensor data collected via the Verily Study Watch. By combining traditional gold standard assessments with raw signals acquired in real world settings, this unique dataset allows us to bridge the gap between subjective clinical snapshots and objective, real-world motor quantification. [10]. The project was designed to overcome the limitations of traditional, episodic clinical evaluations by enabling continuous, real-world monitoring of motor behaviour through wrist-worn sensors and multimodal data collection.

This thesis does not merely aim to extract features, but to critically evaluate the translational potential of these digital tools through a rigorous, multi-stage analytical framework. The work seeks to determine whether wearable-derived metrics can transition from research variables to reliable, actionable instruments for precision medicine and individual patient profiling.

1.5 Outline of the thesis

Given the multidisciplinary nature of this work and the sequential methodology, the thesis intentionally diverges from the classical structure (Introduction, Methods, Results, Discussion) in favor of a progressive, pipeline-oriented organization. Following the introduction of the clinical problem and the technological background, the thesis systematically guides the reader through the consecutive stages of the computational workflow. Consequently, the core chapters of this work do not separate methodologies from their outcomes; instead, each chapter integrates both the specific methods and the corresponding experimental results. The overall structure of the analytical framework is summarized in Figure 1.1, which provides a visual overview of the pipeline implemented throughout this thesis.

The research begins with a rigorous preprocessing phase, focused on an initial analysis of data completeness to ensure the integrity of long-term monitoring (Chapter 3). This is followed by the core methodological step: the extraction of digital features via dedicated signal process-

ing algorithms applied to the raw inertial data (Chapter 4). Building upon this extracted dataset, the study proceeds to a statistical screening, designed to filter out redundancy and identify independent metrics (Chapter 5). This leads to a comprehensive discriminative analysis, where hypothesis testing and effect size quantification are employed to evaluate the ability of individual features to distinguish between pathological and normative motor patterns (Chapter 6). The validated biomarkers were subsequently used as input features for a supervised machine learning framework aimed at evaluating their combined discriminative capacity at the individual level (Chapter 7). Finally, the clinical relevance of the framework is assessed through a correlation analysis with standard clinical scores, investigating which specific features most accurately reflect the patient's concurrent severity and real-world functional impairment during that specific monitoring week (Chapter 8).

Figure 1.1: Overview of the analytical pipeline implemented in this thesis.

Chapter 2

Parkinson's Progression Markers Initiative

2.1 The PPMI wearable sub-study

The dataset employed in this thesis originates from the Parkinson's Progression Markers Initiative (PPMI) [11], a landmark observational clinical study sponsored by The Michael J. Fox Foundation for Parkinson's Research. Launched in 2010, PPMI represents a large-scale, comprehensive, and longitudinal investigation conducted across multiple international sites. Its primary objective is to identify robust biomarkers of PD progression to improve therapeutic trials and accelerate the development of disease-modifying treatments. [10].

Unlike traditional studies that rely heavily on episodic clinical observations, PPMI has evolved to integrate a multi-modal approach, combining advanced imaging (DaTscan, MRI), biologic sampling, and detailed clinical phenotyping. Recently, recognizing the limitations of subjective rating scales that capture only a brief snapshot of the patient's condition, the initiative expanded its protocol to include a specific wearable sub-study.

This sub-study was designed to capture the digital phenotype of the disease in a free-living environment. By monitoring participants continuously, the study aims to bridge the gap between the clinic performance, often influenced by the white-coat effect and the patient's actual functional status during activities of daily living [12].

2.1.1 Sensor specifications and capabilities

To achieve high-fidelity continuous monitoring, the PPMI wearable sub-study utilizes the Verily Study Watch, a medical-grade investigational device developed by Verily Life Sciences (a subsidiary of Alphabet). Unlike consumer smartwatches designed primarily for fitness tracking, the Verily Study Watch is engineered for clinical research, prioritizing the collection of raw, uncompressed sensor data over pre-processed metrics [13]. The device is equipped with a multimodal sensor suite, including:

• Inertial Measurement Unit (IMU): Comprising a triaxial accelerometer and gyroscope to capture high-frequency movement dynamics (sampled at high rates to detect subtle tremors and gait irregularities).

• Electrodermal Activity (EDA) Sensor: To measure skin conductance, potentially serving as a proxy for autonomic nervous system function and stress.

• Photoplethysmography (PPG) Sensor: To monitor heart rate and cardiovascular dynamics.

2.1.2 Research potential

While the clinical arm of the PPMI study is widely recognized as a cornerstone of Parkinson's research, the specific dataset generated by the Verily Study Watch remains a relatively novel and under-utilized resource in current literature. To date, only a limited number of studies have fully leveraged this high-fidelity raw data, leaving much of its potential unexplored. This relative scarcity of prior validation represents a significant scientific opportunity. The dataset offers a vast testing ground to rigorously address the persistent methodological challenges that hinder the clinical adoption of digital biomarkers. Therefore, this thesis leverages this resource not merely to replicate existing findings, but to provide new evidence regarding the reliability, stability, and diagnostic utility of wearable technology in the objective quantification of PD.

2.2 Study cohort and dataset

The dataset employed in this thesis was derived from the PPMI Verily data. From the extensive original database, a specific study cohort of 20 participants was selected:10 individuals with a confirmed clinical diagnosis of PD, representing the target population (PD) for the investigation and validation of digital motor biomarkers, and 10 individuals without a history of neurological or motor disorders, serving as a normative baseline (HC) for physiological motor behaviour, allowing the identification of deviations specifically attributable to Parkinsonian pathology.

The inclusion of both patients and controls allows for direct group-level comparisons while also supporting exploratory analyses of inter-individual variability. By constraining the cohort size and maintaining comparable demographic characteristics between groups, the study minimizes confounding effects related to age, education, and general health status, thereby enhancing the interpretability of wearable-derived features.

A summary of demographic characteristics and clinical assessments for both groups is reported in Table 2.1. Continuous variables are expressed as mean $\pm$ standard deviation (SD).

DEMOGRAPHICS	PD (n=10)	HC (n=10)
AGE (YEARS)	71.01 (6.97)	68.85 (9.75)
GENDER (M/F)	3/7	5/5
YEARS OF EDUCATION	16.70 (2.49)	16.30 (2.45)
CLINICAL SCORES		
MOCA	26.9 (2.4)	-
MDS-UPDRS 1	12.7 (4.6)	-
MDS-UPDRS 2	14.2 (5.6)	-
MDS-UPDRS 3 ON	27.8 (12.8)	-
MDS-UPDRS 4	4.11 (2.6)	-
MDS-UPDRS TOT ON	54.7 (17.6)	-
HOEHN & YAHR ON	2 (0.4)	-

Table 2.1: Demographic profile and clinical score comparison (mean (SD)) between the two groups

The groups were balanced in terms of age, gender, and educational level to minimize confounding factors that could influence motor performance. Group differences in age were assessed using a non-parametric Mann-Whitney U-test, confirming no statistically significant difference between the PD and HC groups ($Z = 0.35, p = 0.73$). This comparability ensures that any observed differences in motor features are primarily attributable to disease-specific mechanisms rather than age-related physiological decline.

The clinical status of the population was characterised using widely adopted and standardised clinical scores.

- **Montreal Cognitive Assessment (MoCA)**: a rapid screening instrument used to detect mild cognitive impairment [14]. Scores range from 0 to 30, with a score of 26 or higher typically considered normal. It assesses multiple cognitive domains, including attention, concentration, memory, and language.

- **Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS)**: the current gold standard for assessing PD severity and progression [15]. The scale is divided into four parts:

- *Part 1*: this section evaluates patient's cognitive and neuropsychiatric status. It combines semi-structured interviews with patient questionnaires.

- *Part 2*: Assesses the impact of the disease on functional independence through a questionnaire. It covers daily tasks such as speech, dressing, and hygiene.

- *Part 3*: this is the objective clinician-scored assessment of motor signs: the neurologist evaluates the cardinal signs of PD.

- *Part 4*: Assesses the adverse effects of long-term dopaminergic treatment, including the duration of motor fluctuations and the presence of dyskinesia.
- The MDS-UPDRS part 3 is sensitive to the patient's pharmacological state. Since medication logs were not available during the 7-day free-living monitoring, only the ON state clinical scores (representing the optimal dopaminergic response) were retained for the analysis.
- **Hoehn & Yahr (H&Y) scale**: a clinical staging system used to describe the overall progression of symptoms and functional disability [16]. The scale ranges from stage 1 to stage 5.

The clinical landscape of the analyzed cohort is characterized by a significant degree of inter-individual variability, which is essential for developing robust digital biomarkers. As reported in Table 2.1, the clinical scores exhibit relatively high standard deviations, confirming that the patients are not clustered around a single level of impairment but rather span a diverse spectrum of disease severity.

This clinical heterogeneity is not a methodological limitation but a significant strength of the dataset. By reflecting the real-world variance observed in clinical practice, it increases the ecological validity of the study. An algorithm capable of robustly tracking symptoms across such a diverse group, ranging from early-stage patients to those with more advanced residual deficits, demonstrates a higher potential for generalizability and clinical translation.

2.2.1 The Verily Study Watch

Motor activity was recorded using a medical-grade wearable wrist-worn sensor: the Verily Study Watch (Figure 2.2).

Figure 2.1: The Verily Study Watch, a medical-grade investigational wearable device used for data collection in the PPMI wearable sub-study.

The device integrates multiple inertial and physiological sensors that operate at high frequencies, enabling detailed characterization of movement patterns. In this study, the main sensors of interest were:

- **Triaxial accelerometer**: recording linear acceleration along three orthogonal axes to quantify precisely movement intensity and gravity. Accelerometry data provide critical information to detect tremor and gait irregularities, which are hallmark motor symptoms of PD
- **Triaxial gyroscope**: recording angular velocity along three axes to capture rotational dynamics. This information complements accelerometer data by providing a more complete description of wrist kinematics in order to recognize complex motor patterns specific of PD.

The device continuously recorded raw data, rather than features, at a high sampling frequency of 100 Hz, ensuring sufficient temporal resolution to detect fluctuations in motor activity throughout daily activities. A lower sampling frequency would risk undersampling phenomena such as tremor or bradykinesia, potentially leading to misestimation or even complete omission of clinically relevant features.

From the full recording period of each participant, we selected only a monitoring window of 7 days for the analysis. This one-week observation window was selected to capture the natural variability of daily living activities and to provide sufficient data density for the extraction of stable digital biomarkers. Participants were instructed to wear the device continuously throughout the day, removing it only for necessary maintenance.

To focus the analysis on active motor behaviour, a temporal filter was applied to exclude data recorded between 22:00 and 07:00, corresponding to the typical nocturnal rest period. Since the primary objective of this thesis is the characterization of voluntary motor activity, sleep data were considered non-informative and excluded to avoid underestimation of daytime activity levels.

Chapter 3

Data completeness

3.1 Extraction of data completeness

The efficacy and the reliability of health monitoring systems based on wearable sensors that collect free living measures depend strictly on the quality and quantity of the collected data: before extracting digital biomarkers from the data, a rigorous assessment of data completeness must be performed. Unlike controlled laboratory settings, unsupervised free-living environments can introduce significant noise, interruptions and discontinuities: real world factors such as battery issues or device removal may fragment the continuity of recordings.

In the context of PD, a further critical aspect concerns the possible influence of motor and cognitive impairments on participants' ability to properly use and maintain the wearable device over extended periods. While it is not possible to deterministically identify the specific cause of every interruption, it is fundamental to assess whether the patterns of data missingness differ significantly between the PD cohort and HC. If the presence of motor or cognitive impairments leads to a different adherence profile in patients compared to controls, this could introduce a systematic bias, distorting group-level comparisons of digital biomarkers.

The primary objective of this analysis was to quantify the effective wear time of the wrist sensor compared to the theoretical maximum, and to systematically compare data completeness profiles between the two groups, ensuring that any observed differences in motor features are not artifacts of unequal data density. To achieve this, data were organized into hourly bins, and for each hour the total number of recorded samples was computed. Data completeness was then quantified as the completeness ratio, defined as the ratio between the observed number of samples and the expected number of samples per hour. Given the fixed sampling frequency of 100 Hz, the expected sample count was set at 360,000 samples/hour. This formulation yields a normalized metric bounded between 0 and 1, facilitating comparisons between participants, days, and time intervals, regardless of the absolute duration of the recording. The use of hourly bins provides a temporal resolution sufficient to identify systematic patterns of data loss, such as recurring daily gaps or prolonged interruptions, while preserving interpretability.

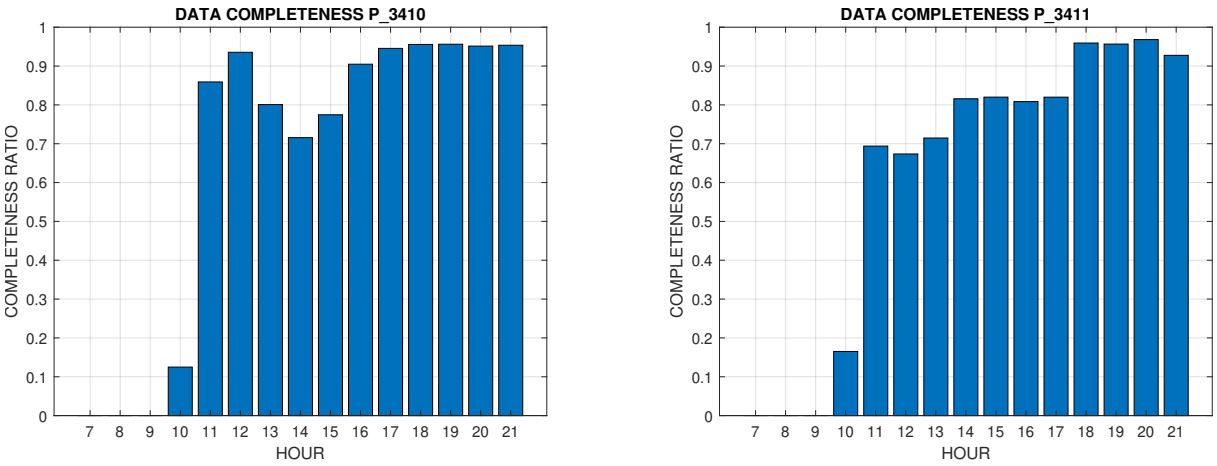


Figure 3.1: Representative evolution of the data completeness over time in two participants (two HC). Histograms show the mean hourly completeness ratio over seven consecutive days of recording.

This analytical framework enabled an initial visualization of the temporal evolution of data quality both within and across subjects. By examining completeness profiles over time, it was possible to identify periods characterized by reduced data density evaluating whether specific time windows consistently suffered from poor adherence and, crucially, to evaluate whether these patterns of data loss were uniform across the cohort or specific to the disease group. These observations informed subsequent methodological decisions, including the definition of valid analysis windows and the implementation of data cleaning procedures aimed at improving overall dataset consistency and robustness.

3.2 Analysis of hourly consistency

The first optimization step focused on defining a valid and reliable daily time window to be included in the analysis: the primary objective was to determine if certain time intervals were consistently underpopulated, which could compromise the stability and comparability of the extracted features.

To quantify data density, hourly histograms were generated for each patient by computing the mean completeness ratio for each hourly bin. This approach allowed the visualization of the typical daily recording profile of each participant, highlighting potential recurring gaps or periods of low adherence and identifying systematic patterns in data availability. A representative example of the temporal evolution of data completeness for two participants is shown in Figure 3.1.

Visual inspection of all the 20 individual histograms revealed a recurrent pattern across both study cohorts and statistical analysis of the hourly completeness profile confirmed a significant temporal disparity. A Mann-Whitney U-test performed on the aggregated hourly means revealed that the completeness ratio during the morning interval (07:00–11:00) was significantly lower compared to the remainder of the active day ($Z = -2.8, p = 0.005$). Quantitatively, while ad-

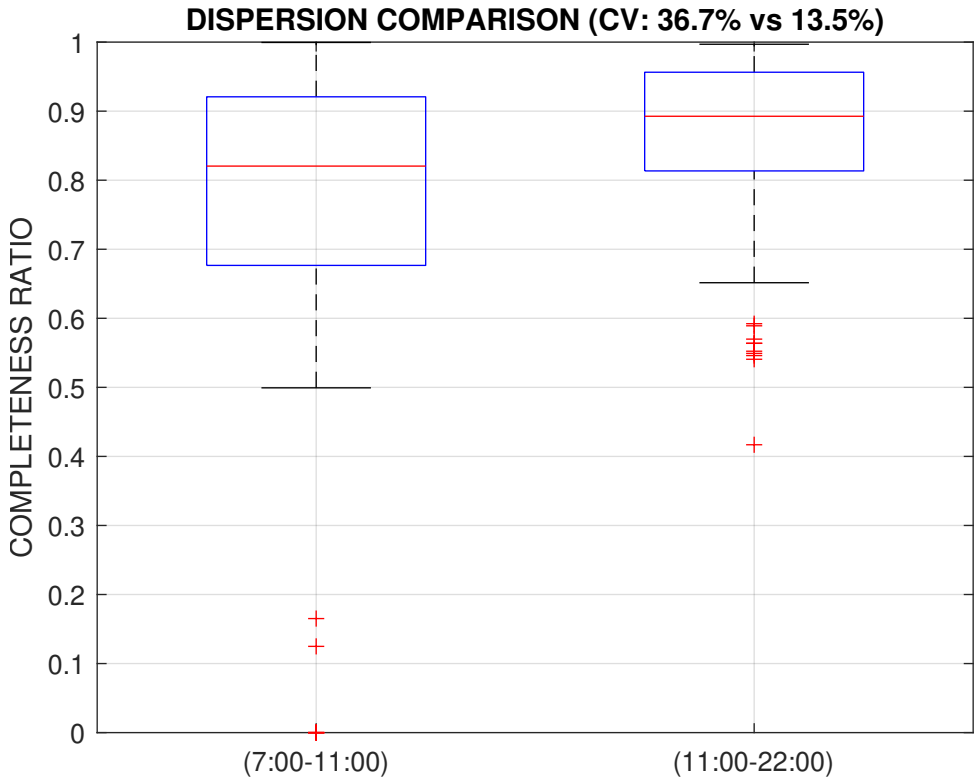


Figure 3.2: Comparison of data stability between the 07:00–11:00 and 11:00–22:00 time windows. The boxplots illustrate the distribution of the aggregated hourly completeness ratios. The morning interval (left) displays significantly greater dispersion (wider interquartile range) and a lower median compared to the afternoon interval (right).

herence stabilized above 90% in the afternoon, the mean completeness ratio during the morning hours remained consistently below 75%. In many participants, this specific interval was found to be nearly devoid of data, presenting a substantial gap in the daily recordings (Figure 3.1). These data gaps were systematically present in both the PD and the HC groups, indicating that the phenomenon was not disease-specific but primarily attributable to two main factors: the time required for the device to reach full charge, often performed overnight or during early morning, or simply the patient forgetting to put the sensor on immediately after waking up.

To further corroborate the decision to exclude the 07:00–11:00 interval, we performed a specific statistical analysis focusing on data stability and signal dispersion, rather than simple data volume. We compared the variability of the recording density between the excluded morning window (07:00–11:00) and the selected analysis window (11:00–22:00) using two metrics: the Coefficient of Variation (CV) and Levene's Test for equality of variances.

The Coefficient of Variation analysis revealed a contrast in data reliability between the two time segments. The morning interval exhibited a high degree of instability, with a CV of 36.66%. In contrast, the afternoon window demonstrated significantly higher consistency, with the CV dropping to 13.53%. The analysis highlights a critical lack of consistency in the data recording flow during the early hours. Processing such erratic data would likely bias the resultant digital biomarkers, leading to misleading conclusions about the patients' motor performance.

To formally assess whether this difference in variability was statistically significant, a Lev-

ene's test was conducted on the distribution of the completeness ratios. The test strongly rejected the hypothesis of homogeneity of variances between the two time windows ($F = 47.78, p = 2.89\times10-11$). Based on these observations, we decided to systematically exclude the 7:00-11:00 interval from the analysis. Focusing the study on the 11:00-22:00 window we enhanced data density and we improved inter-subject comparability: by removing the period with the highest variance and lowest compliance, the resulting dataset consisted of more homogeneous and data-rich segments across all participants. Although this decision may have excluded clinically relevant early-morning OFF-state manifestations, this trade-off was strictly necessary to ensure the algorithmic robustness of the pipeline. This preprocessing step was essential to ensure that downstream feature extraction and group comparisons were based on reliable and representative motor activity data, thereby enhancing the robustness of the overall analysis.

3.3 Analysis of daily consistency

To further evaluate the temporal stability of data completeness across the monitoring period, a daily consistency analysis was performed. We created a 2D heatmap for every patient to visualize the evolution of the completeness ratio over the monitoring week. This method allowed us to simultaneously visualize the completeness through all hours (y-axis) and days (x-axis), highlighting small interruptions, prolonged data gaps, and days characterised by lack of data.

Figure 3.3 illustrates the evident disparity between the best case and the worst case. In the best-case scenarios, completeness remained consistently close to the theoretical maximum (completeness atio=1), maintaining values systematically above 0.9 throughout the active monitoring period. These profiles exhibited only minimal interruption probably corresponding to the device maintenance: for example, participants must periodically remove the sensor for battery recharging and for uploading the data to the server. These interruptions were short in duration and not clustered around specific hours of the day, indicating that they did not compromise the overall validity of the daily recordings.

In contrast, in the worst cases we can observe significant data discontinuities, with extended periods of missing data: data scarcity can modify the results of the analysis because it can compromise the comparison between subjects. Comparing biomarkers with different recording volumes is statistically inappropriate, as symptom severity and motor variability may be underestimated simply due to insufficient sampling rather than genuine clinical differences. Consequently, days with unequal density of data cannot be treated as equivalent units in our analysis.

