Towards Data-Driven Gait Analysis with a Special Focus on Individuals with Multiple Sclerosis

A Thesis submitted to the University of Sheffield

for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Alexandru Stihi, MEng.

Department of Mechanical Engineering

University of Sheffield

March 2025

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors, Dr. Tim Rogers, Professor Elizabeth Cross and Professor Claudia Mazzà, for making this work possible and for their continuous support and guidance throughout these four years.

Secondly, I am grateful for all those who have made the IMSB and Dynamics Research Group in Sheffield such a friendly and interesting place to work. Special thanks must also go to Dr. Pinaki Bhattacharya, Dr. Jennifer Rowson and especially to Dr. Tecla Bonci for allowing me to take part in the Mobilise-D project and for the support offered during the final years of my PhD.

Also, I would like to thank my family for their unwavering love, support and constant encouragement, always motivating me to be my best. For this, I am deeply thankful.

Finally, to Iris, I could not have done this without you. You are my partner in crime and have always steered me in the right direction. This is my proudest achievement and I am so grateful to have you by my side!

Dedicated to my family.

Abstract

The healthcare sector is witnessing a paradigm shift, driven by the transformative potential of digital technologies. This digital revolution fundamentally redefines data collection, interpretation, and utilization, impacting both clinical practice and research endeavours. Within this evolving landscape, wearable sensors have emerged as a promising tool for unobtrusive monitoring of physiological kinematic data. This holds particular promise for the understanding of the progression of neurological conditions, where mobility assessments play a significant role. One such neurological condition where data-driven sensory assessments can be particularly impactful is multiple sclerosis (MS). MS is a slowly progressive heterogeneous neurodegenerative disease which primarily manifests through reduced mobility. Considering the complexity of the disease, traditional clinical evaluations, relying on subjective observation and intermittent testing, may overlook subtle gait changes over time. This highlights the need for more objective and sensitive assessment methods. Recent advancements in machine learning technology—which have been specifically developed to extract meaningful information from high-dimensional data and model complex non-linear biomechanical signals—appear well-suited to augment clinical assessment with data-driven insights. As such, this thesis proposes a data-driven assessment pipeline, which includes detection of gait impairment, severity assessment, as well as methodologies for longitudinal monitoring.

Within the context of gait impairment detection, the first contribution of this thesis consists in the proposal of a novel gait anomaly detection technique, using the Mahalanobis squared distance, along with the minimum covariance determinant. This approach offers robust estimates of the true healthy participant condition and improves sensitivity of gait impairment detection for the MS population.

The second contribution of this thesis is the proposal of a novel neural network-based framework for disease severity assessment using a single wearable sensor worn on the lower back. The thesis introduces contrastive learning approaches, which aim to effectively cluster individuals with similar gait patterns, improving model generalisation. Additionally, the thesis employs layer-wise relevance propagation to identify key gait features associated with severity assessment, aiding interpretability and building trust into the model predictions.

The third contribution addresses a critical gap in the current practice: the lack of reliable methods for monitoring disease progression over time. Therefore, the thesis introduces a novel technique for assessing the consistency of movement patterns between clinical visits. Using residual patterns as a sensitive feature, computed as the difference between the predictions of autoregressive with eXogenous inputs models and the true measured data, together with kernel two-sample hypothesis testing using the maximum mean discrepancy, this thesis provides some fundamental considerations and identifies the challenges for longitudinal data analysis, highlighting the need for more generalisable models.

The final developments in this thesis approach gait analysis from a different perspective by introducing a novel Bayesian framework for probabilistically modelling the shank angular velocity as a proxy for lower limb distal motion. Given the heterogeneous MS gait pattern, a probabilistic approach can offer valuable insights in the challenging problem of assessing and quantifying the degree of gait impairment and its changes over time, especially in the context of neurological disorders, such as MS, which is marked by intrinsically unpredictable disease progression. As such, the last contribution presented here is the extension of hierarchical Gaussian process models to effectively handle heteroscedasticity and facilitate scalability to large datasets through sparse inference. By acknowledging the hierarchical nature of wearable sensor data—collected from contralateral limbs, individuals, and groups of individuals comprising a population- this modelling approach allows a granular analysis of the gait patterns. The idea is to make a departure from understanding gait with respect to a set of summary features. Instead, the shank angular velocity is modelled functionally, across the entire gait cycle, with automatic uncertainty estimation.