To maintain the validity of the study and to permit us to compare all the subjects in a rigorous way we implemented a data cleaning protocol. Specifically, any recording day that contained a cumulative data gap bigger than 3 hours was systematically excluded from the clean dataset. The 3-hour threshold was chosen as a trade-off between preserving sample size and ensuring com-

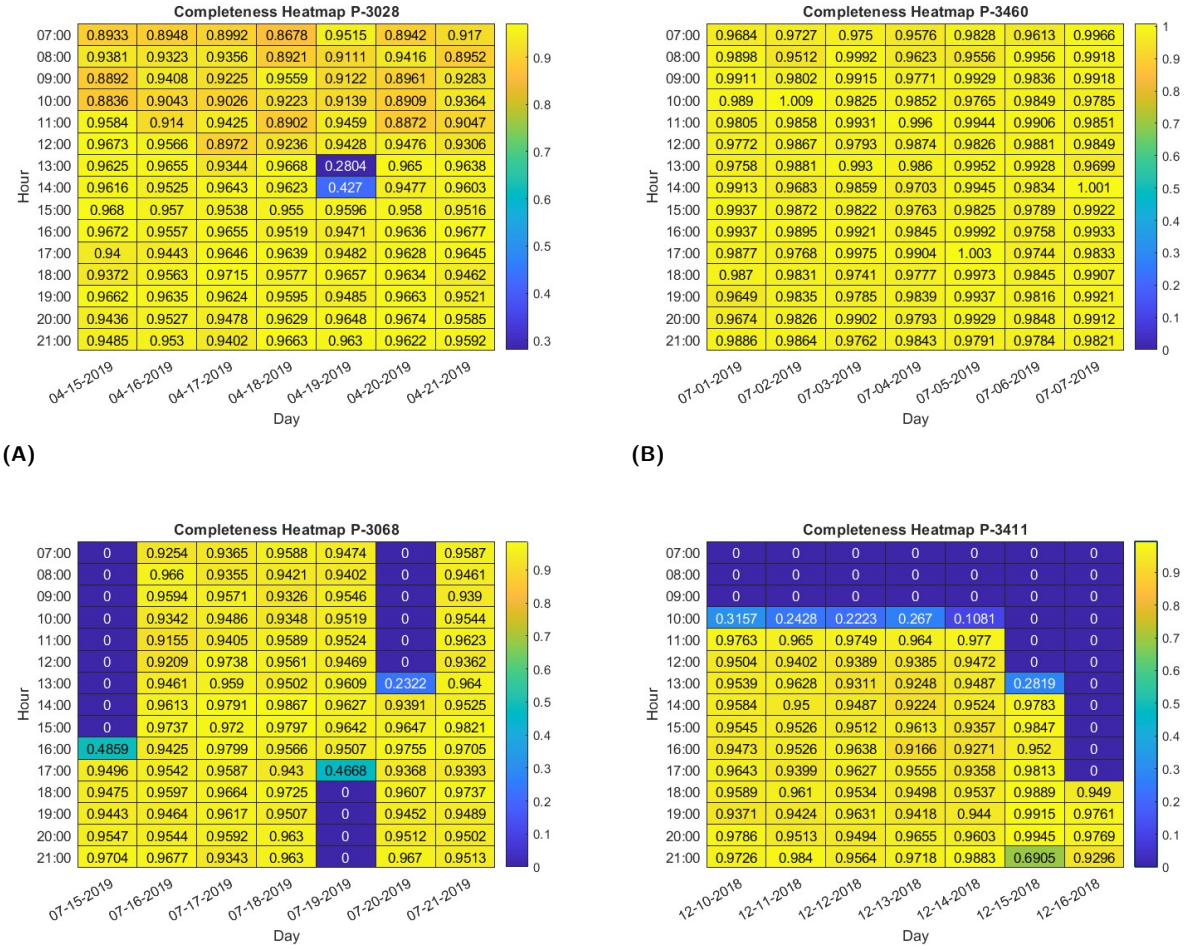


Figure 3.3: Comparative visualization of data completeness profiles. The panels display representative heatmaps of hourly adherence over the monitoring period: (A) Best case scenario for a PD patient; (B) Best case scenario for a HC; (C) Worst case scenario for a PD patient, exhibiting fragmented recording; and (D) Worst case scenario for an HC subject. The color gradient represents the completeness ratio (from 0 to 1), where brighter colors indicate high data availability and darker regions represent missing data.

parability across subjects. While this procedure decreased the total amount of available data, it ensured that the final dataset consisted exclusively of comparable and high-density recordings suitable for robust biomarkers extraction. As a result of this pruning process, eight daily recording were excluded from the dataset: specifically, 3 days were removed from HC group and 5 from the PD cohort. This trade-off between data quantity and data quality was deemed necessary to preserve statistical validity and to support meaningful group-level and individual-level analyses.

3.4 Completeness group comparison of adherence profiles

Following the comprehensive assessment of data completeness and the implementation of the data filtering and cleaning procedures, a group-level comparative analysis was performed to discuss about the potential differences in data quality between HC and PD participants. This comparison is a crucial step to ensure that any subsequent differences observed in wearable-derived digital biomarkers cannot be attributed to systematic differences in device adherence between the two cohorts.

3.4.1 Mean comparison

The primary objective of this analysis was to determine if the motor and non-motor impairments associated with PD negatively impact device adherence and overall data completeness.

Mean completeness ratios were computed for each participant, and group-level averages were visualized as mean $\pm$ standard error (Figure 3.4). Although the raw number of samples was slightly higher in the HC group, a non-parametric Mann-Whitney U Test was conducted to assess whether this difference was statistically significant. The test yielded a p-value that revealed no statistically significant differences between the two groups ($z=-0.7223$, $p=0.4701$), suggesting that overall data completeness and adherence in patients was statistically comparable to controls.

Crucially, this finding indicates that the motor impairments characterizing the pathological cohort did not significantly influence data completeness. Despite the physical challenges associated with PD, the patient achieved an adherence level statistically equivalent to that of the HC group. This result is particularly important, as it supports the feasibility of long-term, unsupervised wearable monitoring in PD populations. In addition, it strengthens the validity of the analysis by ensuring that any difference found in digital biomarkers can be attributed to the disease itself rather than artifacts introduced by unequal recording duration or compliance.

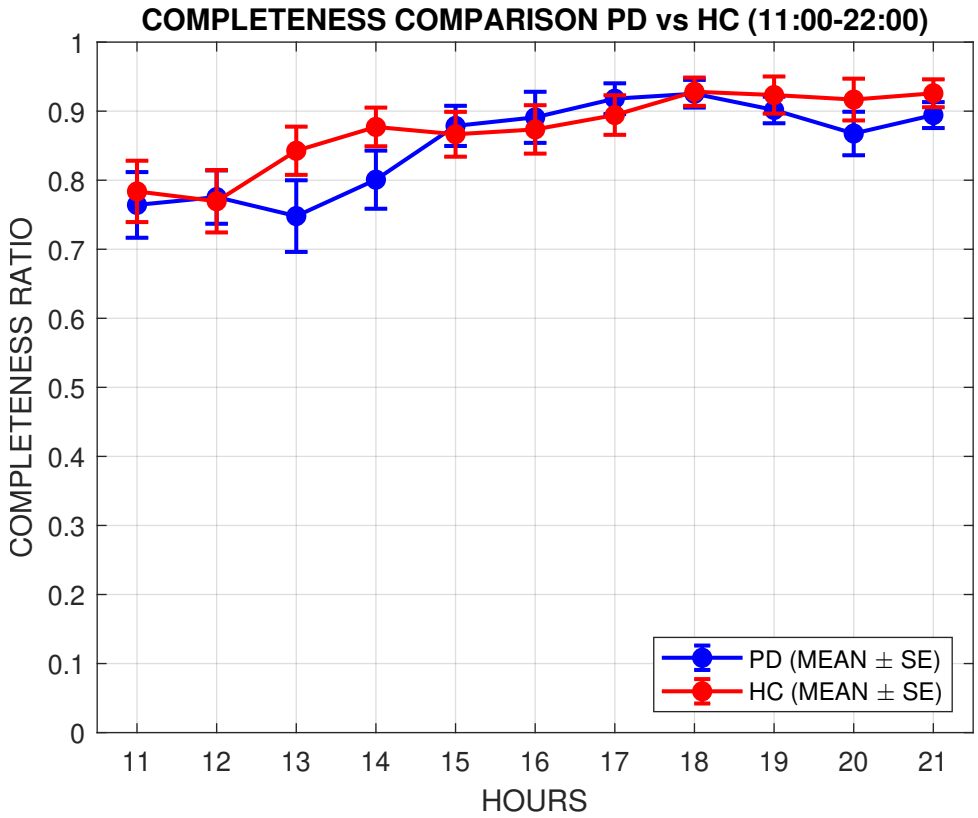


Figure 3.4: Comparison between the average completeness ratio (mean ± standard error) between PD and HC groups.

3.4.2 Variance comparison

Beyond mean differences, variability in data completeness represents an additional and equally important dimension of data quality. Increased variability may indicate inconsistent device usage, irregular wear patterns, or intermittent adherence, all of which could introduce bias and reduce the reliability of extracted digital biomarkers. Therefore, a second analysis was performed to compare the variance of completeness ratios between PD and HC groups.

Figure 3.5 illustrates the comparison between the variance of the two groups across our filtered time interval (11:00-22:00). This visualization provide an intuitive overview of the dispersion, central tendency, and presence of outliers within each group.

To quantitatively assess differences in variance, a Levene's test for equality of variances was applied. The test in the restricted 11:00-22:00 window revealed no statistically significant differences in variance between the PD and HC groups ($F = 0.0045, p = 0.9416$). Quantitatively, both cohorts exhibited low dispersion, with comparable standard deviations (PD: ±0.1214; HC: ±0.1111).

This result demonstrates that the variability of the data in the selected window is statistically indistinguishable across both groups. From a clinical perspective, this is a crucial finding: it indicates that the pathological condition did not introduce erratic behavior in device usage. Despite potential motor limitations or daily fluctuations typical of PD, the stability and precision

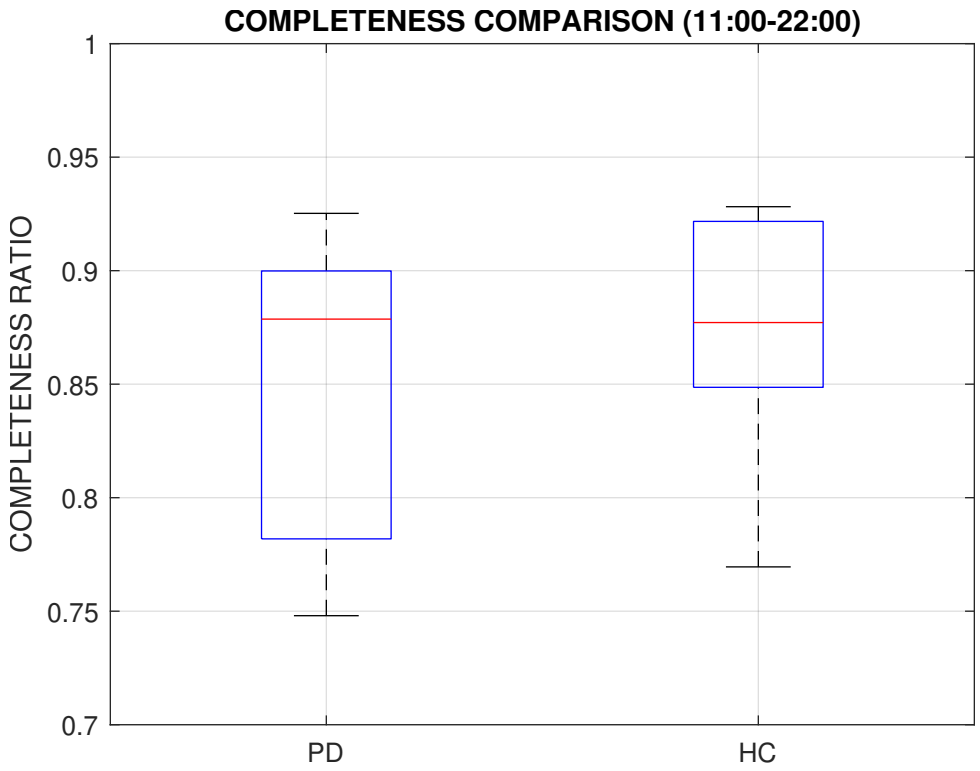


Figure 3.5: Boxplots illustrating the distribution of completeness ratios for PD and HC groups across the 11:00-22:00 interval.

of the data collected from patients were strictly comparable to those of healthy subjects.

3.5 Temporal effects on data completeness

While the previous analyses established the global quality of the dataset, it was also essential to understand the temporal dynamics of device usage within a typical day. In unsupervised home monitoring, adherence is not static, but may fluctuate depending on daily routines, sleep–wake cycles, and practical aspects of device handling. The specific objective of this section is to investigate about the effect of time on data completeness and to determine whether such temporal variations differed between PD and HC.

To achieve this, we performed a repeated-measures ANOVA (RM-ANOVA) using hours as the within-subject (repeated) factor and the group as the between-subject factor. The Mauchly Test of Sphericity was conducted to validate assumptions and, when violated, the Greenhouse-Geisser correction was applied to ensure a conservative and statistically valid inference.

The analysis revealed a statistically significant effect of time ($F = 8.36, p = 0.0001$) indicating that data completeness is not uniform through the day and, specifically, it varies significantly on specific hours of the day. In Figure 3.4 we can follow easily the trend line of data completeness: it progressively increased during the second half of the day for almost all subjects. This pattern is likely driven by behavioural and practical factors rather than technical issues. In particular, the risk of data loss is concentrated in the early morning hours, when par-

ticipants may delay wearing the device after waking up or may be charging the watch. However, once the device is worn, it tends to remain on the wrist for the remainder of the day, resulting in sustained high levels of completeness until bedtime.

The interaction effect was not statistically significant ($F = 0.9471, p = 0.4318$) confirming that the temporal pattern of daily usage is statistically similar between the groups: although data completeness varies as a function of time, this variation affects both groups in a comparable manner.

In summary, this rigorous data completeness analysis not only ensured the quality of the input for subsequent feature extraction but also provided valuable insights: the comparable adherence between PD and HC groups supports the feasibility of long-term wearable monitoring in this patient population, free from significant compliance-related biases.

Chapter 4

Feature extraction

While rich in information, raw accelerometer and gyroscope data are not directly interpretable from a clinical or physiological perspective. To gain a meaningful understanding of the disease and to enable a rigorous comparison between PD and HC, it is necessary to perform a structured feature extraction process.

Raw inertial data contains hundreds of data points per second that are not directly informative of specific clinical symptoms. Raw signals do not distinguish between voluntary movement and pathological tremor; features must be extracted to isolate and identify specific digital signatures that directly map clinical symptoms. In this context, feature extraction serves as a critical translation step between high-dimensional sensor data and interpretable clinical descriptors.

It is crucial to emphasize that simple statistical features, such as the global mean or variance computed over the entire raw signal, are inadequate for our clinical objectives. These aggregate metrics tend to obscure transient or frequency-specific phenomena central to PD, such as rhythmic tremor activity or progressive reductions in movement amplitude. Moreover, global statistics are highly sensitive to confounding factors such as overall activity level, lifestyle, or environmental context, limiting their specificity for disease-related motor impairments. Instead, more targeted features are required to capture distinct motor signatures and enhance discriminative power between PD and HC groups.

Following the pre-processing phase, we extracted a set of digital features from the raw accelerometric and gyroscopic signals. The feature extraction process was driven by clinical relevance and designed to encompass four core motor domains characteristic of PD: Gait, Tremor, Bradykinesia, and the Global Activity Profile (Table 4.1). Each domain reflects a different aspect of motor dysfunction and contributes complementary information to the overall characterization of disease severity.

The extraction process relied on validated, open-source algorithms available on GitHub and primarily implemented in Python. These algorithms operate in both time and frequency domains in order to isolate clinically meaningful patterns. By leveraging reproducible and widely adopted computational tools, the feature extraction process ensures methodological transparency, facili-

tates comparison with existing studies, and supports the development of robust and generalizable digital biomarkers.

Parameter	Symptom	Toolbox
STEP COUNT	Gait impairment	OxWearables
PERCENTAGE WINDOWS TREMOR	Tremor	ParaDigMa
MEDIAN RANGE OF MOTION ARM SWING	Hypokinesia	ParaDigma
MEAN HAND MOVEMENT AMPLITUDE	Bradykinesia	Mahadevan et al.
LIGHT PHYSICAL ACTIVITY	Global activity	ActiNet

Table 4.1: This table illustrates the explicit link between the engineered kinematic features, their corresponding target clinical domains (reflecting specific Parkinsonian motor symptoms or behavioral alterations), and the specific signal processing algorithms utilized for their derivation.

4.1 Gait metrics

Gait related metrics were extracted by the step count algorithm developed by the Oxford Wearables Group (OxWearables) [17]. This framework tries to robustly detect gait events using the information provided by a wrist-worn accelerometer. These features quantify both the volume and stability of the patient's ambulation, which are essential for assessing general mobility and identifying gait biomarkers for PD. Despite the indirect measurement location, wrist-based gait analysis provides a reliable proxy for overall mobility and walking behaviour when robust signal processing and machine learning approaches are employed

The model used in this algorithm is based on self-supervised machine learning, which allows the algorithms to learn discriminative representation of walking patterns without requiring extensive manually labelled data. The signal was first segmented into 10-second epochs, then a deep learning model was used to classify these windows into walking or non-walking periods. The effect of gravity was removed, and the signal was filtered using a low-pass fourth-order Butterworth filter at 5 Hz to isolate the frequencies characteristic of human gait. Steps were identified through peak detection algorithm applied only to the gait-classified segments.

From the output of the OxWearables algorithm, several primary gait-related features were extracted to characterize both the quantity and quality of ambulation:

- **STEP COUNT**: the total number of steps accumulated by the patient during the filtered monitoring hours. This is the primary metric for ambulatory volume, directly reflecting the patient's daily mobility level.

- **WALKING DURATION**: the cumulative time (in minutes) spent in active walking bouts. This metric specifically reflects the temporal duration of this physical activity.

- **STEP CADENCE**: the average stepping rate, measured in steps per minute. This feature can be altered, in PD, as a mechanism caused by a reduced stride length.

- **ENMO MG**: Euclidean norm minus one, measured in milli-g. It represents the measure for movement intensity. It provides a raw measure of physical exertion during gait, independently of the step count.

In addition to the primary outputs of the algorithm, three secondary biomarkers were derived from the outputs provided by the algorithm to capture more subtle aspects of gait quality and stability. These features were computed by post-processing the detected walking bouts and step-level information:

- **MEAN BOUT DURATION**: the average length of a single continuous walking bout. A reduction in this metric indicates gait fragmentation suggesting that patients are unable to sustain extended periods of walking and instead break their activity into short intervals.

- **MEAN INTRA BOUT VARIABILITY**: a measure of the irregularity of the steps within a single bout: high variability is a marker of gait instability and has been associated with fall risk and disease severity [18].

- **CADENCE CV INTER BOUT**: the coefficient of variation of step cadence calculated between the different walking bouts throughout the day. This metric quantifies how consistent the patient's walking cadence is across different time of the day. High inter bout variability suggests that the patient's walking ability fluctuates significantly.

These gait metrics provide a multidimensional characterization of ambulatory behaviour, spanning activity volume, temporal structure, and motor stability, and are well-suited to capture gait abnormalities associated with PD in real-world conditions. This comprehensive representation supports group comparisons between PD and HC and enables the investigation of gait-related digital biomarkers with potential relevance for disease monitoring and progression assessment.

4.1.1 Step count algorithm validation

Before applying the step-count algorithm to the clinical dataset, a validation procedure was performed to verify the accuracy of the algorithm in detecting steps from raw accelerometer. Although the OxWearables step-count algorithm has been previously validated in the literature, it is crucial to test its performance to ensure methodological robustness, particularly when the algorithm is used in a clinical context and applied to a dataset with different acquisition characteristics. The primary objective of this validation was to compare the algorithm's output against a ground truth in two different scenarios: a standardized epidemiological dataset and a controlled real-world experiment.

UK Biobank Dataset We first tested the algorithm on a sample subset from the UK Biobank accelerometer dataset: the gold standard in large scale accelerometery research due to its rigorous data collection protocols and extensive validation [19]. A selection of sample recordings was processed using the OxWearables step-count pipeline. The estimated step counts produced by the algorithm were then directly compared with the corresponding ground truth values provided in the dataset. Table 4.1 reports the comparison between algorithm output and reference step counts for representative subjects.

Patient ID	Algorithm Output	Ground Truth
p31	859	1044
p32	1957	1820
p33	4813	4804
p34	102	94
p35	1884	1884
p36	251	293
p37	189	248
p38	3178	3356
p39	725	605

Table 4.2: Visualization of the comparison of the OxWearables step-count algorithm output and the ground truth values from the UK Biobank dataset.

As detailed in table 4.2, the algorithm demonstrated a strong alignment with the ground truth across subjects. For subjects with significant ambulatory activities (e.g. p33, p35) the algorithm showed a very high precision, with near-identical estimated and reference values. In some subjects (e.g. p31) moderate discrepancies were observed, but the overall linear relationship confirms that the algorithm correctly identifies gait events across a broad range of activity levels. The script successfully processed the UK Biobank files, yielding step counts consistent with the real values reported in the literature, confirming that the algorithm performs as intended on standardized epidemiological data.

Real-World Controlled Test While the Biobank test confirmed software consistency and methodological correctness, it was necessary to validate the algorithm's accuracy on raw data collected in a setting that mimics our specific experimental conditions: to this end, a controlled real-world validation experiment was conducted. Raw triaxial accelerometer data were recorded using a smartphone (iPhone 12) equipped with the Physics Toolbox Sensor Suite application [20], which allows high-frequency logging of inertial sensor signals. The device was securely positioned on the subject's hand to mimic the placement of the wristwatch used in our clinical dataset. The subject of the study performed a standardized walking protocol consisting of five distinct trials. In each trial, the subject walked for exactly fifty steps, manually counted. This manual count was treated as the absolute Ground Truth. The raw data file extracted from the smartphone were then processed using the same step-count pipeline intended for the clinical

study.

Accelerometer	Algorithm Output	Ground Truth
accelerometer1	47	50
accelerometer2	44	50
accelerometer3	49	50
accelerometer4	48	50
accelerometer5	47	50

Table 4.3: Visualization of the comparison of the OxWearables step-count algorithm output and the ground truth in the controlled real-world experiment.

As reported in Table 4.3, the algorithm demonstrated high sensitivity and accuracy across all trials. The estimated step counts ranged from 44 to 49 steps against a target of 50, corresponding to a mean accuracy exceeding 90%. Minor underestimations were observed, likely attributable to subtle variations in wrist motion amplitude or brief pauses during turning, which are known sources of error in wrist-based step detection.

The successful completion of both validation stages confirmed that the feature extraction pipeline is robust and reliable for application to the clinical dataset of PD. The algorithm performed consistently across a large-scale epidemiological dataset and a controlled experimental setup that closely mimics the conditions of the present study. These results support the application of the OxWearables step-count framework to the clinical PD dataset, ensuring that subsequent analyses of gait-related digital biomarkers are grounded in accurate and validated step detection.

4.2 Tremor metrics

To objectively quantify the phenomenon of tremor in a naturalistic environment, we employed the spectral analysis algorithm provided by the open-source Parkinson's disease Digital Markers (ParaDigMa) toolbox [21]. Tremor represents one of the most distinctive motor symptoms of PD, yet its assessment in daily life remains challenging due to its intermittent nature and the confounding influence of voluntary movements. The objective was to detect the presence of tremor in unconstrained real-world conditions and to characterize both its intensity and its temporal constancy throughout the day.

The ParaDigMa toolbox is specifically designed to process wrist-worn inertial sensor data collected during everyday activities and to extract clinically meaningful digital biomarkers related to motor and non-motor manifestations of PD. Its architecture makes it particularly suitable for longitudinal monitoring outside controlled laboratory environments. The analysis focused on the raw gyroscope signal and the processing pipeline consisted in the following pipeline.

The raw gyroscope signal was segmented into non-overlapping 4-second windows. Using gyroscope data over accelerometer data is advantageous as it eliminates the need to filter out

the gravitational component. For each window, the Power Spectral Density (PSD) was calculated using Welch's method and summed across the three axes. The 12 Mel-frequency cepstral coefficients (MFCCs) were extracted and selected as the primary features because they are scale-invariant and capture the overall shape of the power spectrum. This scale invariance makes the algorithm robust to variations in sensor placement and signal amplitude, which is crucial for home monitoring. Based on video annotations, windows extracted from the dataset were labelled as tremor if this was present for at least 50% of the time. A logistic regression classifier was trained to detect tremor based on the MFCCs features. To ensure high specificity for Parkinsonian rest tremor and minimize false positives from daily activities, two post processing filters were applied to the positive windows: the dominant peak frequency had to fall strictly within the 3-7 Hz range, and a 'non-tremor arm movement detector' based on the power in the 0.5-3 Hz band was used to discard windows containing slow voluntary movements.