The methodological development of the algorithms presented in this thesis leaves the user with a toolbox of methods which can facilitate not only a better understanding of the gait patterns exhibited by people with MS, but can be also extrapolated to other pathological conditions affecting gait.

Publications

Author Publications to Date

Journal Papers - in Print

A. Stihi, T.J. Rogers, C. Mazzà, E.J. Cross, "On gait consistency quantification through ARX residual modeling and kernel two-sample testing", *IEEE Transactions on Biomedical Engineering*, 71(3):720–731, 2024, doi: 10.1109/TBME.2023.3316474

Journal Papers - Under Review

A. Stihi, C. Mazzà, E.J. Cross, T.J. Rogers, Tentative Title: "Hierarchical Gaussian processes for characterising gait variability in multiple sclerosis", *Data-Centric Engineering*

Conference Abstracts

A. Stihi, T.J. Rogers, C. Mazzà, E.J. Cross, "Identifying gait anomalies for people affected by multiple sclerosis using autoregressive residual analysis and the maximum mean discrepancy", *BioMedEng 2022*, University College London, United Kingdom.

A. Stihi, T.J. Rogers, C. Mazzà, E.J. Cross, "Detection of anomalies in the gait of people affected by multiple sclerosis", *INSIGNEO Showcase 2022*, University of Sheffield, United Kingdom.

A. Stihi, C. Mazzà, E.J. Cross, T.J. Rogers, "Revealing Gait Anomalies Through Hierarchical Bayesian Machine Learning", *BioMedEng2023*, Swansea University, United Kingdom

P. Tasca, A. Küderle, C. Kirk, C. Hinchcliffe, D. Megaritis, **A. Stihi**, B. Caulfield, L. Rochester, A. Cereatti, "From development to deployment: introducing MobGap, the open-source tool for mobility assessment with wearable devices by Mobilise-D", *SIAMOC Congress 2024*, Stresa, Italy.

Contents

Chapter 1

Introduction

The healthcare sector is undergoing a significant paradigm shift, driven by the transformative power of digital technologies. This digital transformation fundamentally alters how data is collected, interpreted and utilised, impacting both clinical practice and research endeavors. Within this evolving landscape, wearable sensors have emerged as a promising tool for unobtrusive monitoring of physiological kinematic data. This holds particular promise for the understanding of the progression of neurological conditions, where mobility assessments play a significant role [1].

One such neurological condition where this approach can be particularly impactful is multiple sclerosis (MS). MS is a slowly progressive neurogenerative disease which is characterised by the inflammatory-mediated demyelination of axons in the central nervous system [2, 3]. The disease affects over 2.3 million people globally [4] and is currently one of the leading causes of disability in young and middle-aged adults [5]. While clinical signs of MS are highly diverse, mobility constraints in the form of gait impairments are one of the most common symptoms experienced by patients with multiple sclerosis (PwMS) [6, 7]. As a result of gait disorders, PwMS experience a loss of functional mobility, along with a decline in physical independence and overall quality of life [8]. This has a considerable escalating impact on a personal, societal and economic level. [9].

MS-related symptoms exhibit significant inter- and intra- patient variability and change at different rates over different timescales. At times, a sudden deterioration in disability level is observed, which is caused by the rapid demyelination, for reasons

which are not completely understood to this day [10]. Alternatively, there can also be a plateau in the motor abilities of PwMS, or slight improvements, as a result of reduction of inflammation [11]. This heterogeneity poses a substantial challenge in quantifying the efficacy of therapeutic interventions and disease management strategies. Although MS demonstrates a highly individualised disease course, existing literature recognises four main clinical phenotypes based on disease progression (see Figure 1.1) [12–15]:

- Relapsing-remitting MS (RRMS): This phenotype can be initially experienced by approximately 85% of the MS-affected population. It is characterised by acute episodes of neurological dysfunction, termed *relapses*, followed by periods of partial or complete recovery, referred to as *remissions*.
- Primary-progressive MS (PPMS): Approximately 10% of individuals diagnosed with MS experience a progressive worsening of symptoms from disease onset, without clear relapses or remissions.
- Progressive-relapsing MS (PRMS): This is the least frequent phenotype, affecting less than 5% of PwMS. It is characterised by a progressive disease course with superimposed acute relapses, but without distinct remission periods.
- Secondary-progressive MS (SPMS): This phenotype manifests only in some individuals initially diagnosed with RRMS. Throughout the course of the disease, the initially relapsing-remitting phase transitions to a progressive worsening of symptoms, with or without occasional relapses. Roughly half of RRMS patients are estimated to progress to SPMS.

While gait impairments have conventionally characterised the more advanced stages of the disease, previous studies revealed that subtle gait alterations may also be apparent in mildly affected MS patients [17, 18]. This suggests that the deterioration of the motor function represents a *prodromal* phase of the disease, prior to the appearance of clinical symptoms [19], requiring further attention. As a result, measuring and quantifying gait anomalies could provide a potential biomarker for MS progression. This has the potential of supporting clinical decision-making strategies, while offering early opportunities for therapeutic interventions. To this end, periodic and accurate assessment outcomes are crucial for monitoring of MS progression and verifying the effectiveness of rehabilitative and pharmacologic treatment plans [20].

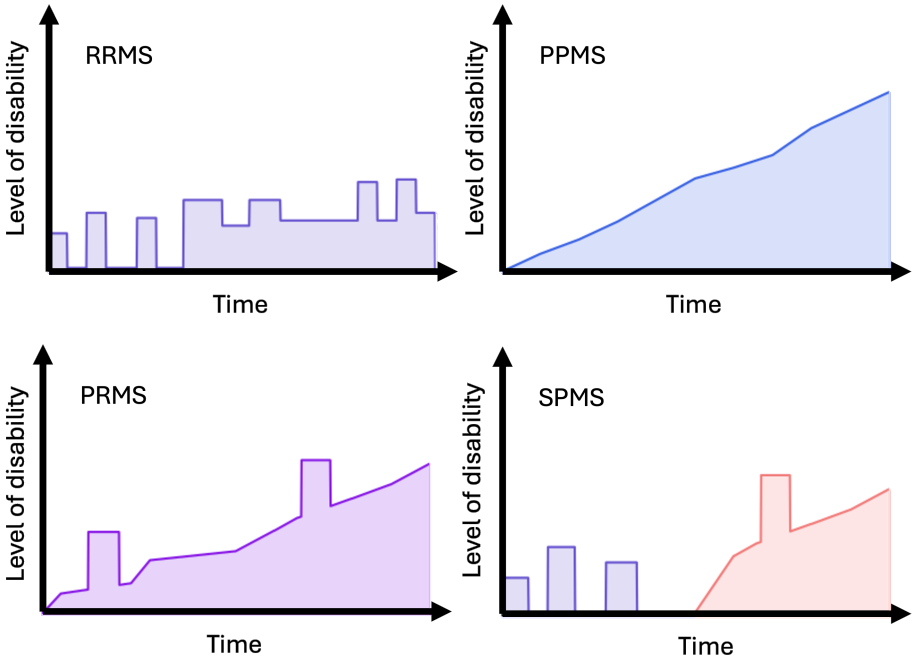


Figure 1.1: Typical disease course for the four main multiple sclerosis (MS) phenotypes. RRMS - Relapsing-remitting MS; PPMS - Primary-progressive MS; SPMS - Secondary-progressive MS; PRMS - Progressive-relapsing MS. Figure adapted from [16].