Based on the classified windows and the calculated power values the following clinical metrics were extracted:

- **MEDIAN TREMOR POWER**: represents the central tendency of the daily tremor intensity. It serves as a robust measure of the symptom severity because it filters out the outliers.
- **MODAL TREMOR POWER**: represents the most frequently occurring power value (the mode of the distribution).
- **X90P TREMOR POWER** (90th percentile): this metric captures the peak intensities of the symptom. It represents the power value below which 90% of the measurements fall.
- **PERCENTAGE WINDOWS TREMOR**: this is calculated as the ratio between the number of time windows classifies as tremor positive and the total number of valid windows in the analysis period. This feature is a direct indicator of tremor constancy: a high value indicates a persistent tremor.

These features provide a comprehensive characterization of tremor, capturing not only its magnitude but also its temporal expression. This approach enables reliable tremor quantification in real-world, supporting its use as a digital biomarker for PD.

4.3 Bradykinesia and hypokinesia metrics

Bradykinesia (slowness of movement) and hypokinesia (reduced in movement amplitude) are clinical hallmarks of PD and essential components of clinical diagnosis and disease staging. Unlike tremor, which is characterized by rhythmic oscillations, bradykinesia and hypokinesia manifest as subtle, distributed alterations in the speed, amplitude, and fluidity of voluntary movements. Their objective quantification in free-living conditions therefore requires analytical

approaches capable of capturing both task-specific motor impairments and global reductions in spontaneous activity. To capture these phenomena in a comprehensive manner, we implemented a dual-pipeline approach using two independent and validated toolboxes.

4.3.1 ParaDigMa framework

For the extraction of the features related to the arm swing we followed the pipeline provided by the ParaDigMa toolbox [22]. Arm swing abnormalities are clinically recognized as sensitive markers of Parkinsonian motor dysfunction, often appearing early in the disease.

A sliding Hann window of 6 seconds with a 5-second overlap was applied to the accelerometer signal. The algorithm extracted 34 features per window to capture characteristics of gait; a random forest classifier was then employed to distinguish gait from non-gait activities. Thanks to a logistic regression classification, gait segments are then processed to filter out periods where the hands were engaged in non-swinging tasks. This ensures that the analysis is restricted to gait-induced arm swing. On the filtered segments, raw gyroscope data are filtered and combined, and periodic peaks corresponding to the forward and backward swing of the arm are identified. The features related to the arm swing are then calculated.

- **MEDIAN PEAK VELOCITY ARM SWING**: calculated as the 50th percentile of all maximum velocity detected during walking bouts. Lower values indicated bradykinesia, reflecting the slowness of the movements.
- **MEDIAN RANGE OF MOTION ARM SWING**: it is calculated by integrating the velocity signal over the swing phase. A restricted ROM is a direct manifestation of hypokinesia.

It is important to underline that in specific monitoring days the filtering process may remove all detected gait segments. Consequently, if the duration of valid segments drops below a minimum threshold for a given day, the arm swing features are not extracted. In our dataset, this resulted in the exclusion of two specific days (July 8th and 13th) for patient 3023 and two days (May 8th and May 9th) for patient 3237. These exclusions ensure that the extracted features are based on sufficiently reliable and representative data.

4.3.2 Mahadevan et al. algorithm

To capture the generalized poverty of movement and the loss of motor control we implemented the pipeline described by Mahadevan et al. [23], which is specifically designed to quantify generalized bradykinesia and hypokinesia from wrist-worn accelerometer data.

The raw triaxial accelerometer data was converted into a single vector magnitude to ensure orientation-independent analysis. A band-pass filter (0.2-4.0 Hz) was applied to isolate voluntary human movements from gravity and high frequency noise. The filtered signal was

segmented into non-overlapping 1-minute epochs. An adaptive threshold was applied to the signal to classify the data into movement periods and rest periods. After the processing of the signal, three specific clinical metrics were extracted to map the bradykinetic profile.

- **MEAN HAND AMPLITUDE**: it is calculated as the mean of the acceleration magnitude during movement periods within the epochs. A lower mean hand amplitude quantifies the inability to generate large movements during daily tasks suggesting hypokinesia.
- **HAND MOVEMENT SMOOTHNESS**: it is estimated using the spectral arc length (SPARC) of the speed profile derived from the acceleration signal: a method that works in the frequency domain. A lower smoothness reflects the breakdown in motor programming and the loss of fluid movement execution showing us the motor fragmentation typical of the disease.
- **PERCENTAGE NO HAND MOVEMENT**: it quantifies the proportion of the monitoring window during which the signal amplitude remains below the movement threshold. This reflects akinesia and difficulty in movement initiation, capturing prolonged periods of inactivity that are commonly observed in PD patients.

This dual-pipeline strategy provides a comprehensive and multi-scale characterization of bradykinesia and hypokinesia in daily life conditions. The resulting digital biomarkers capture complementary dimensions of Parkinsonian motor impairment, enabling a better comparison between PD and HC, and supporting the clinical interpretability of wearable-derived features.

4.4 Global activity profile

The global activity profile characterizes how much and in what manner the patient occupies their daily time, offering an integrated view of motor function, lifestyle, and functional independence. In the context of PD, alterations in daily activity patterns are not limited to isolated motor symptoms but reflect the cumulative impact of bradykinesia, fatigue, motor fluctuations, and reduced mobility on everyday life. Quantifying these patterns in free-living conditions is therefore essential to capture disease burden beyond episodic clinical assessments.

To obtain a classification of daily behaviours and to quantify the lifestyle alteration caused by the motor impairment associated with PD, we employed ActiNet, a validated framework developed by the OxWearables group [24]. Unlike traditional methods that relies on simple intensity threshold, ActiNet utilizes advanced deep learning architectures to recognize complex behavioural patterns under real-world conditions.

The classification pipeline works on the raw triaxial accelerometer signal. It obtained the Euclidean norm minus one (ENMO): a measure that serves as the primary input for estimating energy expenditure. The ENMO signal was segmented into non-overlapping epochs which

served as the input window for the classifier. The ActiNet model utilized a Convolutional Neural Network (CNN) architecture trained on large-scale datasets. The model scanned the epoch to identify specific signatures associated with different human activities. For each time epoch, the model assigned a classification into one of four behavioural domains:

- **SEDENTARY HOURS**: time spent in behaviours characterized by minimal movement and very low energy expenditure, such as sitting or reading.
- **LIGHT HOURS**: time spent in low intensity physical activities such as slow walking and self-care tasks.
- **MODERATE VIGOROUS HOURS**: time spent in activities requiring significant physical effort such as brisk walking or sustained dynamic movements.
- **SLEEP HOURS**: estimated total sleep duration. The algorithm distinguished sleep from sedentary wakefulness.

By aggregating epoch-level classifications across the monitoring period, the global activity profile provides quantitative, clinically interpretable measures of how individuals distribute their time across behavioural states. Shifts toward increased sedentary time, reduced moderate-to-vigorous activity, and altered sleep patterns may serve as sensitive digital biomarkers of disease severity.

Chapter 5

Feature selection

The feature extraction phase, as detailed in the previous chapter, yielded a high-dimensional set of digital biomarkers describing various aspects of the Parkinsonian motor phenotype, including gait dynamics, tremor characteristics, bradykinesia-hypokinesia related features, and global activity patterns. While this richness of information represents a major strength of wearable-based monitoring, it also introduces substantial analytical challenges. In particular, high-dimensional feature spaces increase the risk of overfitting, reduce statistical power in small cohorts, and complicate the clinical interpretability of the results. Not all extracted features are equally suitable for longitudinal clinical monitoring. Some metrics may be highly sensitive to noise, daily behavioural variability, or contextual factors unrelated to disease pathology, while others may be redundant or strongly correlated with each other, providing overlapping information. Furthermore, certain features may exhibit systematic temporal trends driven by adherence patterns, circadian effects, or recording artefacts rather than genuine motor changes. Without an appropriate selection strategy, these factors can obscure true physiological signals and lead to misleading conclusions.

To address these limitations and to ensure a robust analysis, we performed a rigorous feature selection process based on three fundamental and complementary pillars.

Reliability: ensuring that the metric is stable over time (test-retest stability) and reflects a consistent motor characteristic rather than random fluctuations.

Redundancy reduction: ensuring that each feature provides independent information reducing multicollinearity and enhancing interpretability.

Temporal bias check: ensuring that the metric is not influenced by systematic trends or experimental confounds.

It is important to emphasize that the initial assessments of reliability (Section 5.1) and redundancy (Section 5.2) do not lead to immediate pruning. Instead, they provide the quantitative parameters required for the final decision-making stage. The actual selection is performed through a scatterplot (Section 5.3) that merges both reliability and independence criteria into a single visual framework.

The goal of this feature selection was to identify a subset of biomarkers that are reproducible across days, largely independent, and clinically interpretable. By constraining the analysis to features that satisfy these criteria, the resulting dataset becomes more suitable for robust statistical comparisons and machine-learning-based discrimination between PD and HC.

All statistical computations described in this chapter were performed using the MATLAB R2024b environment.

5.1 Reliability analysis

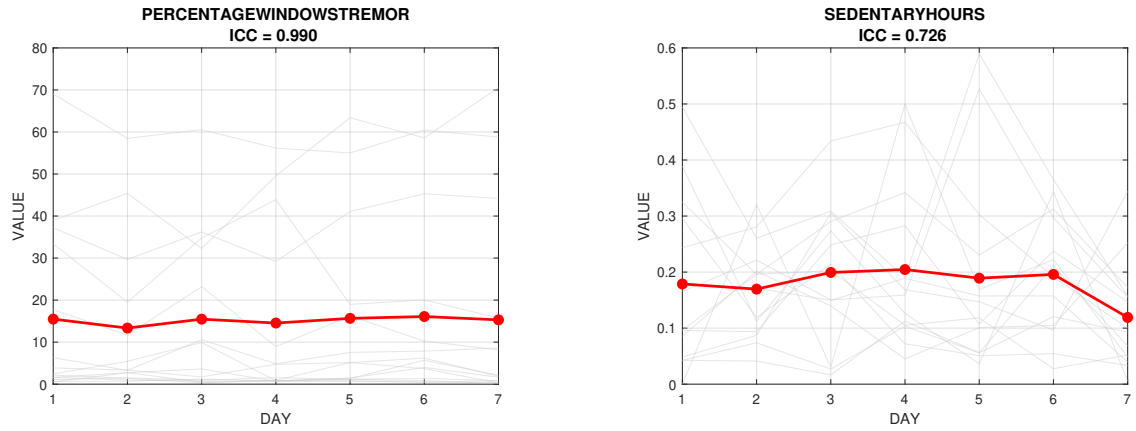
To assess the stability and robustness of our digital biomarkers, a reliability analysis was conducted using the Intraclass Correlation Coefficient (ICC) [25]. Reliability represents a fundamental prerequisite for any candidate digital biomarker, particularly in the context of continuous, unsupervised home monitoring, where measurements are repeatedly acquired under varying environmental and behavioural conditions. A feature that is not stable over time cannot be meaningfully interpreted as a disease-related marker, as its variability would be dominated by noise rather than underlying physiological characteristics.

In the context of continuous remote monitoring, test-retest reliability refers to the ability of a metric to provide consistent values for the same individual across repeated measurements collected at different time points. This aspect is particularly critical in PD, where motor performance naturally fluctuates throughout the day due to medication cycles, fatigue, and contextual factors. A robust digital biomarker must therefore be capable of distinguishing true inter-subject differences from intra-subject day-to-day variability. High test–retest reliability indicates that repeated measurements of the same subject yield statistically comparable values, implying that the feature captures a stable pathological trait rather than transient behavioural states or recording artefacts.

In our framework, the ICC was derived using a two-way ANOVA model without replication. The calculation is based on decomposing the total variance of the dataset into three primary components: the total sum of squares, representing the overall variation observed in the entire dataset, between-subject sum of squares, quantifying the variance attributable to real physiological differences between subjects, and the within-subject sum of squares, representing daily fluctuations and noise, capturing the residual variance.

After obtaining these values we computed the final ICC of the features as the ratio of the true variance to the total variance. The ICC values range from 0 to 1: an ICC close to 1 indicates that the measurement error is negligible compared to the differences between subjects, identifying the feature as a stable digital signature, an ICC close to 0 indicates that daily fluctuations mask the difference between subjects.

Features with ICC values above 0.90 were considered to exhibit excellent reliability, values between 0.75 and 0.90 good reliability, values between 0.50 and 0.75 moderate reliability, and



(A) (B)

Figure 5.1: Visualization of feature reliability through global mean values across the monitoring week. (A) Best-case scenario: PERCENTAGE WINDOWS TREMOR exhibits high test-retest reliability (ICC > 0.90), with stable daily values. (B) Worst-case scenario: SEDENTARY HOURS exhibits poor reliability (ICC < 0.75), with values dominated by daily noise rather than stable individual traits.

values below 0.50 poor reliability [26]. In the context of clinical monitoring and feature selection, a conservative threshold of $ICC > 0.75$ was applied in the following scatterplot analysis to retain only features demonstrating at least good test–retest stability.

Figure 5.1 illustrates representative examples of features with high and low ICC values by visualizing their global mean across the monitoring week for the best- and worst-performing cases. The analysis revealed heterogeneity in feature reliability. Features such as PERCENTAGE WINDOWS TREMOR demonstrated excellent test-retest reliability ($ICC > 0.90$). This suggests that despite hourly fluctuations, the overall daily burden of tremor is a highly consistent trait for a given patient, making it a strong candidate digital biomarker.

Conversely, other features were found to be dominated by daily noise rather than pathological traits. A notable example is the SEDENTARY HOURS metric, which exhibited an ICC value lower than 0.75: estimated time in a sedentary state was heavily influenced by noise and was considered unstable for robust clinical interpretation.

5.2 Redundancy and multicollinearity assessment

The second criterion applied for feature selection was information independence. In high-dimensional datasets, it is common to extract multiple features that measure the exact same physiological phenomenon. Including such redundant features leads to multicollinearity, which inflates the variance of statistical models without adding clinical insights [27]. This introduces several critical issues:

- **Overfitting:** it increases the complexity of the model without adding new information,

making the algorithm more likely to memorize noise rather than learning generalizable patterns, especially in cohorts with a limited sample size.

- **Unstable estimates:** it inflates the variance of the regression coefficients, making the model unstable and sensitive to minor changes in the input data.

- **Reduced interpretability:** it obscures the true relationship between clinical symptoms and digital metrics.

To avoid these pitfalls and ensure that each feature provides unique clinical information, we performed a correlation analysis across all extracted biomarkers to filter out redundant information while preserving the most clinically relevant metrics.

Before performing this analysis, it was essential to assess the distributional properties of the extracted biomarkers to determine the appropriate statistical method: we applied the Shapiro-Wilk test to each feature to verify the assumption of normality.

Most of the digital features did not follow a Gaussian distribution and returned a significant p-value ($p < 0.05$), suggesting the adoption of the non-parametric Spearman Rank Correlation for redundancy analysis. Unlike Pearson's method, Spearman's rho is robust to outliers and does not require the assumption of normality, making it the ideal choice for this dataset.

A correlation matrix (Figure 5.2) was generated to visualize the pairwise relationship between the features, using a threshold of $r > 0.85$ to identify highly redundant clusters. Pairs of features exceeding this threshold were considered "highly redundant," implying that they convey mathematically interchangeable information regarding the patient's state. The actual selection of the most representative feature from each redundant pair is performed in the integrated analysis described in the next section, where independence is weighed alongside temporal reliability to preserve the most robust and clinically informative metrics.

The analysis yielded two critical insights regarding the structure of our dataset.

Features belonging to different physiological domains generally showed low correlation coefficients ($r < 0.5$), confirming that our pipeline successfully captures distinct and independent Parkinsonian symptoms. For instance, a patient can exhibit severe tremor without necessarily having bradykinesia. This justifies the need for a multivariate approach: no single feature can capture the full complexity of the disease, and combining these independent domains is necessary for comprehensive phenotyping.

Conversely, some features from the same domain were highly redundant (see Figure 5.2).

5.3 Integrated reliability and redundancy analysis

To finalize the feature selection process, we integrated the results of the ICC and Correlation analyses into a single decision-making framework using a scatterplot. This approach enabled a

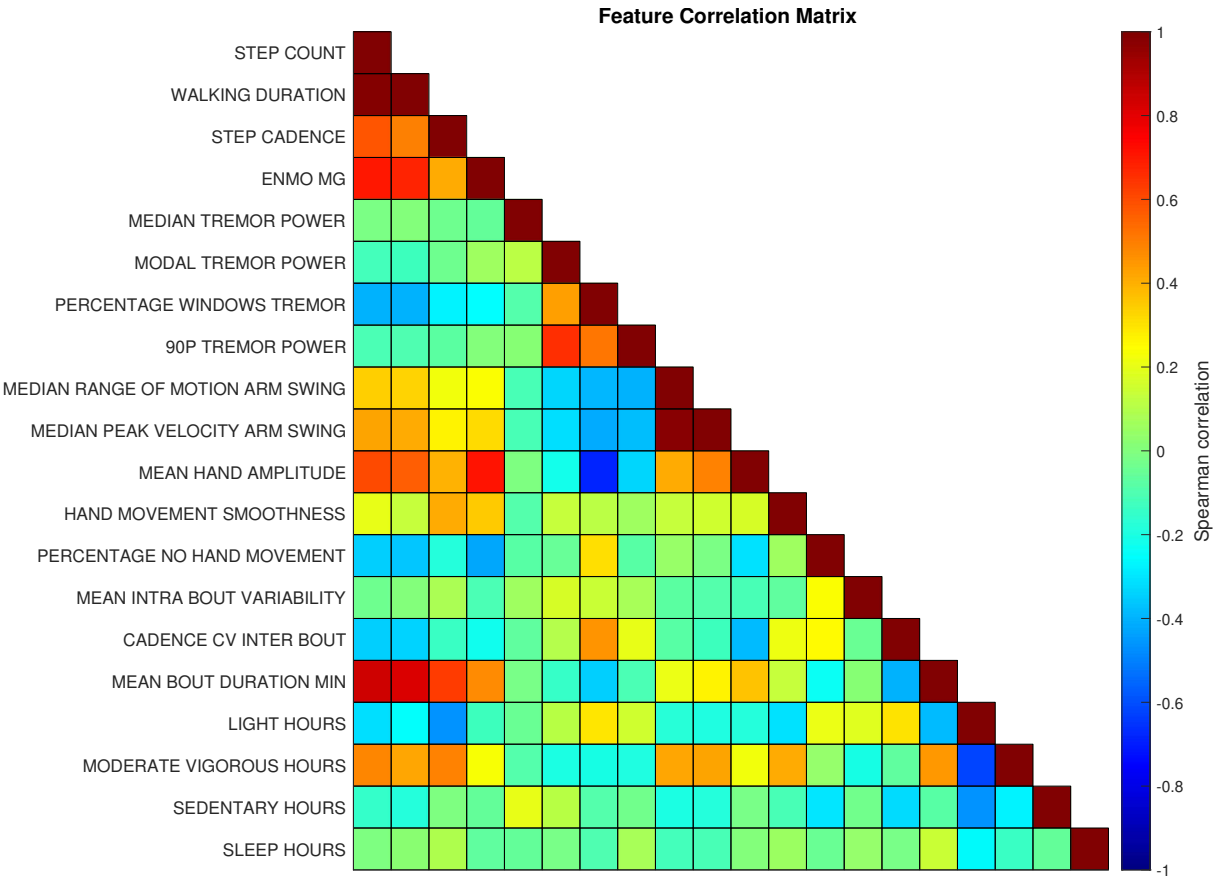


Figure 5.2: Heatmap of the Spearman correlation matrix of our features. Red areas indicate high positive correlation; blue areas indicate high negative correlation.

direct and intuitive visualization of the trade-off between temporal stability and informational redundancy as illustrated in the feature selection map (Figure 5.3).

This integrated map serves as the definitive filter for our dataset. Each extracted feature was represented as a single point in the scatterplot, with the Intraclass Correlation Coefficient (ICC) plotted on the Y-axis and the maximum absolute Spearman correlation with any other feature plotted on the X-axis. This representation allowed us to simultaneously assess whether a feature was both stable over time and sufficiently independent from the rest of the feature set.

To identify the subset of features suitable for downstream analysis, strict validation thresholds were defined based on methodological and clinical considerations:

- Features were required to exhibit good to excellent reliability, defined as $ICC > 0.75$, ensuring temporal stability across repeated daily measurements.
- Features were required to show limited redundancy, defined as a maximum absolute correlation < 0.85 with any other feature, ensuring that each retained metric contributed unique information.

These thresholds defined an optimal quadrant in the upper-left region of the scatterplot. Features falling within this region were considered both reproducible and non-redundant and were therefore retained for the final feature set.

Features with high stability but high redundancy were plotted in Figure 5.3 as orange dots. Although these features demonstrated strong temporal stability, their high correlation with other metrics indicated overlapping physiological information. The most representative feature of the correlated couple was kept, and the other was pruned to reduce dimensionality and minimize multicollinearity. A very strong correlation was observed between WALKING DURATION and STEP COUNT and to reduce dimensionality, we decided to discard WALKING DURATION retaining STEP COUNT as the primary feature for gait volume. Similarly, MEDIAN RANGE OF MOTION ARM SWING was selected over MEDIAN PEAK VELOCITY ARM SWING to minimize multicollinearity.

Features with low ICC were plotted in Figure 5.3 as red dots. All these features were discarded regardless of their correlation, as they lacked the fundamental stability required for clinical monitoring. This decision reflects the principle that temporal stability is a non-negotiable requirement for clinical monitoring: a feature that is not reproducible cannot serve as a reliable digital biomarker, even if it appears independent. The behavioural metrics SLEEP HOURS and SEDENTARY HOURS were removed from the final dataset because the analysis revealed an ICC below the threshold indicating that these features are subject to high daily variability driven by noise rather than reflecting stable disease-related traits.

The combination of ICC filtering and correlation-based pruning ensured that our final dataset is comprised exclusively of robust biomarkers. This selection process transformed our initial features into a refined clinical profile, maximizing the statistical power for the subsequent analysis.

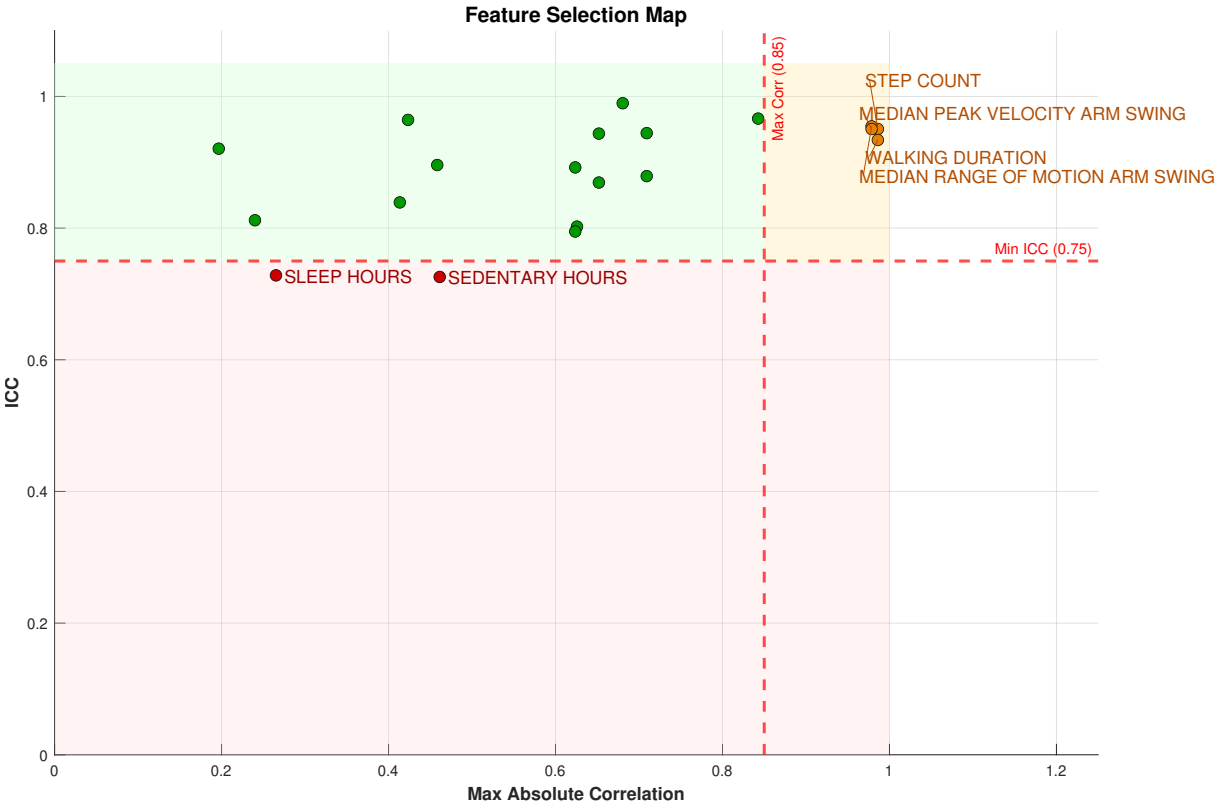


Figure 5.3: Scatterplot integrating reliability (ICC, y-axis) and redundancy (maximum absolute Spearman correlation, x-axis) analyses. Green dots indicate features satisfying both stability and independence criteria. Orange dots indicate features with high stability but high redundancy; only the most representative feature from each redundant pair was kept. Red dots represent features with low ICC, which were discarded regardless of correlation.

5.4 Temporal stability of extracted features

While the ICC assesses the temporal reliability of the measurement relative to random daily fluctuations, it does not detect potential systematic trends over time. Such trends, if present, could bias the interpretation of digital biomarkers and reduce their validity. To ensure that our digital biomarkers were not subject to significant temporal bias, we performed a Repeated Measure Analysis of Variance (RM-ANOVA) for each feature, with time as the repeated measure. This analysis evaluates whether the mean value of a feature significantly changes across the measurement days. Features exhibiting a significant main effect of time ($p < 0.05$) were flagged to inspection because this would indicate a systematic temporal bias, implying that the biomarker is not stable across the monitoring window.

The analysis revealed no statistically significant effect of time for our metrics (all p-values > 0.05). This confirms that the digital profile of our patients remained stable throughout the week, suggesting that our features are robust against systematic temporal bias.

The absence of a time effect provides several important insights:

- **Consistency of patient behaviour:** the stability of our features indicates that participants did not alter their activity patterns due to the burden of wearing the device. This suggests habituation to the wearable device and that the data captured represents a true snapshot of free-living behaviour.

- **Protocol validation:** from a methodological point of view, these results validate both the monitoring window and the experimental protocol. A one-week monitoring period is sufficient to capture representative behaviour without being confounded by adaptation effects, fatigue, or novelty.

- **Feature robustness:** features that remained stable across the week are more likely to reflect intrinsic disease-related traits rather than day-to-day variability or external confounders. This strengthens the validity of the selected digital biomarkers for clinical monitoring, disease characterization, and potential longitudinal studies.

5.5 Final set of features

As a result of this rigorous multi-stage selection process the initial feature space was reduced to a parsimonious set of key digital biomarkers. Starting from an initial set of 20 extracted metrics, 2 were excluded for poor reliability and 2 were removed to minimize redundancy yielding a final robust subset of 16 highly discriminative features structured as follows:

- **Gait domain:** STEP COUNT, STEP CADENCE, ENMO_MG, MEAN INTRA BOUT VARIABILITY, CADENCE CV INTER BOUT, MEAN BOUT DURATION MIN.

- **Tremor domain:** MEDIAN TREMOR POWER, MODAL TREMOR POWER, PERCENTAGE WINDOWS TREMOR, x90P TREMOR POWER.
- **Bradykinesia and Akinesia domain:** MEDIAN RANGE OF MOTION ARM SWING, MEAN HAND AMPLITUDE, HAND MOVEMENT SMOOTHNESS, PERCENTAGE NO HAND MOVEMENT.
- **Global activity domain:** LIGHT HOURS, MODERATE VIGOROUS HOURS.

The multi-stage feature selection process implemented in this study allowed us to systematically refine a large set of initially extracted digital biomarkers into a parsimonious, robust, and clinically interpretable dataset. By combining reliability analysis (ICC), redundancy reduction (correlation analysis), and temporal bias evaluation (RM-ANOVA), we ensured that each retained feature satisfies three fundamental criteria: temporal stability, independence, and resilience against systematic trends. The outcome of this feature selection pipeline is a well-curated and validated set of digital biomarkers that balances statistical rigor, clinical interpretability, and practical applicability. These features form the foundation for the subsequent analysis chapters, providing confidence that observed differences between PD and HC reflect true pathological signatures rather than methodological artifacts.

Chapter 6

Discriminative statistical analysis

The extraction of digital features from raw inertial data represents only the preliminary phase of this research. While advanced signal processing and feature engineering enable the quantification of complex motor behaviours in unsupervised real world settings, these engineering-derived metrics acquire clinical value only if they can be shown to meaningfully differentiate pathological from physiological motor patterns. Unlike controlled laboratory environments, where gait and movement are performed under standardized protocols, free-living data is inherently noisy and influenced by environmental contexts. Therefore, the central objective of this chapter is therefore to rigorously evaluate the discriminative capability of the selected digital biomarkers, assessing whether they can reliably distinguish between individuals with PD and HC.

However, comparing these two cohorts presents specific statistical challenges identified during the data analysis:

- **Sample size constraints:** the study cohort is relatively small ($N = 20$), which reduces the statistical power and increases the risk of type II errors (false negatives). In such settings, meaningful group differences may fail to reach conventional levels of statistical significance despite having clinical relevance.

- **Non-Gaussian distributions:** as verified by the Shapiro-Wilk test, most motor features exhibited skewed distributions, precluding the use of standard parametric tests.

- **High dimensionality:** the simultaneous analysis of multiple features increases the risk of type I errors (false positives) due to multiple comparisons. Without appropriate analytical safeguards, spurious differences may be incorrectly interpreted as disease-related effects.

To address these challenges, this thesis adopts a multi-layered statistical framework aimed at investigating the magnitude of differences, quantifying diagnostic classification capabilities, and studying the differences in the longitudinal temporal dynamics. This integrative approach ensures that the conclusions drawn in this chapter are not based on a single statistical criterion, but rather on the convergence of evidence across multiple analytical dimensions. By jointly

considering statistical significance, effect magnitude, classification performance, and temporal stability, the framework enhances methodological robustness while preserving clinical interpretability.

Ultimately, this chapter aims not only to identify statistically different features but to establish a coherent and clinically meaningful digital motor phenotype of PD that is resilient to sample size limitations, distributional violations, and temporal confounds.

6.1 Effect size estimation

In comparisons involving small sample sizes, relying solely on p-values can be misleading; a result may be statistically non-significant due to low power despite a potentially relevant clinical difference. For this reason, the present analysis framework explicitly incorporates effect size estimation to complement hypothesis testing and to distinguish random variability from differences that are potentially meaningful from a clinical and physiological perspective. Effect size quantifies the magnitude of the difference between groups independently of sample size and significance thresholds. In this study, Cohen's d was used to estimate the standardized difference between the PD and HC groups for each extracted feature. Cohen's d is defined as:

$$d = \frac{\bar{x}_{pd} - \bar{x}_{hc}}{s_{pooled}}$$

Where $\bar{x}$ represents the group mean and s_{pooled} is the pooled standard deviation of the two groups. This normalization allows direct comparison of effect magnitudes across features expressed in different units and across heterogeneous physiological domains. The magnitude of the effect was interpreted according to the standard threshold defined by Cohen:

- $|d| < 0.2$: negligible effect.
- $0.2 < |d| < 0.5$: small effect.
- $0.5 < |d| < 0.8$: medium effect.
- $|d| > 0.8$: large effect.

The results, summarized in Figure 6.1, reveal a heterogeneous landscape of discriminative power across the different physiological domains. The distribution of Cohen's d values confirms that PD does not manifest as a uniform alteration across all motor domains, but rather as a constellation of domain-specific impairments with different magnitudes of deviation from HC.

While several features exhibited a moderate effect size, a subset of metrics demonstrated a discrimination capability exceeding the 1.0 Cohen's d value, representing the strongest biomarkers in this framework:

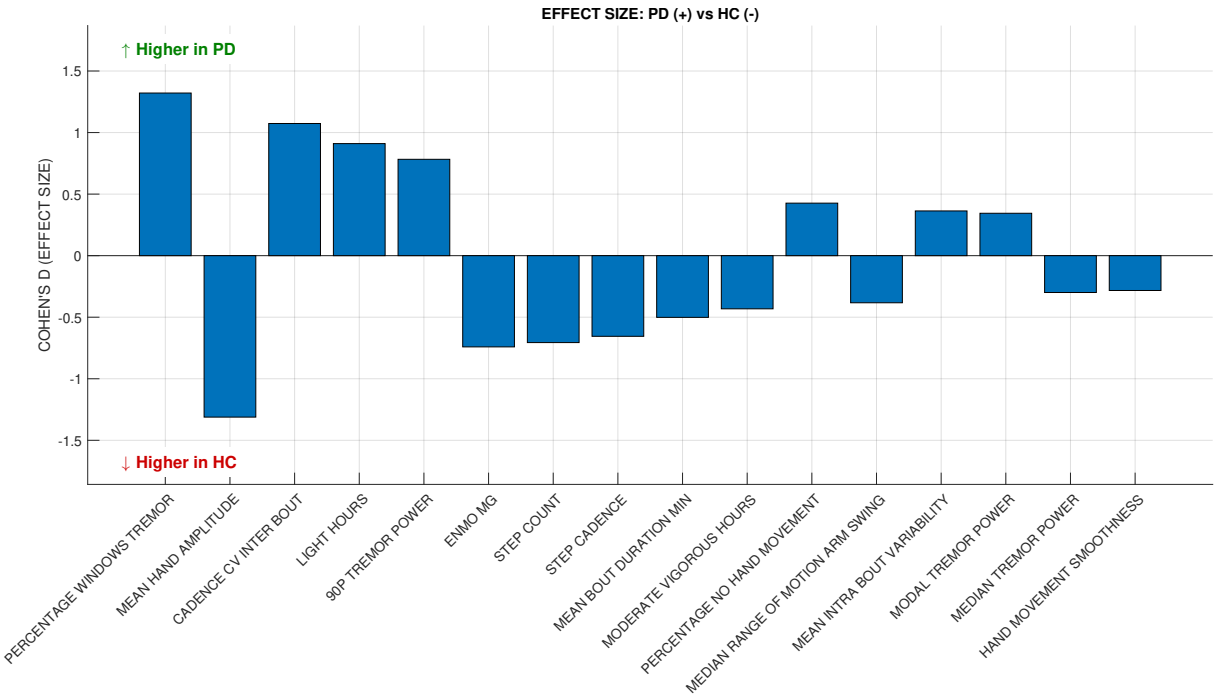


Figure 6.1: Effect size (Cohen's d) calculated between the PD and HC groups. Positive values indicate higher feature values in PD, negative values indicate lower values in PD.

- **Tremor constancy** (PERCENTAGE WINDOWS TREMOR, $d = 1.32, 95\%CI : 0.94 - 1.70$): this feature emerged as the most powerful discriminator. Tremor-related features displayed a clear dissociation between tremor intensity and tremor persistence. While MEDIAN TREMOR POWER ($d = -0.30, 95\%CI : -0.64 - 0.05$) and MODAL TREMOR POWER ($d = 0.34, 95\%CI : -0.00 - 0.69$) showed only small effect sizes, the PERCENTAGE WINDOWS TREMOR emerged as the strongest discriminator across all domains. This finding suggests that, in free-living conditions, the temporal burden and consistency of tremor is far more informative than instantaneous amplitude-based measures [28]. Clinically, this aligns with the observation that Parkinsonian tremor is often intermittent and context-dependent; therefore, metrics capturing how often tremor occurs may better reflect disease severity than metrics capturing how strong tremor is at isolated moments. The large effect size confirms tremor constancy as a robust digital signature of PD even in moderate disease stages. The high effect size observed for x90P TREMOR POWER ($d = 0.78, 95\%CI : 0.43 - 1.14$) further supports the relevance of distribution-based descriptors, which are less sensitive to outliers and capture the upper tail of tremor activity rather than its average behaviour.

- **Hypokinesia** (MEAN HAND AMPLITUDE, $d = -1.31, 95\%CI : -1.69 - -0.93$): the strong negative effect size of this feature reflects a marked reduction in the amplitude of spontaneous hand movement in PD patients. This aligns perfectly with the clinical definition of bradykinesia and reduced arm swing, validating the sensor's ability to capture the poverty of movement in daily life conditions.

- **Gait instability** (CADENCE CV INTER BOUT, $d = 1.07, 95\%CI : 0.71 - 1.44$): while simple gait metrics like cadence showed moderate separation, the variability of the rhythm between walking bouts proved to be a superior biomarker. This result suggests that gait instability and inconsistency, rather than absolute gait speed, better characterize pathological walking behaviour in free-living conditions. The elevated effect size implies that PD patients struggle to maintain a stable walking strategy throughout the day, reflecting impaired motor control, reduced adaptability, and fluctuating motor states. MEAN INTRA BOUT VARIABILITY ($d = 0.36, 95\%CI : 0.02 - 0.71$) showed smaller effects, reinforcing the idea that inter-bout variability metrics capture higher-level motor control deficits, while intra-bout metrics may be more influenced by contextual factors.

The observed large Cohen's d values for specific features demonstrate that the selected digital biomarkers capture substantial physiological differences between PD and HC, even when statistical power is limited. Moreover, the heterogeneity of effect sizes across domains highlights the importance of a multivariate approach. No single feature fully characterizes PD; rather, the combination of the motor symptoms provides a richer and more discriminative motor phenotype. The effect size analysis confirms that the refined feature set possesses not only statistical relevance but also clinically meaningful discriminative strength. The magnitude of the observed effects suggests that the differences identified in the following chapters can be confidently interpreted as genuine pathological signatures, rather than artefacts of noise, sampling variability, or methodological bias.

6.2 Discriminative power and diagnostic value

While effect size analysis provides a quantitative estimate of the magnitude of differences between PD and HC, it does not formally assess statistical significance, nor does it directly quantify the discriminative capability of individual features. To complement the effect size results, we performed a set of non-parametric statistical tests and classification-oriented analyses [29].

Specifically, this section integrates two complementary approaches:

1. A Mann-Whitney U-test to evaluate differences in central tendency between groups.

2. A Receiver Operating Characteristic (ROC) analysis was performed, using the Area Under the Curve (AUC) to quantify the ability of each feature to discriminate PD from HC.

This multi-pronged strategy allows us to move beyond a binary notion of statistical significance and to interpret digital biomarkers in terms of robustness.

Feature	EFFECT SIZE	AUC	RAW P-VALUE	ADJUSTED P-VALUE
STEP COUNT	0.706	0.830	0.014	0.044
STEP CADENCE	0.655	0.730	0.088	0.177
ENMO MG	0.741	0.750	0.064	0.146
MEDIAN TREMOR POWER	0.299	0.580	0.571	0.608
MODAL TREMOR POWER	0.344	0.630	0.344	0.459
PERCENTAGE WINDOWS TREMOR	1.321	0.860	0.007	0.029
X90p TREMOR POWER	0.782	0.820	0.017	0.046
MEDIAN RANGE OF MOTION ARM SWING	0.382	0.620	0.384	0.473
MEAN HAND AMPLITUDE	1.311	0.880	0.004	0.029
HAND MOVEMENT SMOOTHNESS	0.282	0.610	0.427	0.488
PERCENTAGE NO HAND MOVEMENT	0.426	0.680	0.185	0.330
MEAN INTRA BOUT VARIABILITY	0.363	0.640	0.307	0.447
CADENCE CV INTER BOUT	1.073	0.860	0.007	0.029
MEAN BOUT DURATION	0.501	0.670	0.212	0.339
LIGHT HOURS	0.910	0.920	0.001	0.027
MODERATE VIGOROUS HOURS	0.432	0.550	0.733	0.733

Table 6.1: Visualization of the results of the statistical test of group comparison. P-values in red are statistically significant. Blue values in the effect size column indicate a large effect (> 0.8). Green values in the AUC column are the five highest of the analysis.

6.2.1 Mann-Whitney U-test for feature comparison

Group differences were assessed using the Mann–Whitney U-test, a rank-based non-parametric test that evaluates whether two independent samples originate from the same distribution. Importantly, it tests for differences in distributional location (median) rather than means, which is often more representative of central tendency in skewed motor data.

However, conducting 16 simultaneous statistical tests on the same cohort inherently inflates the probability of committing a Type I error (false positives). To rigorously address this issue, the Benjamini-Hochberg procedure for False Discovery Rate (FDR) correction was applied alongside the raw p-values. A feature was considered strictly discriminative only if its FDR-adjusted p-value remained below $\alpha = 0.05$. Both raw and adjusted p-values are reported together in Table 6.1 to allow direct comparison and full transparency.

As summarized in table 6.1, several features reached statistical significance ($p < 0.05$). Notably:

- **STEP COUNT** ($z = -2.452, p = 0.014, adjustedp = 0.044$) showed a significant reduction in PD, reflecting a lower overall volume of ambulation in daily life.
- **PERCENTAGE WINDOWS TREMOR** ($z = 2.683, p = 0.007, adjusted\ p = 0.029$) and **x90P TREMOR POWER** ($z = 2.381, p = 0.017, adjusted\ p = 0.046$) showed significant differences, confirming that tremor-related features are among the most robust markers differentiating PD from HC.
- **MEAN HAND AMPLITUDE** ($z = -2.384, p = 0.004, adjusted\ p = 0.029$) demonstrated a highly significant reduction in PD patients, reflecting hypokinesia and reduced spontaneous movement.

- **CADENCE CV INTER BOUT** ($z = 2.683, p = 0.007, adjusted\ p = 0.029$) showed significant group separation, indicating that variability in walking rhythm across the day is a hallmark of pathological gait control.
- **LIGHT HOURS** ($z = 3.13, p = 0.001, adjusted\ p = 0.027$) also exhibited significant differences, highlighting alterations in the global activity profile of PD.

The survival of all originally significant features after FDR correction is a noteworthy finding, suggesting that the identified group differences are unlikely to be statistical artifacts. In summary, the Mann–Whitney U-test confirms that a subset of wearable-derived digital biomarkers exhibits statistically significant differences between PD patients and HC, even under conservative non-parametric assumptions. Tremor constancy, gait rhythmic variability, hypokinesia-related metrics, and global activity features emerge as the most robust discriminators providing converging evidence that the wearable-derived metrics capture a multifaceted and clinically coherent digital motor signature of PD. These results form a solid and rigorously validated foundation for the subsequent classification and correlation analyses, with the FDR correction providing additional confidence that the observed discriminative patterns reflect genuine pathological differences rather than chance findings inflated by multiple testing..

6.2.2 Receiver Operating Characteristic (ROC) analysis

To evaluate the diagnostic classification capability of each feature at individual level, Receiver Operating Characteristic (ROC) curves were computed and the Area Under the Curve (AUC) was used as a summary metric. ROC analysis assesses how well a feature can discriminate PD from HC across all possible decision thresholds, independently of statistical significance testing. ROC analysis evaluates the trade-off between sensitivity and specificity across all possible decision thresholds and is independent of distributional assumptions. An AUC of 0.5 indicates chance-level performance, whereas values approaching 1.0 indicate excellent discriminative capability. Several features demonstrated excellent discriminative performance:

- **LIGHT HOURS** (AUC=0.92) emerged as the strongest individual classifier, suggesting that alterations in daily activity structure provide a strong signature of Parkinsonian pathology.
- **MEAN HAND AMPLITUDE**(AUC=0.88) showed exceptional diagnostic power. This confirms that the reduction in spontaneous upper-limb movement (hypokinesia/bradykinesia) during free-living conditions is a core and highly discriminative hallmark of the disease [30].
- **PERCENTAGE WINDOWS TREMOR** (AUC=0.86) and **CADENCE CV INTER BOUT** (AUC=0.86) showed high sensitivity–specificity trade-offs, identifying them as highly informative biomarkers of tremor burden and gait instability.

Conversely, other gait-related metrics such as STEP CADENCE (AUC = 0.73) and MEAN BOUT DURATION MIN (AUC = 0.67) showed more moderate ROC performance. This suggests that while these parameters differ significantly on average between groups, there is a greater degree of individual overlap in these specific dimensions.

In summary, the ROC analysis confirms that the extracted digital features possess significant diagnostic information. The high AUC values obtained for multiple independent domains validate the use of wearable-derived metrics as objective tools for discriminating PD patients from healthy individuals in real-world settings.

6.2.3 Integrated interpretation of statistical findings

When jointly considering the results of the Mann–Whitney U-tests, effect size estimates, and ROC analyses, a coherent and clinically meaningful picture of Parkinsonian motor impairment emerges. Rather than relying on a single statistical indicator, this integrated approach allows us to evaluate each feature across three complementary dimensions: statistical significance at the group level, magnitude of the pathological effect, and individual-level discriminative capability. This multidimensional validation is particularly important in PD, where motor symptoms are heterogeneous and fluctuate over time.

Univariate analysis revealed that features such as LIGHT HOURS, MEAN HAND AMPLITUDE, PERCENTAGE WINDOWS TREMOR, and CADENCE CV INTER BOUT (Figure 6.2) emerged as the most robust individual discriminators between the PD and HC. They exhibited both a substantial effect size—indicating a clinically meaningful magnitude of difference between the groups—and a high Area Under the Curve (AUC) in the Receiver Operating Characteristic (ROC) analysis.