Traditionally, characterisation of gait impairments is performed with the aid of clinical evaluation scales such as the Expanded Disability Status Scale (EDSS) [21], which measures the severity of the disease following a neurological assessment and observing the walking range as well as the level of walking assistance needed. The scale is rated from 0 (normal healthy status) to 10 (MS-related death) in 0.5-unit increments. Scores up to 3.5 indicate no visible gait impairment, while scores between 4.0 and 5.5 denote individuals capable of walking limited distances independently. Scores up to 6.5 indicate the necessity of assistive walking devices, while higher scores are denoting restricted mobility. Although, clinical evaluation scales provide an intial indication of the degree of gait deficit, they are limited by subjectivity and clinical expertise, and are incapable of detecting subtle alterations in the gait of PwMS [22]. Their concerning accuracy and precision could have a negative impact on diagnosis and treatment strategies [23]. Moreover, it has been shown that clinical outcomes may also be influenced by the patient's performance and fatigue at the time of the assessment [24, 25]. It becomes clear that clinical assessment scales are not sensitive enough for monitoring gait disability which accumulates not just in the short term, but also during the course of the disease. Thus, alternative, objective approaches are

needed.

The development of new technologies and devices has enabled a more objective approach to gait analysis through instrumentation. This instrumented approach offers the advantage of capturing detailed, quantitative gait data with high accuracy and precision. By analyzing this data, end users can gain valuable insights that can objectively inform clinical decisions. The instrumented gait analysis technologies currently available can be classified into two distinct categories: those based on non-wearable devices and those based on wearable devices [23]. The next few paragraphs will focus on briefly explaining the differences between these technologies, providing advantages and disadvantages, with the aim of justifying the selection of wearable devices as the preferred technology for MS gait assessments used throughout this thesis.

Within this context, non-wearable devices can be further decomposed into two different subtypes: optical motion capture systems (OMCS) and those based on floor sensors (FS). The OMCS-based approach captures gait data through one or more optical sensors, utilising advanced digital image processing, and objectively outputting a comprehensive set of gait parameters. This approach is considered to be the gold standard in laboratory-based gait analysis [26], due to its ability to precisely determine the orientation and position of an object in the global reference frame [27]. However, the OMCS-based approach comes with the major disadvantages of being a high-cost technology, requiring a dedicated laboratory setting and skilled personnel to acquire the data. The FS-based approach, on the other hand, consists of an instrumented floor platform, which is located along the pathway on which subjects are asked to walk as part of their clinical assessment. When a patient walks on the platforms, data is collected with the aid of pressure or ground reaction force sensors (also known as GRF sensors) [28]. Similar disadvantages as the ones previously mentioned for OMCS-based approaches are present for the FS-based approaches. Nonetheless, their major limitation lies in the inability to assess and monitor gait outside laboratory settings, in free living conditions. As a drawback, extrapolating the results and conclusions based on a short period study might not reveal the true condition of PwMS [23]. Therefore, a different approach is preferred.

In the recent years, there has been an increase in the interest in wearable devices [29], especially because they enable continuous and remote gait monitoring outside of the laboratory or clinical context [30]. Because of their practicality and cost-effectiveness, inertial measurement units (IMUs) have become the go-to method for

collecting mobility data [29]. IMUs typically integrate accelerometers, gyroscopes, or magnetometers to record kinematic gait data and therefore capture various aspects of gait mechanics. Specific to MS, IMUs have been successfully used in the past in order to analyse various locomotor tasks performed by PwMS, such as walking in a straight line [19, 31–35], walking on a treadmill [36], the Timed Up and Go (TUG) test [37, 38], walking over obstacles [39], walking while texting concomitantly [40] or ascending stairs [41]. Although additional validation procedures are necessary, recent studies have also further demonstrated the potential of wearable technologies for enabling successful differentiation of abnormal MS gait patterns from healthy ones, and even some potential for distinguishing across different disability levels [17, 18, 42–44].

A significant limitation in current gait analysis research is the widespread reliance on cross-sectional data - where measurements are taken from different participants at a single point in time. While such data provides valuable insights into differences between individuals, it cannot capture how the disease evolves within the same person over time. This contrasts with longitudinal assessments, where the same participants are tracked through repeated measurements over extended periods, enabling researchers to monitor individual disease trajectories [1, 45].

This gap in long-term monitoring is particularly problematic for MS for two key reasons: first, the disease is characterised by its highly variable and unpredictable course [16, 46], making it difficult to understand progression patterns from single timepoint measurements. Second, current analysis procedures lack sufficient generalisability to account for this variability [47, 48]. While recent studies are beginning to incorporate longitudinal research designs [16, 45, 48–50], there remains a pressing need for robust analysis pipelines that can effectively capture and interpret MS-related gait fluctuations over time. To address this gap, this thesis proposes the development of a comprehensive end-to-end pipeline for MS assessment, using a data-driven approach.