From a physiological perspective, these results are highly consistent with the known pathophysiology of PD. The high discriminative power of PERCENTAGE WINDOWS TREMOR confirms that the constancy and temporal burden of tremor are more informative than average tremor amplitude alone [30]. This suggests that Parkinsonian tremor is best characterized as a persistent motor state rather than a sporadic high-power event. Similarly, CADENCE CV INTER BOUT reflects impaired gait automatization and reduced motor control stability, hallmarks of PD motor symptoms. The strong performance of LIGHT HOURS highlights how PD extends beyond isolated motor symptoms, influencing global activity patterns and daily energy expenditure. Crucially, MEAN HAND AMPLITUDE, also emerged as a highly robust discriminator, underscoring its direct physiological relevance in capturing bradykinesia and generalized hypokinesia. Importantly, the integrated analysis underscores that no single digital feature is sufficient to fully characterize PD. The disorder manifests as coordinated alterations across multiple motor domains, including gait, tremor, upper-limb kinematics, and global activity behaviour. Features that individually show moderate discriminative power may collectively provide a much richer and more robust representation of the disease phenotype. This observa-

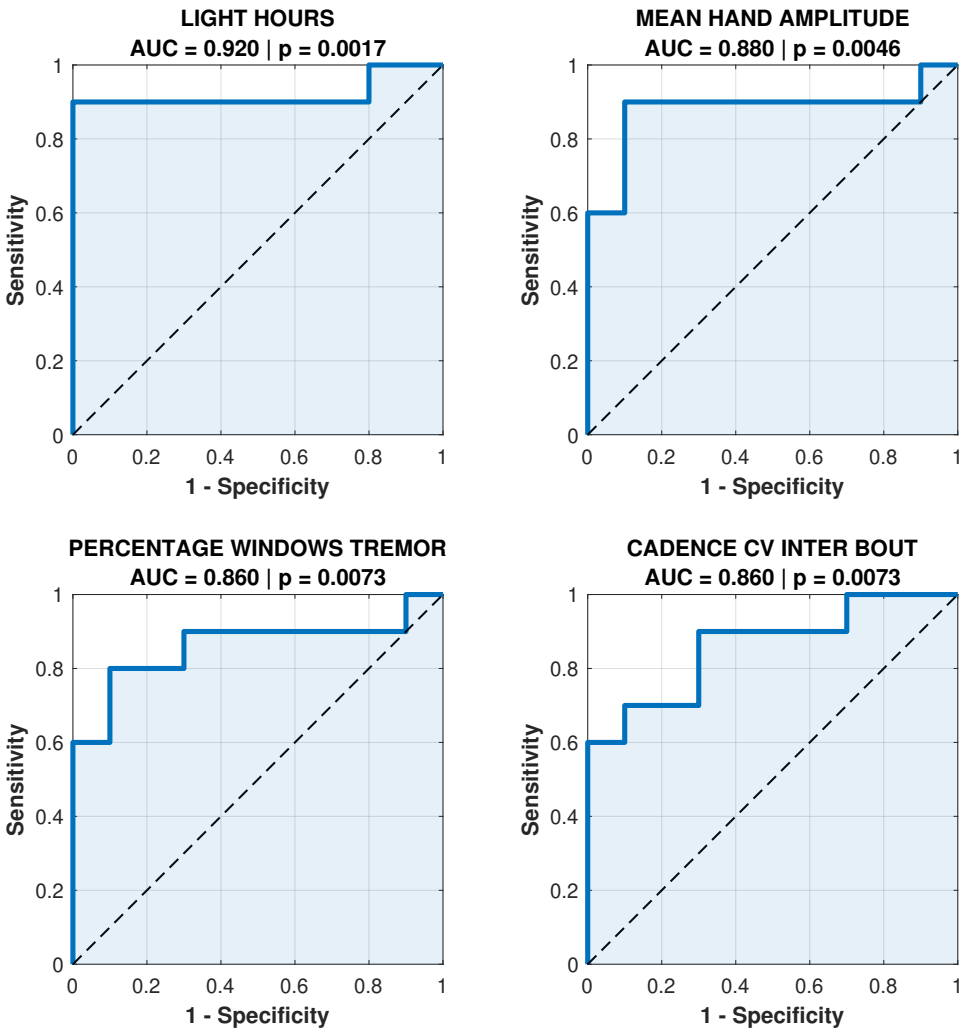


Figure 6.2: Visualization of the four highest AUC.

tion strongly supports the adoption of a multivariate modelling strategy in subsequent analyses. From a methodological standpoint, the combination of non-parametric hypothesis testing and ROC analysis ensures that selected biomarkers are not only statistically valid but also clinically relevant. While p-values quantify the existence of group-level differences, effect sizes provide insight into their clinical magnitude, and ROC curves evaluate their potential utility for individual-level classification. The convergence of these metrics strengthens confidence that the observed differences reflect genuine disease-related mechanisms rather than noise or methodological artefacts. In summary, this integrated interpretation confirms that the selected feature set captures meaningful and complementary aspects of Parkinsonian motor dysfunction. The results justify the transition from univariate statistical comparisons to multivariate classification and predictive modelling, where these biomarkers can be combined to enhance diagnostic accuracy and robustness. The validated features identified in this chapter form a solid foundation for the subsequent machine learning analyses aimed at automated discrimination between PD and HC.

6.3 Intra-week temporal stability

An additional and clinically relevant question concerns whether PD and HC exhibit different longitudinal trajectories in our monitoring week. In other words, beyond static group differences, it is crucial to determine whether the daily evolution of motor behaviour over time differs between groups, which would suggest adaptation to the device or distinct day-to-day fluctuations. To address this question, a repeated measures ANOVA (RM-ANOVA) was conducted for each retained feature, including time as the within-subject factor and group (PD vs HC) as the between-subject factor. Of particular interest was the time × group interaction term, which tests whether the temporal evolution of a given feature differs between the two cohorts. A significant interaction would indicate that PD and HC subjects change differently over time, potentially reflecting differential fatigue accumulation or different daily fluctuations.

Feature	TIME EFFECT (p)	INTERACTION (p)
STEP COUNT	0.75	0.39
STEP CADENCE	0.84	0.76
ENMO MG	0.33	0.18
MEDIAN TREMOR POWER	0.35	0.35
MODAL TREMOR POWER	0.55	0.54
PERCENTAGE WINDOWS TREMOR	0.59	0.56
X90p TREMOR POWER	0.75	0.27
MEDIAN RANGE OF MOTION ARM SWING	0.37	0.74
MEAN HAND AMPLITUDE	0.43	0.24
HAND MOVEMENT SMOOTHNESS	0.60	0.68
PERCENTAGE NO HAND MOVEMENT	0.30	0.48
MEAN INTRA BOUT VARIABILITY	0.50	0.62
CADENCE CV INTER BOUT	0.44	0.32
MEAN BOUT DURATION	0.66	0.25
LIGHT HOURS	0.08	0.75
MODERATE VIGOROUS HOURS	0.23	0.57

Table 6.2: Visualization of the results of the RM-ANOVA. The time effect's p-value reflects the stability of the feature, the interaction's p-value shows if there are differences in the temporal evolution of the features.

Table 6.2 summarizes the p-values associated with the main effect of time and the time $\times$ group interaction for each feature. Across all analyzed features, no statistically significant main effect of time and, critically, no significant time $\times$ group interaction ($p > 0.05$) was observed. This result was consistent across all physiological domains, including gait, tremor, upper-limb movement, and global activity metrics. The absence of significant interaction effects has several important implications. First, it indicates that PD and HC subjects exhibit parallel temporal trajectories over the monitoring period. Although the two groups differ in absolute levels for several features, the shape of their temporal evolution remains comparable. In practical terms, this means that PD patients do not show systematic intra-week deterioration, or compensatory behavioural drift relative to controls within the observed time window. Second, this finding supports the interpretation that the group differences identified in earlier chapters are stable traits rather than transient states. The discriminative features, such as tremor burden, gait variability, and reduced activity levels, reflect enduring characteristics of the Parkinsonian motor phenotype, rather than effects driven by short-term adaptation, learning, or fatigue.

From a domain-specific perspective:

- **Gait-related features** remained temporally stable in both groups, suggesting that the observed gait impairments in PD represent persistent control deficits rather than progressive daily changes.

- **Tremor-related metrics** showed no differential temporal modulation, reinforcing the interpretation of tremor burden as a stable motor hallmark in free-living conditions.

- **Movement features** did not exhibit divergent temporal patterns, , indicating consistent hypokinetic behaviour across the week.
- **Global activity metrics** remained parallel between groups, further confirming that reduced activity levels in PD are structural rather than transient.

In summary, the repeated measures analysis revealed that, although PD and HC differ significantly in the magnitude of several digital biomarkers, their short-term temporal stability over the monitoring period is comparable. The absence of time × group interactions confirms that the observed group differences are robust, stable, and not driven by temporal confounds.

6.4 Multivariate group comparison (MANOVA)

The univariate analyses presented in the previous sections demonstrated that several digital biomarkers extracted from wearable inertial data exhibit strong discriminative power between PD and HC. However, PD is inherently a multidimensional disorder, characterized by concurrent alterations across gait, tremor, upper-limb motor control, and global activity behaviour. Evaluating each feature independently, while informative, does not capture the joint structure of these impairments nor their combined discriminative potential. To address this limitation, a Multivariate Analysis of Variance (MANOVA) was performed to assess whether the overall multivariate motor profile differs between PD and HC when considering multiple digital biomarkers simultaneously.

Although a large number of features were initially extracted, including all of them in a multivariate model would be statistically inappropriate given the limited sample size (N = 20). To mitigate these risks while preserving clinical interpretability, a parsimonious feature set was selected. Specifically, one highly discriminative feature per motor domain was retained, based on prior effect size, statistical significance, and ROC performance. The selected features were:

- **STEP COUNT** – representative of the gait and locomotor activity domain.
- **PERCENTAGE WINDOWS TREMOR** – representative of tremor burden and constancy.
- **MEAN HAND AMPLITUDE** – representative of upper-limb hypokinesia.
- **LIGHT HOURS** – representative of global daily activity patterns.

This selection strategy ensured that each major Parkinsonian motor domain was represented while maintaining a favourable variable-to-subject ratio and minimizing redundancy. A one-way MANOVA was conducted with group (PD vs HC) as the independent variable and the four selected digital biomarkers as dependent variables (Figure 6.3). The analysis revealed a statistically significant multivariate group effect(p-value = 0.0181, lambda=0.475).

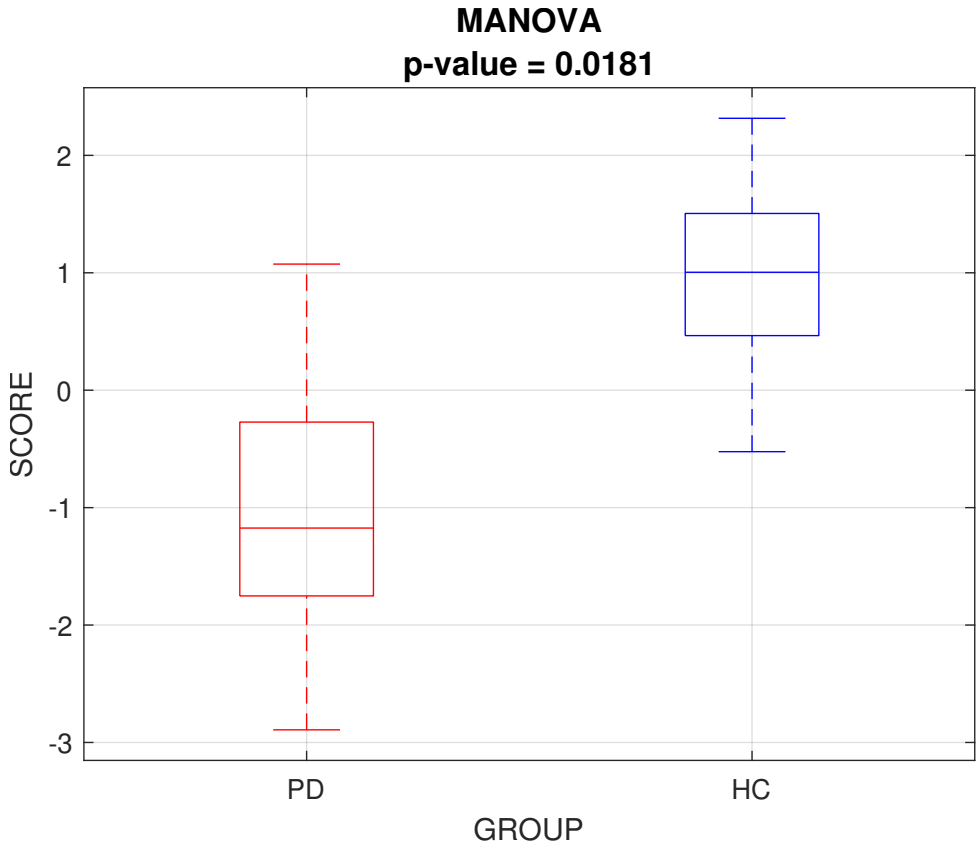


Figure 6.3: Results of the multivariate analysis of variance (MANOVA). A clear separation between PD and HC groups is observed, with a statistically significant multivariate effect.

This result indicates that, when considered jointly, the selected features define a multivariate motor signature that differs significantly between PD and HC. The significant MANOVA outcome confirms that the observed motor impairments in PD are not isolated phenomena but rather emerge as a coordinated pattern across multiple motor domains. While individual features such as tremor burden or hand movement amplitude already demonstrated strong univariate discrimination, the multivariate analysis shows that their combined expression provides additional discriminatory power. Clinically, the MANOVA findings reinforce the concept that PD should be interpreted through a multidomain digital phenotype rather than through single isolated metrics.

Chapter 7

Machine learning for group classification

The increasing availability of high-frequency digital biomarkers collected in real world settings has created new opportunities for the application of machine learning (ML) techniques in the study of PD.

The primary objective of this chapter is to describe the machine learning framework developed to discriminate between PD and HC using digital biomarkers extracted from wearable inertial sensors. Particular emphasis is placed on methodological rigor, prevention of data leakage, feature selection strategies, model optimization, and clinically meaningful evaluation at the patient level.

Multiple supervised learning algorithms were implemented and systematically compared, including Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM) (linear and radial basis function kernels), Random Forests, and Gradient Boosting (XGBoost). The performance of each model was evaluated under a unified experimental protocol designed to reflect realistic clinical deployment scenarios.

The classification results should be interpreted as exploratory. Given the limited sample size, the models may capture cohort-specific patterns rather than generalizable disease signatures. External validation on larger cohorts would be necessary to confirm robustness.

7.1 Hybrid machine learning approach

A central methodological challenge in the analysis of digital biomarkers for PD lies in the intrinsic mismatch between the unit of data acquisition and the unit of clinical decision-making. While wearable sensors generate high-frequency, repeated measurements at the daily level, clinical diagnoses and therapeutic decisions are made at the patient level. Treating these daily observations as independent samples in both training and testing phases can lead to overly optimistic performance estimates and limited clinical relevance.

To address this issue, a hybrid learning strategy was implemented, explicitly separating the learning and evaluation stages across different hierarchical levels of the dataset. In this frame-

work, machine learning models are trained using daily-level data, thereby exploiting the richness and variability of free-living motor behavior, while final performance is assessed at the patient level using aggregated representations. This approach ensures methodological rigor while preserving clinical interpretability.

A key motivation for adopting this hybrid strategy is the prevention of information leakage. In conventional cross-validation schemes applied directly to daily observations, samples from the same patient may appear in both training and testing folds. Since intra-subject motor patterns are often more similar than inter-subject patterns, this can artificially inflate classification performance without reflecting true generalization to unseen patients.

The hybrid evaluation was implemented using a leave-one-subject-out cross validation. For each iteration:

- One patient was selected as the test subject.
- All daily observations from this patient were excluded from the training set.
- The model was trained on daily observations from the remaining patients.
- The trained model was applied to the aggregated feature vector of the excluded patient.
- The predicted label was compared to the true clinical diagnosis.

This process was repeated for all subjects, yielding unbiased patient-level predictions for the entire cohort. Performance metrics such as accuracy, sensitivity, specificity, and confusion matrices were computed based on these predictions.

Rather than fixing the number of features a priori, an iterative procedure was employed:

- Models were trained using the top k ranked features found by the feature importance of the algorithm, with k varying from 1 up to our total number of features.
- For each k, hyperparameters were optimized via cross-validation using a grid search.
- Performance was evaluated using subject-aware validation schemes.

Following this iterative exploration, the final model configuration was defined by selecting the optimal k and the hyperparameter set that maximized the validation performance. This selection process ensured that the final architecture achieved the best trade-off between model complexity and predictive accuracy. This approach revealed that the optimal number of features differed across algorithms and evaluation levels, highlighting the importance of task-specific feature selection.

7.2 Machine learning algorithms for classification

7.2.1 Random Forest

Random Forest (RF) was employed both as a classifier and as a tool for estimating feature importance. Its ensemble structure, based on multiple decision trees trained on bootstrapped samples, makes it particularly robust to noise and multicollinearity, which are common in high-dimensional wearable-derived datasets. Random Forests were employed as a primary tool for feature ranking due to their robustness, ability to model nonlinear relationships, and inherent handling of feature interactions. As an ensemble of decision trees trained on bootstrapped samples with randomized feature subsets, Random Forests provide a natural measure of feature importance based on the reduction of impurity across splits. Feature importance was quantified using the mean decrease in impurity criterion. For each feature, its importance was computed as the average reduction in node impurity (Gini index) across all trees in the ensemble, weighted by the number of samples reaching each node. However, to ensure that the identified top features were genuinely tied to the underlying pathology rather than resulting from an algorithm-specific mathematical bias, a validation approach was adopted. The analysis was not limited solely to RF; when other classification algorithms possessing native variable importance metrics were employed their respective feature hierarchies were also extracted and recorded. Ultimately, the feature importance rankings produced by these diverse algorithms were compared . Verifying that a specific metric is consistently recognized as highly discriminative across mathematically distinct algorithms provides evidence that the feature represents a true, clinically relevant pathological marker, rather than a mere computational artifact. This process yielded a global ranking of features, reflecting their overall contribution to class separation between PD and HC (Figure 7.1).

Among all extracted features, the most informative biomarkers consistently identified by the Random Forest were:

- PERCENTAGE WINDOWS TREMOR
- MEAN HAND AMPLITUDE
- STEPCOUNT
- CADENCE CV INTER BOUT
- X90P TREMOR POWER

These features capture complementary aspects of PD, spanning tremor consistency, bradykinesia, and gait quantity and variability, reflecting the multifaceted nature of parkinsonian motor deficits.

Model performance was obtained following a dedicated hyperparameter optimization procedure conducted within the hybrid training framework. The optimal configuration identified

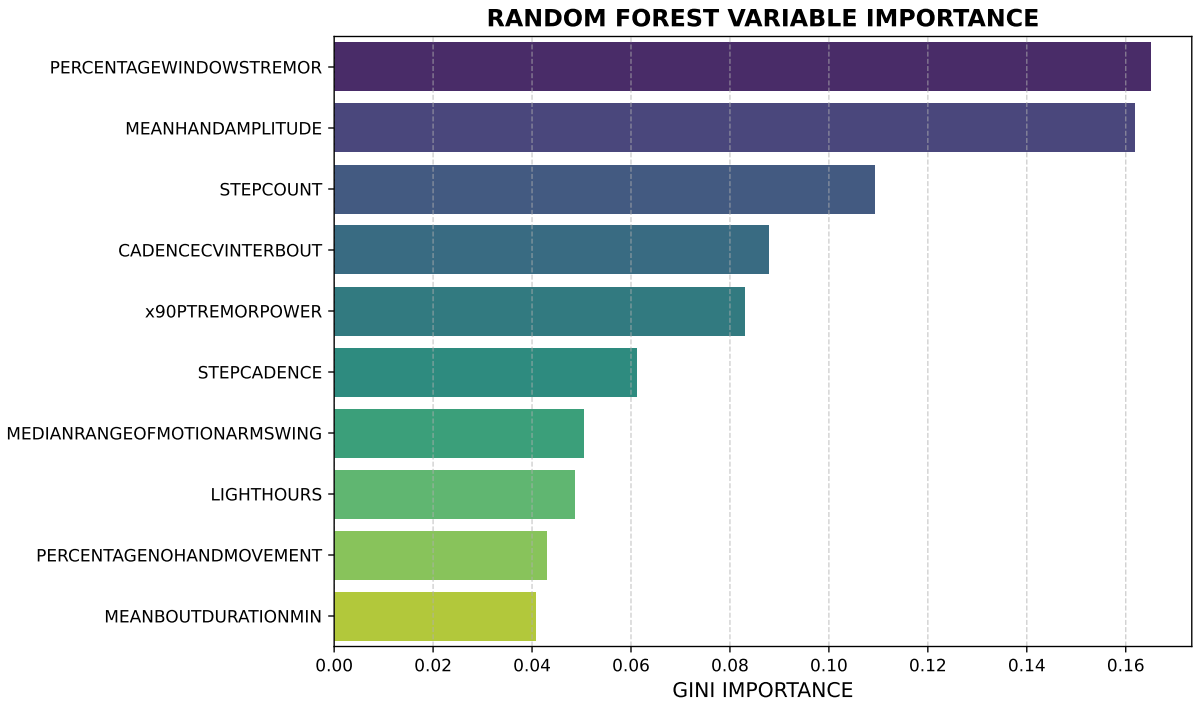


Figure 7.1: Variable importance ranking with the Random Forest approach. Importance is quantified as the mean decrease in Gini impurity across all trees.

five features as the most informative subset, balancing discriminative power and model simplicity. The Random Forest architecture, after a grid search, was finalized with 100 decision trees (`n_estimators = 100`) and no constraint on tree depth (`max_depth = None`), allowing each tree to fully capture complex feature interactions. This configuration provided the best trade-off between bias and variance, maximizing classification performance at the patient level.

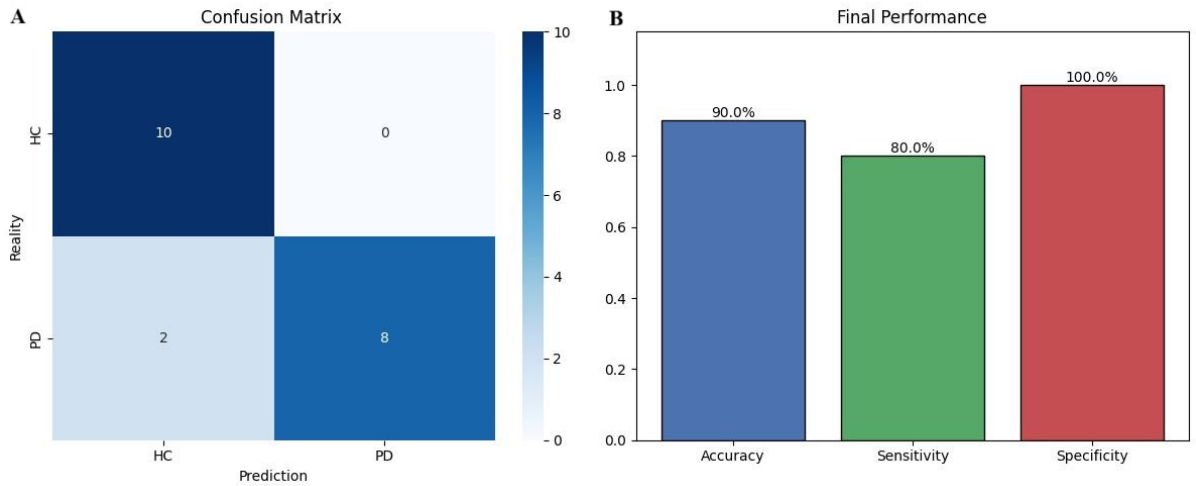


Figure 7.2: Visualization of the performance of the algorithm of random forest. A displays the confusion matrix, while B displays the accuracy, the sensitivity and the specificity. The model successfully avoids false positives (Specificity = 100%), while two PD patients were misclassified as HC.

The Random Forest classifier demonstrated strong performance within the proposed hybrid

evaluation framework, achieving an overall classification accuracy of 90.0% at the patient level (see Figure 7.2). This result shows the effectiveness of ensemble-based methods in capturing the complex, non-linear relationships embedded in wearable-derived digital biomarkers.

Beyond accuracy, the model exhibited a sensitivity of 80.0% and a specificity of 100.0%, highlighting a clinically relevant asymmetry in classification performance. The perfect specificity indicates that the Random Forest model correctly identified all HC subjects, with no false-positive classifications, minimizing the risk of incorrectly labelling healthy individuals as affected by Parkinson's disease.

The slightly lower sensitivity suggests that a small number of PD patients were misclassified as healthy. While this reflects a degree of conservativeness in the model, it is consistent with the intrinsic heterogeneity of PD, particularly in early or mildly affected patients whose motor patterns may overlap with physiological variability. Importantly, the achieved sensitivity remains high considering the challenging nature of free-living data and the limited sample size at the patient level.

7.2.2 Linear Support Vector Machine

The Linear Support Vector Machine (SVM) classifier was implemented as a complementary, interpretable model within the proposed framework. Unlike ensemble-based approaches, the linear SVM relies on a single separating hyperplane, making it particularly suitable for evaluating the discriminative contribution of individual features and for assessing model robustness under linear assumptions.

A key advantage of the linear SVM in this study was the direct exploitation of SVM-derived feature importance (Figure 7.3), which provides a fundamentally different and complementary perspective compared to the Random Forest approach. While the RF model computes importance based on the mean decrease in impurity across multiple decision splits, the linear SVM derives importance directly from the absolute magnitude of its coefficients. This parametric approach not only quantifies the absolute discriminative power of each metric but also intrinsically provides directionality. By analyzing the sign of the SVM coefficients, it is possible to explicitly determine whether an increase in a specific feature drives the algorithm's prediction toward the PD or the HC group, enhancing the clinical interpretability of the entire framework.

The best configuration was obtained with a regularization parameter $C = 0.01$, enforcing a strongly regularized solution. This choice effectively reduced sensitivity to noise and subject-specific variability, which are common challenges in free-living wearable data. The optimal feature subset that reached the highest accuracy consisted of six features, selected based on their ranked importance from the SVM coefficients: PERCENTAGE WINDORS TREMOR, MEAN BOUT DURATION MIN, CADENCE CV INTER BOUT, LIGHT HOURS, STEP CADENCE, STEP COUNT.

Within the hybrid evaluation strategy (training on daily observations and testing at the pa-

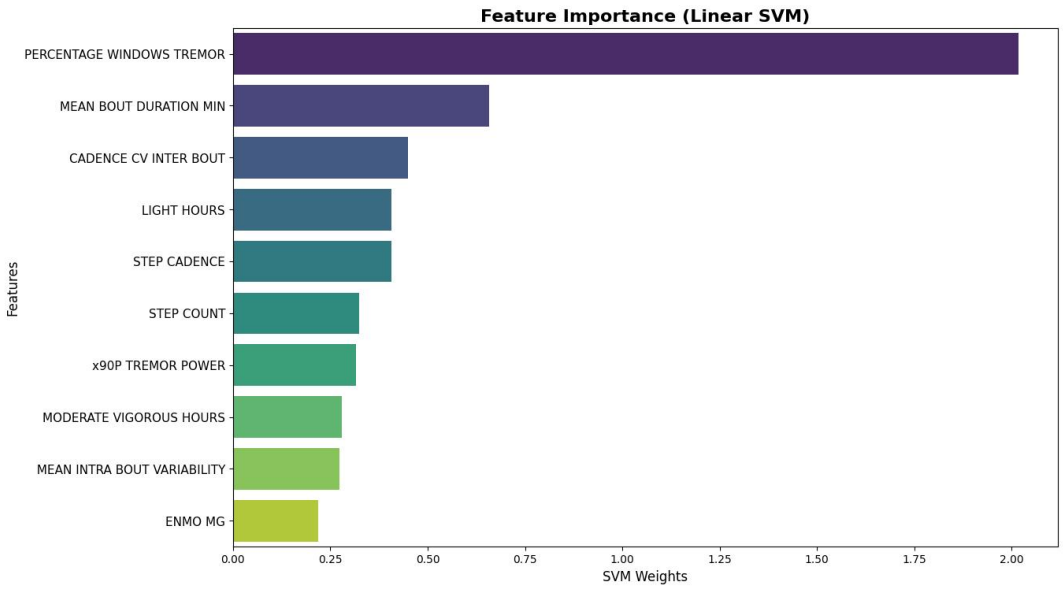


Figure 7.3: Feature importance obtained from the SVM algorithm. Importance is derived from the absolute magnitude of the SVM coefficients.

tient level), the linear SVM achieved an overall classification accuracy of 85.0% . The model demonstrated a sensitivity of 80.0%, while maintaining a specificity of 90.0%, reflecting a low false-positive rate among HC. Although slightly lower than the performance achieved by the random forest method, these results are notable given the simplicity and strong regularization of the linear model.

7.2.3 Support Vector Machine with Radial Basis Function

The Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel was employed as a non-linear classification approach to distinguish between HC and PD. Unlike the linear SVM, the RBF kernel allows the model to capture non-linear relationships between features, potentially improving discriminative performance when the separation boundary between classes is complex.

Hyperparameter tuning was conducted to identify the optimal values for the regularization parameter C and the RBF kernel parameter γ. The tuning process indicated that the best performance was obtained with $C = 1$ and $\gamma = scale$ and six features (the same of the linear function).

The algorithm achieved an accuracy of 90.0%, demonstrating a sensitivity of 80.0% and a specificity of 100.0% . These results demonstrate that the RBF SVM was able to perfectly identify HC while maintaining high sensitivity for PD patients. The slightly higher performance compared to the linear SVM suggests that non-linear relationships between features contribute significantly to discriminating between the two groups.

7.2.4 K-Nearest Neighbors

The k-Nearest Neighbors (k-NN) algorithm was implemented as a non-parametric method for classifying PD from HC. k-NN classifies a sample based on the majority label among its k nearest neighbors in the feature space, making it highly interpretable and sensitive to local patterns in the data. Based on the random forest feature importance ranking, to reach the best performance, the top four features were selected for k-NN. Hyperparameter tuning was then performed to identify the optimal number of neighbors (k). The best cross-validation performance was achieved using $k = 4$ and the Euclidean distance, ensuring a balance between sensitivity to local data structures and robustness to noise. The classification results at the patient level demonstrated excellent predictive capabilities, achieving an overall accuracy of 90.0% . Specifically, the classifier yielded a sensitivity of 80.0% and a perfect specificity of 100.0%. These results demonstrate that k-NN was able to perfectly identify HC while maintaining high sensitivity for PD. The high specificity reflects the model's effectiveness in avoiding false positives, while the choice of a small number of neighbors allowed the classifier to capture subtle variations between patients.

7.2.5 Logistic Regression

Consistent with the unified feature engineering strategy adopted throughout this study, the model was trained using a subset of the most informative digital biomarkers. Feature importance was derived from the Random Forest ranking, and the final configuration retained six features, which provided the best trade-off between model complexity and generalization performance. These features capture complementary aspects of motor impairment, including tremor burden, bradykinesia, and gait dynamics. The Logistic Regression model yielded excellent classification performance (Accuracy: 90.0%; Sensitivity: 80.0%; Specificity: 100.0%), confirming the linear separability of the features. These results indicate an excellent ability to correctly identify HC while maintaining high sensitivity toward PD. The high specificity suggests that the model is particularly conservative in assigning a pathological label, minimizing false positives—an important property in clinical screening contexts. One of the primary advantages of logistic regression lies in the interpretability of its coefficients. Each model coefficient can be transformed into an odds ratio, quantifying how a one-unit increase in a given feature affects the probability of belonging to the PD group while holding all other variables constant.

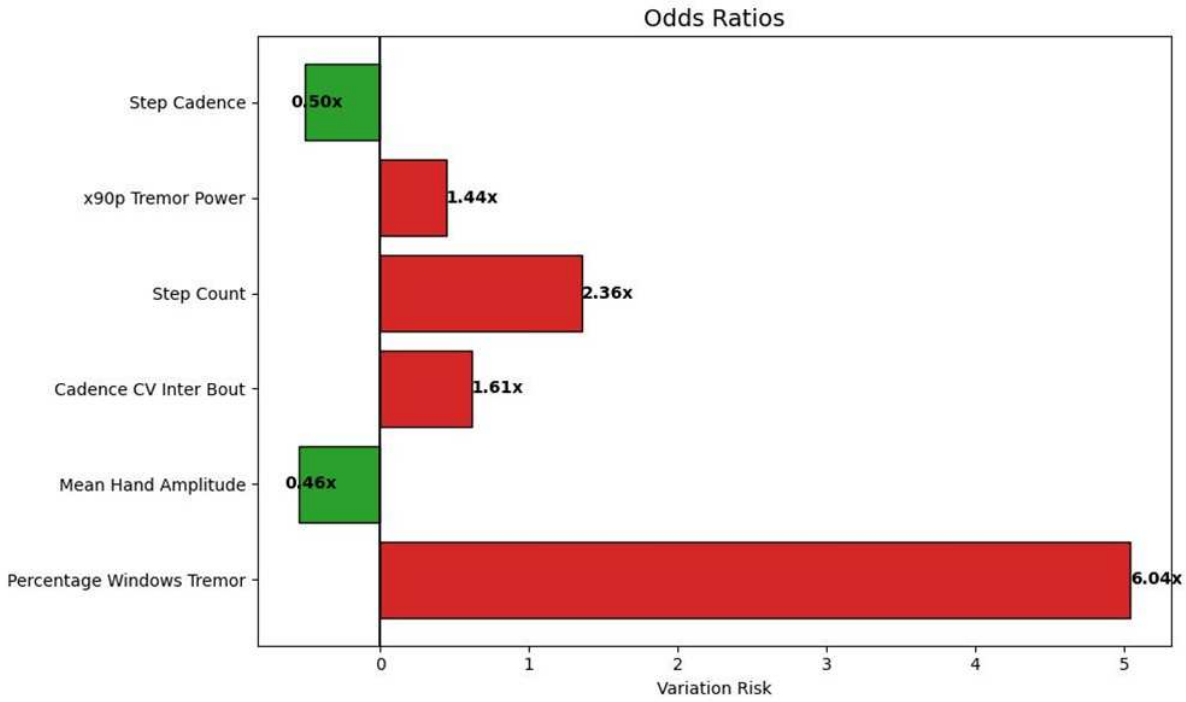


Figure 7.4: Visualization of the Odds Ratio of the logistic regression algorithm. Bars represent the change in odds of belonging to the PD group for a one-unit increase in each feature.

The odds ratio plot provides an intuitive visualization of feature relevance, highlighting both the direction and magnitude of each predictor's contribution to the classification decision (Figure 7.4).

The analysis highlighted PERCENTAGE WINDOWS TREMOR as the most influential predictor, with an odds ratio of 6.04, indicating that higher tremor prevalence dramatically increases the likelihood of PD classification. This result is consistent with tremor being one of the hallmark motor manifestations of Parkinson's disease and confirms the clinical relevance of tremor-related digital biomarkers.

Overall, the odds ratio profile underscores the complementary nature of the selected features: tremor-related metrics strongly increase PD likelihood, while amplitude- and rhythm-related features reflect motor preservation. This balanced contribution reinforces the interpretability of the Logistic Regression model and provides clinically intuitive insights into how distinct motor domains jointly define the digital phenotype of PD.

7.2.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm based on gradient-boosted decision trees, designed to model complex non-linear relationships through the sequential combination of weak learners. Unlike Random Forests, where trees are trained independently, XGBoost builds trees iteratively, with each new tree correcting the errors of the previous ensemble. This makes XGBoost particularly effective in capturing subtle interactions between

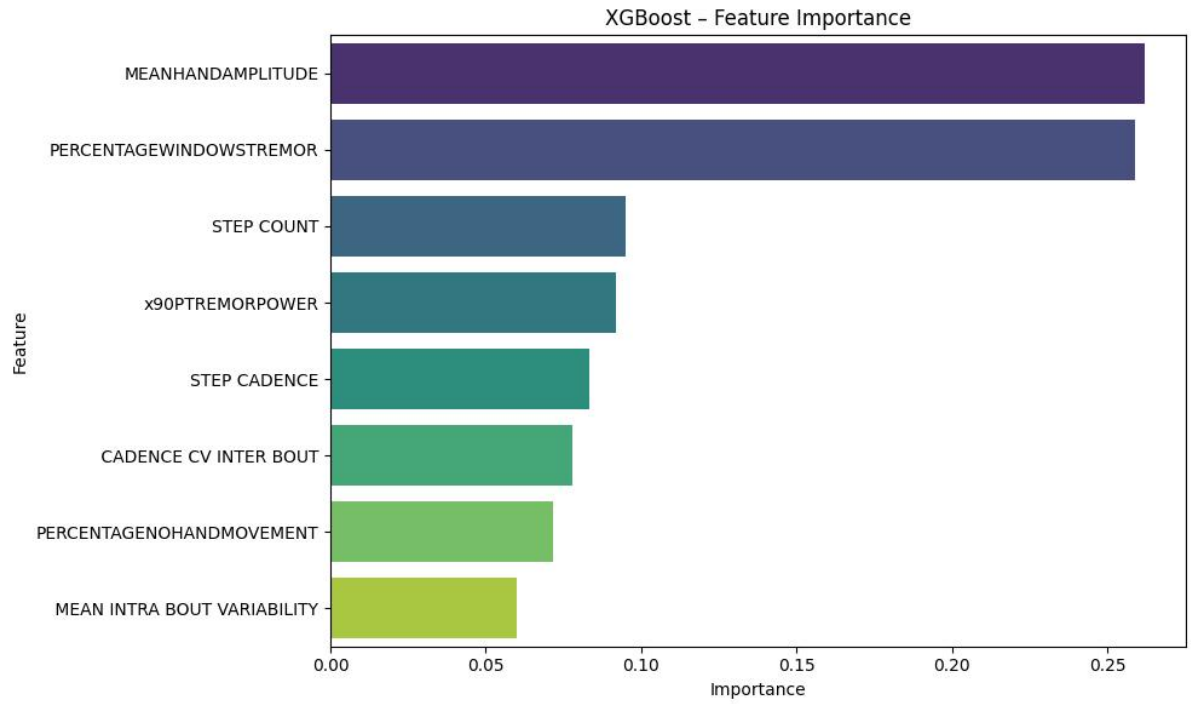


Figure 7.5: Feature importance provided by the XGBoost algorithm, derived from the frequency of feature usage across boosting iterations.

features, at the cost of increased sensitivity to noise and class imbalance.

The optimal configuration identified through the tuning procedure, estimating the feature importance using the XGBoost algorithm (Figure 7.5), was:

- Number of features: 8
- Learning rate: 0.05
- Maximum tree depth: 2
- Number of estimators: 200

The relatively low learning rate and shallow tree depth reflect a conservative boosting strategy, aimed at limiting overfitting and promoting stable generalization in a dataset characterized by a limited number of subjects and longitudinal measurements.

The final XGBoost model achieved a solid classification performance at the patient level, yielding an overall accuracy of 85.0% , with a sensitivity of 70.0% and a specificity of 100.0%. The model demonstrated perfect specificity, correctly identifying all HC, while showing reduced sensitivity compared to other classifiers. Surprisingly, XGBoost underperformed compared to Random Forest, likely due to overfitting. This performance profile, characterized by perfect specificity but reduced sensitivity, indicates that the model adopts a conservative decision boundary, correctly identifying all HC at the cost of misclassifying a subset of PD patients with milder or less pronounced symptom manifestations.

7.3 Comparative analysis of machine learning models

To provide a comprehensive evaluation of the proposed classification framework, a comparative analysis of all implemented machine learning models was conducted. Rather than focusing exclusively on individual algorithmic performance, this section aims to assess the relative strengths and limitations of each model in terms of classification accuracy, sensitivity, specificity, model complexity, and discriminative capacity. This comparative perspective is essential to identify the most suitable approach for the discrimination between PD and HC using digital motor biomarkers derived from wearable sensors.

A critical aspect of this comparative analysis is understanding how different algorithms prioritize physiological features, as illustrated in the cross-algorithm feature importance heatmap (Figure 7.6). By normalizing the importance scores from 0 to 1, several distinct patterns emerge, providing profound insights into the nature of the extracted biomarkers.

Most notably, PERCENTAGE WINDOWS TREMOR achieved a perfect normalized score of 1.00 across all three mathematically distinct algorithms. This absolute consensus confirms that the temporal constancy of tremor is the strongest and most robust global discriminator in the dataset, completely independent of the underlying classification logic. Similarly, features such as STEP COUNT and CADENCE CV INTER BOUT demonstrated solid predictive relevance across the board, reinforcing their role as intrinsic pathological markers.

Interestingly, the comparative visualization highlights significant non-linear dynamics in specific motor symptoms. MEAN HAND AMPLITUDE exhibited near-maximum importance in tree-based ensemble models (0.98 in RF and 0.89 in XGBoost) but dropped in the SVM (0.23). This discrepancy suggests that the reduction in hand amplitude interacts with PD pathology in a highly complex, non-linear manner, which decision trees can easily isolate but a linear hyperplane fails to separate cleanly. Conversely, the SVM prioritized different temporal metrics, such as MEAN BOUT DURATION MIN (0.66), which was largely ignored by the tree-based models (0.15 in RF and 0.07 in XGBoost). Finally, features located at the bottom of the hierarchy, such as MODAL TREMOR POWER, MEDIAN TREMOR POWER, and ENMO (MG), consistently scored near zero across all algorithms, suggesting a lack of independent discriminative power, likely due to physiological redundancy with higher-ranking metrics.

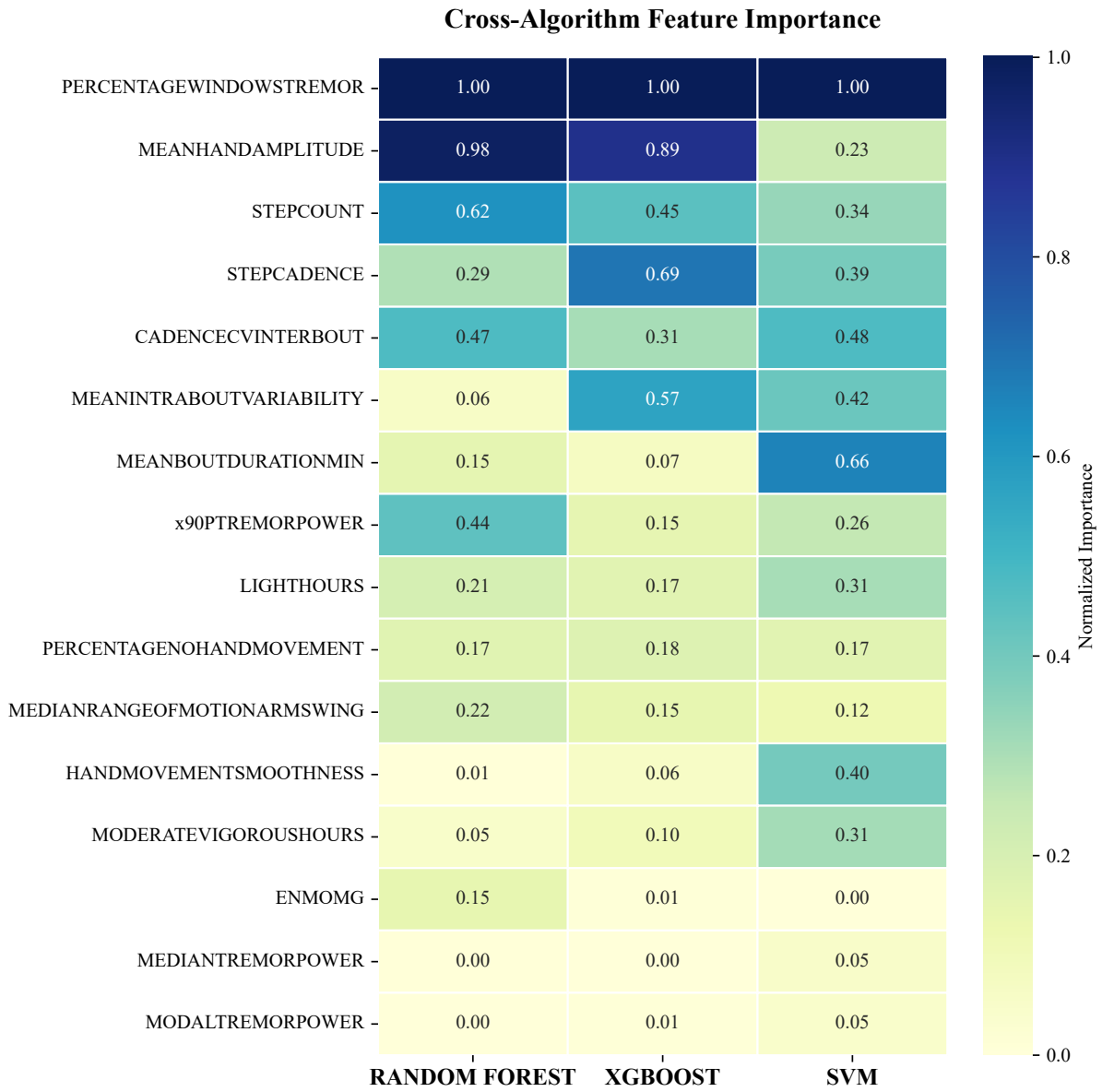


Figure 7.6: Heatmap of the variable importances of the algorithms.

Model	Accuracy	Sensitivity	Specificity	Features	AUC
Random Forest	90.0%	80.0%	100.0%	5	1.00
Linear SVM	85.0%	80.0%	90.0%	6	0.91
RBF SVM	90.0%	80.0%	100.0%	6	1.00
k-NN	90.0%	80.0%	100.0%	4	0.86
Logistic Regression	90.0%	80.0%	100.0%	6	0.89
XGBoost	85.0%	70.0%	100.0%	8	0.72

Table 7.1: Final performance comparison of the machine learning algorithms for the classification of PD and HC. Reported metrics include accuracy, sensitivity, specificity, and number of selected features after hyperparameter optimization.

Table 7.1 summarizes the final performance metrics achieved by each machine learning algorithm after hyperparameter optimization and feature selection. In addition to standard classification metrics, the table reports the number of selected features for each model, highlighting differences in model complexity and feature dependency.

Overall, multiple algorithms achieved high classification performance, with accuracies ranging from 85.0% to 90.0%. Random Forest, RBF-SVM, k-Nearest Neighbors, and Logistic Regression all reached an accuracy of 90.0%, demonstrating that digital motor features extracted from daily life conditions contain sufficient discriminative information to reliably separate PD patients from healthy individuals. A notable observation across models is the consistently high specificity, which reached 100% in most classifiers. This indicates a strong ability to correctly identify HC, minimizing false positive classifications. High specificity is particularly valuable, as it reduces the risk of misclassifying healthy subjects as pathological. Sensitivity values were slightly lower and more variable, ranging from 70.0% to 80.0%. This reflects the intrinsic heterogeneity of Parkinson's disease, especially in early or mildly affected patients, whose motor patterns may partially overlap with normative behavior.

To properly contextualize the performance achieved by the proposed hybrid classification framework, it is essential to compare these results with recent literature focusing on Parkinson's Disease detection using wrist-worn inertial sensors in free-living conditions. While advanced Deep Learning models evaluated on large datasets [31] achieved a comparable accuracy of 91.1%, our framework reaches a highly competitive 90.0% accuracy utilizing strictly interpretable, white-box models. This represents a significant advantage in a clinical setting, where the interpretability of the decision-making process is as crucial as the raw performance metrics.

Furthermore, our approach outperforms standard passive monitoring studies. For instance, recent studies relying solely on passive wrist accelerometry without extensive feature engineering have reported baseline accuracies of approximately 85.0% with the RF and the XGBoost algorithms [32]. Ultimately, the ability to achieve state-of-the-art discriminative power without sacrificing physiological interpretability validates the proposed pipeline as a robust and reliable tool for continuous PD monitoring and for group discrimination.

7.3.1 Accuracy and model complexity

A key dimension of the comparative analysis concerns the relationship between classification accuracy and model complexity, here quantified by the number of selected features. In clinical machine learning, achieving high performance with a minimal number of features is desirable, as it improves generalizability, reduces overfitting risk, and facilitates clinical interpretation and implementation. Although several models achieved comparable accuracy values, they differed substantially in the dimensionality of the feature space required to reach optimal performance. Random Forest, RBF-SVM, k-Nearest Neighbors, and Logistic Regression all attained an accuracy of 90.0%, yet with a varying number of features ranging from four to six.

In particular, k-NN reached peak performance using only four features, while Random Forest achieved similar accuracy with five features, highlighting their ability to extract discriminative information efficiently from a compact digital phenotype. The Support Vector Machines (RBF) and Logistic Regression required 6 features to achieve their optimal performance. While this represents a slightly higher complexity, the feature set remains clinically manageable and still reflects a compact digital phenotype.

In contrast, XGBoost required a larger feature set (eight features) to reach a lower accuracy (85.0%), suggesting that the increased flexibility of boosting-based models may not translate into improved generalization. Instead, it may introduce sensitivity to noise and subtle overfitting effects, even under conservative hyperparameter settings.

These results emphasize that high classification accuracy alone does not necessarily imply an optimal model, especially in small-sample clinical settings. Models that achieve comparable performance with fewer features are generally preferable, as they reduce the risk of overfitting, improve generalizability, and facilitate clinical interpretation.

7.3.2 Sensitivity-specificity trade off

A consistent pattern across all models was a marked asymmetry between sensitivity (70.0-80.0%) and specificity (90.0-100.0%). The models excelled at identifying HC, indicating that the digital profile of normality derived from the selected biomarkers is stable and distinct. The errors were almost exclusively False Negatives where PD patients were misclassified as healthy, identifying these models as conservative classifiers.

Before interpreting this asymmetry from a physiological perspective, it is important to acknowledge that it may partially reflect a statistical artifact introduced by the limited cohort size. With only 10 PD patients and 10 HC subjects, the cross-validation evaluates each model on a single subject at a time, meaning that each individual prediction carries a high weight on the final performance metrics. In this context, the perfect specificity of 100.0% observed across most classifiers must be interpreted with caution: with n=10 HC subjects, a single misclassification would shift specificity by 10 percentage points, making the estimate unstable.
However, beyond this statistical consideration, the asymmetry also carries a genuine physio-

logical interpretation that is consistent with the clinical characteristics of our cohort, suggesting that the conservative behavior of the classifiers is not entirely attributable to sampling bias. Two main factors contribute to this pattern.

First, the influence of medication effects (ON State) plays a critical role: the dataset likely includes prolonged periods where patients were under dopaminergic therapy. During these ON states, motor symptoms such as tremor and bradykinesia may be attenuated to near-physiological levels, making the digital motor signature of PD closely resemble that of HC.

Second, the cohort includes patients with mild disease severity whose motor impairments are subtle and intermittent, rather than continuous and persistent. Crucially, an analysis of the clinical metadata confirmed that the subset of misclassified PD patients exhibited lower MDS-UPDRS Part III scores compared to the correctly identified patients. Specifically, the misclassified subjects P-3061 and P-3119 presented scores of 10 and 19 respectively, significantly below the cohort mean of 27.8. This quantitative evidence suggests that the classifier's errors correspond to subjects with sub-threshold motor signs, where the pathological signal is too weak to be distinguished from physiological variability. Furthermore, the aggregation strategy inherent to the hybrid approach may further contribute to this effect. By averaging daily features to generate a patient-level prediction, sporadic pathological episodes may be diluted by extended periods of near-normal motor behavior. As a result, early-stage patients with fluctuating symptoms may be pushed toward the healthy class in the final decision space.

To explicitly visualize this phenomenon and understand the algorithm's decision-making process, a feature-level profiling was conducted on the False Negatives. Figure 7.11 illustrates a radar chart comparing the standardized digital motor signature of a misclassified patient against the average profiles of HC and PD patients. As visually evident, the multidimensional feature geometries of subject P-3119 collapse almost entirely onto the physiological HC baseline, strongly deviating from the typical pathological PD profile. When examining the core discriminative metrics—specifically Percentage Windows Tremor, Cadence CV Inter-Bout, Step Count, and Mean Hand Amplitude—the patient's values fall squarely within the healthy range. The only notable exception in this digital signature is the 90th Percentile Tremor Power, which still retains a pathological deviation. However, this isolated pathological sign is mathematically overpowered by the physiological behavior of all the other predominant features, causing the algorithm to ultimately misclassify the overall digital motor profile as healthy.

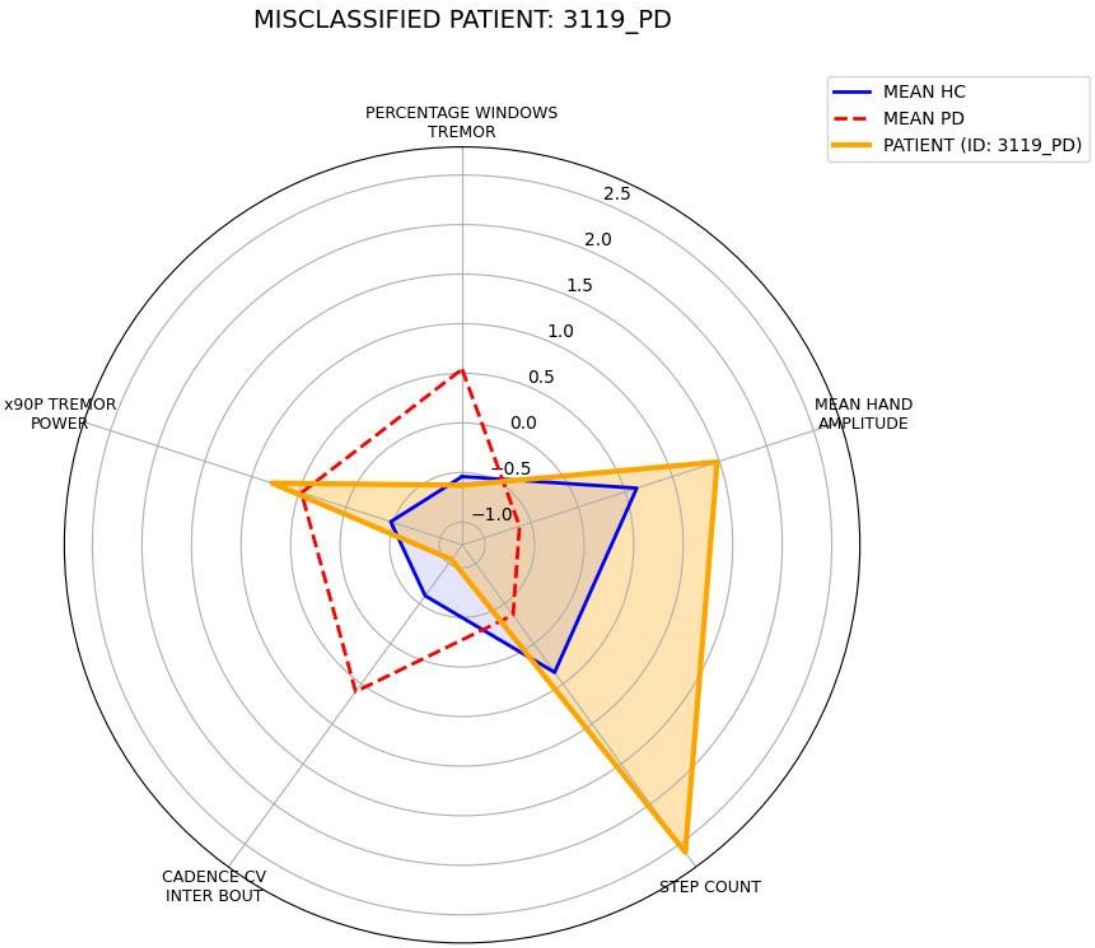


Figure 7.7: Radar chart comparing the standardized motor features (Z-scores) of patient 3119 against the average profile of HC and PD. The radial axes represent standard deviations from the global dataset mean. The patient's feature geometry closely resembles the HC baseline across most discriminative dimensions, explaining the misclassification

These findings suggest that the observed errors do not reflect a fundamental limitation of the machine learning models, but rather the physiological overlap between well-controlled or early-stage PD patients and healthy individuals in daily life conditions. Importantly, this behavior is clinically meaningful: the models appear to correctly flag subjects only when motor impairment is sufficiently and consistently expressed in daily-life movement patterns, reducing the risk of false alarms in a potential screening scenario.

7.3.3 Receiver Operating Characteristic curve analysis

While accuracy, sensitivity, and specificity provide threshold-dependent performance measures, Receiver Operating Characteristic (ROC) curve analysis offers a threshold-independent evaluation of classifier discrimination capability [33].

Random Forest and RBF SVM achieved high AUC (1.00) (Table 7.1), showing a perfect probability separation between the groups. Patient-level accuracy is slightly lower (90.0%), meaning that although the ranking of predicted probabilities is perfect, the threshold used for

classification led to a few misclassifications. The discrepancy arises when the predicted probabilities are perfectly ordered but not perfectly calibrated around this default threshold.

This suggests that while the models are capable of assigning higher disease probabilities to PD patients compared to HC, the default decision threshold resulted in the misclassification of a small subset of patients who likely presented with borderline probability scores.

Linear SVM (AUC = 0.91), Logistic Regression (AUC = 0.89), k-NN (AUC = 0.86) and XGBoost (AUC = 0.72) showed progressively lower discriminative power, although still well above chance. These differences highlight that models with similar accuracy may differ substantially in their probability calibration and ranking ability, an important consideration for clinical decision support systems.

7.3.4 Model selection

The comparative analysis of Accuracy and Area Under the Curve (ROC-AUC) reveals a remarkable consistency across the evaluated classifiers. The fact that multiple algorithms—ranging from simple linear boundaries to complex non-linear ensembles—achieved comparably high accuracies (up to 90.0%) and near-perfect specificities is a crucial clinical finding. This algorithmic consensus demonstrates that the extracted digital motor features contain a strong, intrinsic physiological signal that reliably discriminates PD patients from healthy controls, rather than relying on the mathematical idiosyncrasies of a single classifier.

However, when evaluating the models holistically—balancing pure discriminative power, robustness to noise, and clinical interpretability—Random Forest unequivocally emerges as the superior model and the definitive choice for this diagnostic framework.

Although its top-tier accuracy (90.0%) and outstanding AUC (1.00) were matched by the RBF-SVM, Random Forest proved significantly more efficient and resilient [33]. Random Forest achieved its peak performance using a highly parsimonious set of only 5 features, whereas the RBF-SVM required 6 features to stabilize its decision boundary. This indicates that Random Forest extracted the core digital phenotype of PD more efficiently, minimizing the risk of overfitting and maximizing generalizability to new patients.

Random Forest successfully bridges these gaps, offering the predictive power of a non-linear ensemble. By providing a direct, quantifiable measure of feature importance through the mean decrease in impurity, Random Forest acts as an explainable AI tool. This transparency allows clinicians to continuously validate the model, confirming that the algorithmic output is driven by clinically relevant motor signs rather than spurious background noise.

This chapter validates the diagnostic utility of wearable-derived digital biomarkers for PD detection. While the high classification consistency observed across all tested models serves as a strong validation of the feature engineering process, Random Forest stands out as the optimal predictive engine.

Chapter 8

Concurrent validity and clinical associations of digital biomarkers

The extraction of digital features from inertial sensors, however technically sophisticated, acquires a higher medical value if it demonstrates a meaningful correspondence with current clinical standards. This process, known as concurrent validity, aims to verify whether the objective measures provided by the sensor correlate significantly with the subjective rating scales administered by neurologists. Without this validation step, even the most advanced signal processing algorithm remains a ”black box” with uncertain clinical utility.

A robust digital biomarker must not only detect the presence of pathology but also accurately track its severity gradient. It must demonstrate a monotonic relationship with clinical status: as the patient’s condition deteriorates (higher UPDRS or H&Y scores), the digital feature should exhibit a proportional change (e.g., lower smoothness, higher tremor power). If a digital feature correlates strongly with a clinical scale, it implies that the sensor captures the same physiological construct assessed by the neurologist [34]. Consequently, such features can potentially serve as an objective, continuous surrogate marker of motor impairment, capable of detecting subtle intra-day fluctuations that might be completely missed between episodic clinical visits.

In the present study, the concurrent validity was investigated by comparing digital features extracted during home monitoring with three ”Gold Standard” clinical scales. Since Inertial Measurement Units specifically quantify kinematic and dynamic parameters of physical movement, the correlation analysis was strictly focused on clinical scales assessing motor function and functional mobility. Non-motor scales (such as the MDS-UPDRS Part I for non-motor experiences or the MoCA for cognitive screening) were excluded from this primary validation. While cognitive and psychiatric symptoms are integral to PD, an inertial sensor is not the appropriate instrument to measure them directly.

This comparison is particularly challenging yet crucial because it attempts to correlate continuous data collected in a natural, uncontrolled environment with snapshot assessments performed in a controlled clinical setting. The clinical ground truth was established using the fol-

lowing standardized scales:

MDS-UPDRS Part II: Evaluates motor experiences of daily living (e.g., dressing, walking, getting out of bed), reflecting the functional disability perceived by the patient.

MDS-UPDRS Part III (ON state): Evaluates the objective severity of motor signs (bradykinesia, rigidity, tremor) through a standardized neurological examination.

Given the ordinal and non-Gaussian nature of clinical scores, statistical analysis was performed using Spearman's rank correlation coefficient, which assesses monotonic relationships based on rank ordering. This method is robust to outliers and does not require the assumption of linearity, making it ideal for relating continuous sensor data to ordinal clinical ratings.

The analysis was restricted to the PD cohort to assess the sensor's sensitivity to disease severity gradients, excluding HC to prevent the artificial inflation of correlation coefficients. By analyzing only the PD cohort, this study rigorously tests the sensor's sensitivity to disease severity (i.e., distinguishing mild PD from severe PD), rather than merely confirming its ability to distinguish sick from healthy.

8.1 Correlation with standardized clinical scales

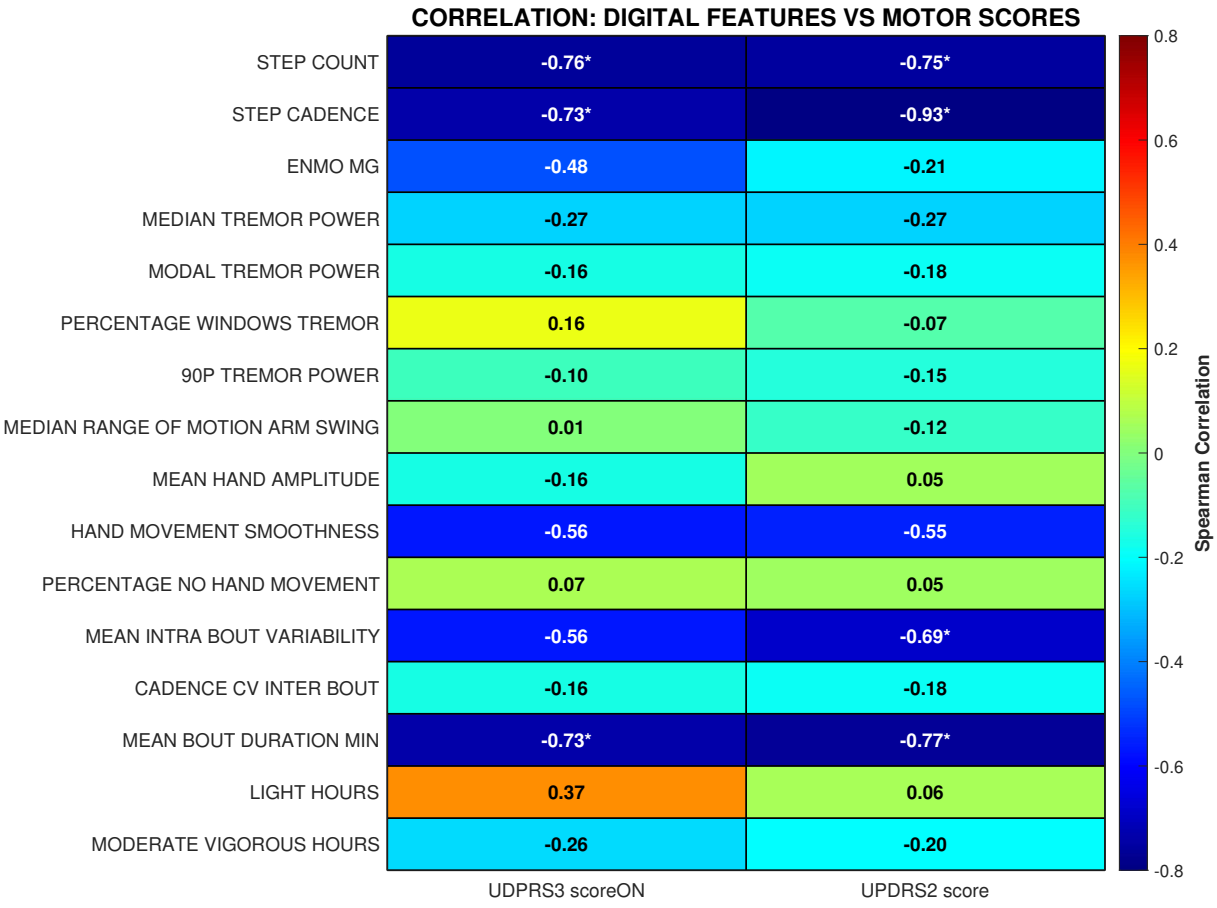
The analysis revealed statistically significant correlations with magnitudes ranging from moderate to strong ($r > 0.7$) between specific digital domains and their clinical counterparts. The results outline a coherent picture in which the sensor is capable of distinctly capturing functional disability and the residual motor deficits present during pharmacological treatment.

Figure 8.1 provides a comprehensive visualization (heatmap) of the Spearman correlation coefficients between the 16 extracted digital features (rows) and the clinical scales (columns). The color intensity encodes the magnitude of the Spearman coefficient, with darker shades representing stronger negative associations. The prevalence of negative coefficients reflects the inverse relationship between motor features and clinical severity.

Broadly, the correlation profile can be divided into two main behaviors. Features related to global mobility, gait rhythm (like STEP CADENCE and MEAN BOUT DURATION MIN), and movement variability exhibit moderate to exceptionally strong associations (r>0.6) across the board. Conversely, tremor-specific features (such as PERCENTAGE WINDOWS TREMOR and MEDIAN TREMOR POWER) do not show significant linear correlations with either of the composite UPDRS scales.

8.1.1 Step cadence as a proxy for functional disability (MDS-UPDRS II)

The most relevant result of the entire study emerges from the analysis of gait related digital biomarkers in relation to functional disability, as assessed by the MDS-UPDRS Part II. Among



*** Correlation is significant at the 0.05 level**

Figure 8.1: Spearman correlation of the final set of features. Color intensity represents correlation magnitude, with darker shades indicating stronger associations. Asterisks (*) denote statistical significance based on nominal p-values ($p<0.05$) without adjustment for multiple comparisons.

all investigated features, gait rhythm parameters demonstrated the strongest and most direct association with patients' self-reported difficulties in activities of daily living, highlighting gait as a central proxy for functional autonomy in PD.

In particular, STEP CADENCE showed an excellent negative correlation with the MDS-UPDRS Part II score ($r = -0.93$, $p = 0.0001$) as is shown in Figure 8.2. The negative sign indicates an inverse relationship: patients with reduced cadence (slow and shuffling gait) report significantly higher UPDRS II scores, indicative of greater functional impairment [35]. The strength of this correlation is remarkable and suggests that cadence captures a fundamental dimension of disease impact that extends beyond isolated motor symptoms. The magnitude of this correlation can be explained by the specific composition of the MDS-UPDRS Part II scale, which heavily weights mobility-dependent tasks such as walking and dressing. Gait cadence represents the temporal organization and automaticity of stepping. When this rhythmic automaticity deteriorates, patients must increasingly rely on conscious control to initiate and maintain movement, resulting in slower gait and greater difficulty performing routine daily activities. Consequently, impairments in cadence translate directly into higher disability scores.

Similarly, MEAN BOUT DURATION showed a strong negative correlation with the MDS-UPDRS Part II score ($r = -0.77, p = 0.0098$). Shorter walking bouts reflect a fragmented locomotor pattern, in which patients are unable to sustain continuous walking for extended periods. This fragmentation is a well-recognized hallmark of thedisease and is often driven by fatigue, postural instability, fear of falling, or freezing of gait [36]. From a functional perspective, the inability to maintain prolonged walking severely limits independence, forcing patients to interrupt daily activities and rely on frequent rest periods. The wearable-derived bout structure thus provides an objective and ecologically valid representation of real-world mobility limitations.

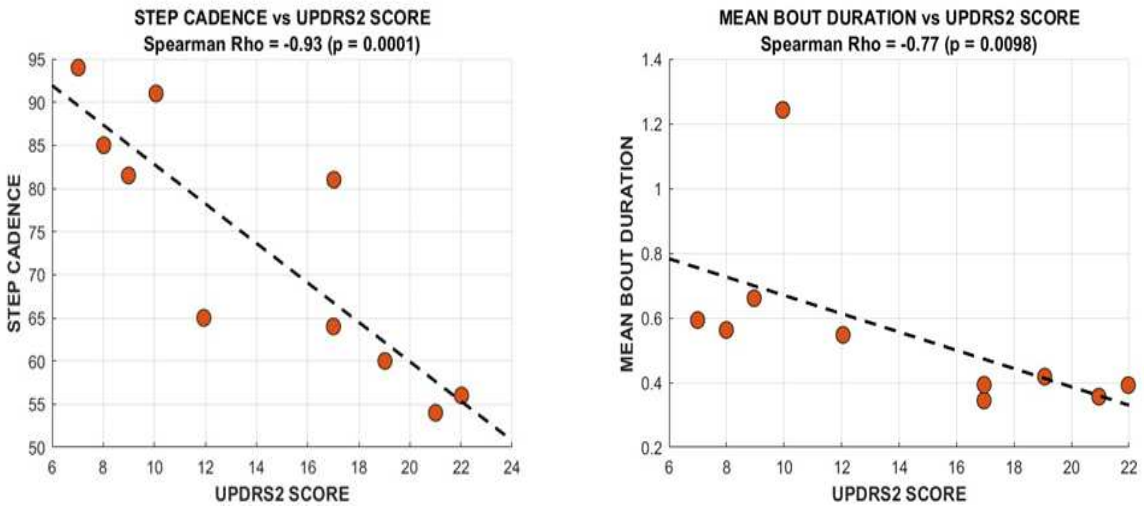


Figure 8.2: Visualization of the correlation between step cadence and mean bout duration with the MDS-UPDRS Part 2 score.

From a clinical perspective, these results have important implications. First, they suggest

that continuous gait monitoring could serve as an objective digital endpoint for functional disability, complementing patient-reported outcomes. Second, cadence and bout-related metrics may enable early detection of functional decline, even before patients perceive or report substantial difficulties. Finally, the strong linkage between gait and daily functioning supports the use of wearable-based gait analysis as a scalable tool for longitudinal disease monitoring and personalized treatment optimization in real-world settings.

In summary, the analysis demonstrates that free-living gait rhythm metrics encode clinically meaningful information about functional disability in PD. Their strong association with MDS-UPDRS Part II underscores the central role of gait as a unifying marker of motor impairment, autonomy loss, and quality of life, positioning wearable-derived gait features as key candidates for digital biomarkers in both clinical practice and research.

8.1.2 Predictors of motor severity (MDS- UPDRS III ON)

To fully characterize the sensor's discriminative capacity, the correlation analysis was extended to include clinical scores assessed during the "ON" medication state (MDS-UPDRS Part III ON). This analysis aimed also to verify whether the digital biomarkers remain valid even when the patient's motor symptoms are mitigated by dopaminergic therapy. In this context, the digital features capture the residual motor deficits that persist despite optimal pharmacological treatment.

When correlated with the MDS-UPDRS Part III assessed under the effect of dopaminergic medication, STEP COUNT emerged as the strongest digital biomarker ($r = -0.76, p = 0.0159$) (see Figure 8.4). This finding indicates that, for medicated patients, the total volume of daily activity is the best indicator of their residual motor severity [37]. A higher step count is strongly associated with a lower UPDRS III ON score. Physiologically, this suggests that when the specific biomechanical deficits (such as rigidity or shuffle) are partially masked by Levodopa, the patient's global wellness is best captured by their activity level. Patients who respond well to therapy are able to maintain higher activity volumes, whereas patients with a poorer therapeutic response remain sedentary despite medication.

Similarly, gait rhythm and continuity proved to be highly robust predictors of global motor severity. Both STEP CADENCE ($r=-0.73$, $p<0.05$) and MEAN BOUT DURATION ($r=-0.73$, $p<0.05$) showed strong negative correlations. This implies that gait rhythmicity is a structural biomarker. While dopaminergic medication successfully improves the amplitude of movement, the underlying temporal organization of the gait cycle—and the ability to sustain it over long bouts—remains closely tied to the patient's true baseline severity. In essence, the wearable sensor effectively "sees through" the medication to capture the core axial motor deficit.

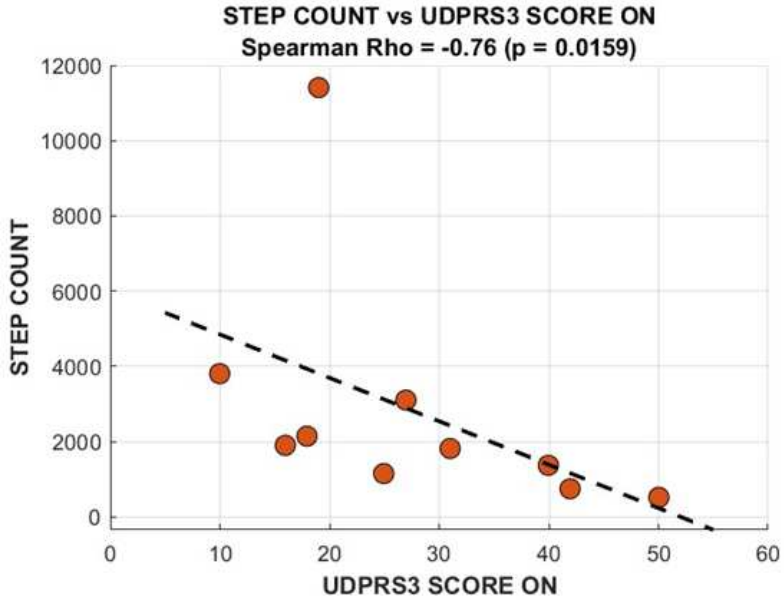


Figure 8.3: Visualization of the correlation between the step count and the MDS-UPDRS Score ON.

8.2 Integrated discussion

The concurrent validity analysis provides strong evidence that the inertial sensor–based monitoring system proposed in this thesis yields clinically meaningful, robust, and interpretable digital biomarkers of PD. The convergence between wearable-derived metrics and established clinical scales demonstrates that continuous motion data collected in real-world conditions can faithfully reflect both functional disability and neurological impairment. This paradigm shift moves the evaluation of PD from episodic to a continuous, objective assessment of the patient's true free-living condition.

The results clearly support the ecological validity of the proposed framework. The near-unitary negative correlation observed between gait cadence and MDS-UPDRS Part II underscores that gait rhythm, measured unobtrusively during daily life, captures the functional consequences of PD. Because UPDRS II predominantly evaluates mobility-dependent activities and the patient's subjective experience of motor deficits, the ability of cadence to mirror perceived disability confirms that free-living gait parameters represent direct surrogates of everyday functioning. This finding highlights the advantage of wearable sensors in overcoming recall bias and contextual limitations inherent to traditional clinical tools, providing a quantifiable metric of the patient's actual independence and mobility at home.

Furthermore, the analysis demonstrates a high degree of phenotypic specificity, proving that the algorithm is capable of disentangling different pathophysiological components of motor impairment. Digital features were shown to map onto specific motor domains with remarkable clinical coherence. On one hand, axial and mobility features (such as STEP CADENCE, STEP COUNT, and MEAN BOUT DURATION MIN) emerged as the most robust and highly signif-

icant predictors ($p<0.05$, $r>0.70$) of both subjective functional disability and objective residual deficits.

On the other hand, upper-limb metrics like hand movement smoothness displayed a moderate effect size ($r=-0.56$) but did not reach statistical significance ($p=0.096$). Rather than a limitation of the tracking algorithm, this lack of significance accurately reflects the pharmacological profile of PD: appendicular bradykinesia is highly levodopa-responsive [38]. The medication effectively normalizes upper limb movement in the ON state, compressing the variance among patients and consequently weakening the linear correlation with global clinical scores. Similarly, tremor-related features did not linearly track the broad composite scales, perfectly mirroring the highly episodic and medication-sensitive nature of this symptom in unconstrained free-living conditions [30].

This statistical divergence confirms that the system is not merely detecting the presence of the disease, but is dynamically profiling its distinct dimensions and their specific responses to dopaminergic therapy. As a result, the proposed approach enables automated and objective patient phenotyping, a key step toward personalized disease monitoring.

These findings demonstrate that the extracted features are not abstract numerical descriptors or statistical artifacts, but faithful digital representations of clinically relevant motor phenomena. The system successfully objectifies both subjective functional impairment (MDS-UPDRS II) and motor signs observed by neurologists (MDS-UPDRS III), bridging the gap between patient experience and clinical evaluation. By mapping raw accelerometer data to standardized medical scales, this framework transforms highly dimensional inertial signals into an interpretable language for clinicians.

Chapter 9

Conclusions

9.1 Key findings

The experimental results obtained in this study offer a comprehensive and rigorous validation of the proposed monitoring framework, providing significant insights into the digital characterization of PD in free-living conditions.

In the first analytical phase, statistical comparisons confirmed that the digital features extracted from inertial data can significantly differentiate PD from HC, delineating a clear digital motor signature of the pathology. Specifically, the temporal constancy of tremor (PERCENTAGE WINDOWS TREMOR) emerged as the strongest discriminative parameter ($Cohen's\ d = 1.32,\ AUC = 0.86$), suggesting that in unsupervised free-living conditions, the persistence of the symptom is a much more reliable pathological marker than its instantaneous intensity [30]. Similarly, MEAN HAND AMPLITUDE accurately quantified the poverty of movement typical of bradykinesia ($Cohen's\ d = 1.31,\ AUC = 0.88$), quantitatively confirming the reduction in spontaneous movement typical of PD [23], while the variability of walking rhythm, represented by CADENCE CV INTER BOUT, proved to be a superior biomarker compared to simple gait speed, highlighting the patients' loss of motor automaticity and subsequent instability ($Cohen's\ d = 1.07,\ AUC = 0.860$) [39]. Although single features demonstrated the ability to separate the two groups, the heterogeneity of PD, characterized by varying combinations of tremor, bradykinesia, and gait instability, means that no single metric can fully describe the pathological state. A comprehensive clinical profile requires the integration of multiple independent domains to capture the full spectrum of the motor phenotype: the disease is best described as a combination of alteration across tremor persistence, movement amplitude, gait variability and global activity patterns. The convergence of large effect sizes, high AUC values, and strong classification performance demonstrates that integrating multiple domains is essential for accurate disease modeling

Moving beyond group-level statistical significance, the application of Machine Learning algorithms demonstrated the system's robustness in making accurate diagnoses at the individual

level. The Random Forest model achieved high diagnostic accuracy (90%), confirming that the selected feature set contains sufficient information to automate disease detection. This result is consistent with previous findings in the literature, where Random Forest has been identified as one of the most robust and high-performing algorithms for the classification of Parkinson's disease based on wearable sensor data [33]. This demonstrates that ML can synthesize information from different motor domains more effectively than traditional statistical methods, compensating for the inter-individual variability observed in the PD cohort. A crucial aspect of these models was the perfect specificity achieved, which allowed the correct identification of all healthy subjects without producing false positives. The 80% sensitivity reflected a conservative behaviour of the algorithms, whose only errors were limited to patients with extremely mild symptomatology. Beyond overall classification performance, a deeper understanding of the model behavior was obtained through feature importance analysis, which provides insight into which digital biomarkers most strongly contribute to distinguishing PD from HC. Across the evaluated models the most influential features consistently belonged to three main domains: tremor consistency, movement amplitude, and gait variability. In particular, PERCENTAGE WINDOWS TREMOR emerged as the most relevant features in the models, MEAN HAND AMPLITUDE showed high importance, indicating that reduce movement amplitude is a robust discriminator, and gait related features as STEP COUNT, STEP CADENCE ranked prominently, supporting that gait irregularity and loss of rhythmic stability are critical indicators of impaired motor control. Features identified as most important by the machine learning models closely overlap with those showing the highest effect sizes and ROC performance. This convergence across analytical methods strengthens the robustness of the findings and indicates that the models are leveraging clinically meaningful biomarkers. At the same time, the distribution of importance across multiple features highlights that no single biomarker dominates the classification process. Instead, the models rely on a distributed representation of the disease, where each feature contributes partially to the decision boundary.

The final phase of the analysis involved the clinical association between the gold standards and our features. The high correlations found provide evidence that the digital biomarkers map directly onto established clinical constructs. Gait-related metrics showed the strongest associations with functional disability. STEP CADENCE emerged as a strong indicator of patient-perceived disability in daily activities, recording a very high inverse correlation with the MDS-UPDRS Part II scale ($r = -0.93,\ p < 0.001$) [35]. This result indicates that a reduction in walking rhythm is directly proportional to the patient's perceived disability. Physiologically, cadence represents the automaticity of stepping; as this automaticity deteriorates, patients report increasing difficulty in routine tasks such as dressing or walking independently, leading to higher UPDRS II scores [40].

The MDS-UPDRS Part III is based on the neurologist's direct observation. Since the free-living data were acquired without medication logs, the analysis focused on validation against the ON state, representing the patient's optimal therapeutic window. STEP COUNT ($r = -0.76$)

showed highly significant correlations with the objective motor score [38]. This suggests that even in the medicated state, axial deficits (gait and stability) remain detectable by the sensors and accurately reflect the global disease severity assessed in the clinic. The high correlation coefficients, found between the features and the clinical scores, demonstrate that the extracted digital biomarkers successfully associate both with the functional disability reported by the patient and the objective assessment conducted by the neurologist. Interestingly, tremor features did not show strong correlations with clinical scales, reflecting their episodic and medication sensitive nature.

9.2 Study limitations and future directions

While the results presented in this thesis demonstrate the robustness and clinical validity of the proposed monitoring framework, several limitations inherent to the study design and dataset must be acknowledged to contextualize the findings correctly.

The primary constraint is the small cohort size (N=20), which, although sufficient for preliminary statistical significance, limits the generalizability of the Machine Learning models. The risk of overfitting in high-dimensional tasks remains a concern; therefore, future work should prioritize the analysis of significantly larger cohorts. Increasing the sample size will not only improve the stability of decision boundaries but also enable the deployment of more sophisticated architectures, such as Deep Learning, to capture complex, multi-dimensional dependencies that a small sample cannot fully represent.

The reliance on a single wrist-worn sensor facilitates patient compliance but introduces a biomechanical blind spot regarding lower limb mechanics and axial symptoms. To reconstruct a more comprehensive digital phenotype, future studies should transition toward multi-sensor configurations. Integrating distributed inertial sensors or non-inertial signals would extend monitoring beyond the motor domain, capturing complex events like freezing of gait or postural instability that a single wrist device may only approximate.

The absence of a precise log for medication intake means that the extracted features represent an average of fluctuating motor states. Currently, this prevents a minute by minute correlation between digital biomarkers and the Levodopa pharmacokinetic cycle. A crucial avenue for future research lies in the automated detection of ON/OFF motor fluctuations. By mapping daily fluctuations against medication timing, future models could provide physicians with an objective, data-driven diary of drug response, enabling personalized medication titration and optimized administration timings.

Finally, while this study provides a ”snapshot” of the patients’ status, the real power of digital biomarkers lies in their longitudinal application. Traditional clinical scales like the MDS-UPDRS lack the resolution to detect subtle changes over short periods. Future research should focus on linking temporal changes in wearable-derived features to long-term clinical assess-

ments. Continuous monitoring of metrics such as STEP CADENCE or HAND MOVEMENT SMOOTHNESS could offer a granular measure of the rate of motor decline, establishing these biomarkers as sensitive surrogate endpoints for clinical trials and long-term disease management.

9.3 Final remarks

In conclusion, this thesis work demonstrates that the use of a single wrist-worn inertial sensor, integrated with rigorous signal processing and Machine Learning techniques, represents a valid, robust, and clinically relevant system for decoding the complexity of motor impairment in PD. By bridging the gap between raw measurement, automated classification, and medical interpretation, the developed framework has proven to be a highly reliable tool for the objective quantification of symptoms in free-living conditions.

Overcoming the intrinsic limitations of traditional clinical assessments, which are typically episodic and subjective, this study lays a solid and concrete foundation for the future integration of wearable technologies into neurological care. Furthermore, the reliance on a single, non-invasive wearable device ensures high patient compliance over extended periods. By minimizing the physical and psychological burden often associated with complex multi-sensor setups, this approach allows high-quality, continuous neurological monitoring to seamlessly integrate into the patient's daily routine, capturing their true motor behavior without causing discomfort or altering their natural habits. Continuous and interpretable digital biomarkers represent a key enabler for personalized medicine.

Bibliography

[1] B. R. Bloem, M. S. Okun, and C. Klein. Parkinson's disease. *The Lancet*, 397(10291):2284–2303, 2021.

[2] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, A. Schrag, and A. E. Lang. Parkinson disease. *Nature Reviews Disease Primers*, 3(1):17013, 2017.

[3] S. Del Din, A. Godfrey, C. Mazzà, S. Lord, and L. Rochester. Free-living monitoring of Parkinson's disease: lessons from the field. *Movement Disorders*, 31(9):1293–1313, 2016.

[4] L. V. Kalia and A. E. Lang. Parkinson's disease. *The Lancet*, 386(9996):896–912, 2015.

[5] E. Rovini, C. Maremmani, and F. Cavallo. How wearable sensors can support Parkinson's disease diagnosis and treatment: a systematic review. *Frontiers in Neuroscience*, 11:555, 2017.

[6] A. J. Espay, P. Bonato, F. B. Nahab, W. Maetzler, J. M. Dean, J. Klucken, et al. Technology in Parkinson's disease: challenges and opportunities. *Movement Disorders*, 31(9):1272–1282, 2016.

[7] F. Porciuncula, A. V. Roto, D. Kumar, I. Davis, S. Roy, C. J. Walsh, and L. N. Awad. Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances. *PM&R*, 10(9):S220–S232, 2018.

[8] L. J. W. Evers, Y. P. Raykov, J. H. Krijthe, A. L. Silva de Lima, R. Badawy, K. Claes, et al. Real-life gait performance as a digital biomarker for motor fluctuations: the Parkinson@home validation study. *Journal of Medical Internet Research*, 22(10):e19068, 2020.

[9] Y. M. Sun, Z. Y. Wang, Y. Y. Liang, C. W. Hao, and C. H. Shi. Digital biomarkers for precision diagnosis and monitoring in Parkinson's disease. *npj Digital Medicine*, 7(1):218, 2024.

[10] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al. The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology*, 95(4):629–635, 2011.

[11] The Michael J. Fox Foundation for Parkinson's Research. Parkinson's progression markers initiative (ppmi) official website, 2026. Accessed: April 8, 2026.

[12] K. Marek, S. Chowdhury, A. Siderowf, S. Lasch, C. S. Coffey, C. Caspell-Garcia, T. Simuni, D. Jennings, C. M. Tanner, J. Q. Trojanowski, et al. The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort. *Annals of Clinical and Translational Neurology*, 5(12):1460–1477, 2018.

[13] M. Burq, E. Rainaldi, K. C. Ho, et al. Virtual exam for Parkinson's disease enables frequent and reliable remote measurements of motor function. *npj Digital Medicine*, 5(1):65, 2022.

[14] Z. S. Nasreddine, N. A. Phillips, V. Bèdirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[15] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170, 2008.

[16] M. M. Hoehn and M. D. Yahr. Parkinsonism: onset, progression and mortality. *Neurology*, 17(5):427–442, 1967.

[17] S. R. Small, S. Chan, R. Walmsley, et al. Self-supervised machine learning to characterize step counts from wrist-worn accelerometers in the UK Biobank. *Medicine & Science in Sports & Exercise*, 2024.

[18] L. Ma, T. M. Mi, Q. Jia, C. Han, J. K. Chhetri, and P. Chan. Gait variability is sensitive to detect Parkinson's disease patients at high fall risk. *International Journal of Neuroscience*, 132(9):888–893, 2022.

[19] S. R. Small, L. von Fritsch, A. Doherty, S. Khalid, and A. Price. OxWalk: Wrist and hip-based activity tracker dataset for free-living step detection and gait recognition, 2022.

[20] Vieyra Software. Physics Toolbox Sensor Suite [Mobile application]. `https://www.vieyrasoftware.net`, 2026. Accessed: 2026-01-20.

[21] N. A. Timmermans, R. Terranova, D. C. Soriano, et al. A generalizable and open-source algorithm for real-life monitoring of tremor in Parkinson's disease. *npj Parkinson's Disease*, 11:205, 2025.

[22] E. Post, T. van Laarhoven, Y. P. Raykov, et al. Quantifying arm swing in Parkinson's disease: a method accounting for arm activities during free-living gait. *Journal of Neuro-Engineering and Rehabilitation*, 22:37, 2025.

[23] N. Mahadevan, C. Demanuele, H. Zhang, D. Volfson, B. Ho, M. K. Erb, and S. Patel. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *npj Digital Medicine*, 3:5, 2020.

[24] H. Yuan, S. Chan, A. P. Creagh, et al. Self-supervised learning for human activity recognition using 700000 person-days of wearable data. *npj Digital Medicine*, 7(1):91, 2024.

[25] G. Bailo, F. L. Saibene, V. Bandini, P. Arcuri, A. Salvatore, M. Meloni, et al. Characterization of walking in mild Parkinson's disease: reliability, validity and discriminant ability of the six-minute walk test instrumented with a single inertial sensor. *Sensors*, 24(2):662, 2024.

[26] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[27] S. Lin, C. Gao, H. Li, et al. Wearable sensor-based gait analysis to discriminate early Parkinson's disease from essential tremor. *Journal of Neurology*, 270:2283–2301, 2023.

[28] Nikolaos Kostikis, Dimitrios Hristu-Varsakelis, Marianthi Arnaoutoglou, and Christos Kotsavasiloglou. A smartphone-based tool for Bradykinesia and Tremor assessment in Parkinson's Disease patients. *Pervasive and Mobile Computing*, 22:114–124, 2015.

[29] X. Zhang, W. Fan, H. Yu, L. Li, Z. Chen, and Q. Guan. Single- and dual-task gait performance and their diagnostic value in early-stage Parkinson's disease. *Frontiers in Neurology*, 13:974985, 2022.

[30] J. V. Pulliam, M. A. Burack, K. E. Aquino, D. A. Heldman, S. H. Mehta, et al. Continuous in-home monitoring of essential tremor and Parkinson's disease motor function using a wearable sensor. *Sensors*, 18(2):607, 2018.

[31] E. Rastegari, H. Ali, and V. Marmelat. Detection of Parkinson's disease using wrist accelerometer data and passive monitoring. *Sensors*, 22(23):9122, 2022.

[32] J. Varghese, A. Brenner, M. Fujarski, C. M. van Alen, L. Plagwitz, and T. Warnecke. Machine learning in the Parkinson's disease smartwatch (PADS) dataset. *npj Parkinson's Disease*, 10(1):9, 2024.

[33] D. Trabassi, M. Serrao, T. Varrecchia, A. Ranavolo, G. Coppola, R. De Icco, C. Tassorelli, and S. F. Castiglia. Machine learning approach to support the detection of Parkinson's disease in IMU-based gait analysis. *Sensors*, 22(10):3700, 2022.

[34] M. Mancini and F. B. Horak. Potential of gait analysis with wearable sensors to assess parkinsonians: a combined motor-cognitive and educational approach. *NeuroRehabilitation*, 26(3):191–205, 2010.

[35] T. Ellis, J. T. Cavanaugh, G. M. Earhart, M. P. Ford, K. B. Foreman, et al. Which measures of physical activity agree with stair climbing and walking difficulty in Parkinson's disease? *Journal of Neurologic Physical Therapy*, 39(4):209–215, 2015.

[36] S. Lord, K. Baker, A. Nieuwboer, D. Burn, and L. Rochester. Gait variability in Parkinson's disease: an indicator of non-dopaminergic contributors to gait dysfunction? *Journal of Neurology*, 258(4):566–572, 2011.

[37] M. Mancini, M. El-Gohary, S. Pearson, J. McNames, H. Schlueter, J. G. Nutt, L. A. King, and F. B. Horak. Continuous monitoring of turning in Parkinson's disease: rehabilitation potential. *NeuroRehabilitation*, 37(1):3–10, 2015.

[38] F. Lipsmeier, K. I. Taylor, T. Kilchenmann, D. Wolf, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Movement Disorders*, 33(8):1287–1297, 2018.

[39] R. C. Helmich. The cerebral basis of Parkinsonian tremor: a network perspective. *Movement Disorders*, 33(2):219–231, 2018.

[40] R. Iansek, M. Danoudis, F. Huxham, and M. E. Morris. Gait automaticity and gait determinants in Parkinson's disease. *Movement Disorders*, 21(2):176–181, 2006.