1.1 How can we explore the condition of a subject?

It is the author's opinion that gait assessment for pathological populations such as MS is inherently hierarchical. At its lowest level, an assessment framework should entail detection of gait anomalies. Then, following detection, the assessment could

include quantification of disease severity and monitoring of its progression, which would then allow end-users to devise optimal personalized treatment plans, or verify the effectiveness of rehabilitative and pharmacologic interventions. This reasoning is similar to the aims of Structural Health Monitoring (SHM), a field which also makes use of sensory data and is concerned with automated monitoring in order to assess the condition of a structure [51]. The ideal behind SHM is that once structural damages have been identified, the faults can be located, and the damage severity can be quantified. At its most sophisticated level, diagnosis in SHM can even estimate the time to failure of a structure. Subsequently, this allows one to make informed decisions regarding the usage of the monitored structure. Therefore, by drawing an analogy to the field of SHM, this work aims to explore the condition of a patient using a four-level hierarchical system[1]. As such, to the best of author's knowledge, this thesis is the first work to consider a similar hierarchical gait analysis path, analogous to that taken in SHM. The levels of the hierarchical gait assessment framework proposed in this thesis are summarised as follows:

1. **Detection** - Does someone have a gait impairment or not?
2. **Classification** - What type of gait impairment is affecting a particular individual?
3. **Quantification** - How severe is the disease / gait impairment, from a functional gait assessment perspective?
4. **Prognosis** - What inference can we make about disease progression?

It is important to note that the hierarchy presented here provides the core objectives of gait assessment in MS. Nonetheless, to achieve these objectives, efficient methods for capturing and extracting the relevant information from the gait data are required. Fortunately, the recent proliferation of wearable sensor technology has led to the accumulation of vast and information-rich datasets, encoding important information relating the health status of an individual[1, 53]. However, analyzing and interpreting the intricate, time-dependent gait patterns presents a significant clinical challenge. Considering both the availability of data and the complexity of the disease, recent advancements in machine learning technology - which have been specifically developed

[1]Although the equivalent hierarchical assessment used throughout SHM contains a damage localisation level [52], this level is not directly transferable to gait analysis applications. However, this work will highlight specific regions within the gait cycle where gait abnormalities are observed.

to extract meaningful information from high-dimensional data and model complex non-linear biomechanical signals - appear well-suited to address the aims listed in the proposed hierarchical framework [54]. Within this work, machine learning technology would be intended as a flexible tool for extracting useful information using intelligent and flexible methods with the aim of knowledge discovery, i.e., of revealing new information which might have been omitted otherwise. The goal is to leverage the latest developments in machine learning technology and provide interpretable results, augmenting standard gait analysis practice with data-driven insights.

A good place to start would be at the lowest level of the hierarchy outlined above and identify the gait features that can give good indication of the presence of gait impairments, as early detection of gait anomalies is crucial for MS, especially for patients who often have no apparent gait impairment. If successful, this allows clinicians to devise optimal treatment strategies at an early stage, being the starting point of the methods presented in this thesis.

Following the successful detection of gait impairments, another classification problem emerges. This level of the hierarchy seeks to determine and homogenise standard atypical patterns affecting an individual. However, this task presents a significant challenge [55], often requiring additional functional assessments. These are not routinely included in standard clinical evaluations, resulting in a shortage of clinician-annotated data for different types of gait impairment. Despite not directly addressing this level within the current work, insights into the specific gait deficits may still be provided by interpreting the outcomes obtained at the subsequent levels of the hierarchy.

Next, the quantification level is included in the proposed framework as an extension of the detection level. While detection of gait impairments is effectively a binary classification task, the quantification of disease severity is further extended to a multiclass classification problem. Specifically, considering the heterogeneity of the disease, disability levels within this work are classified according to EDSS ranges into four primary categories: individuals with no discernible symptoms (healthy), individuals with mild impairment, individuals with moderate impairment, and individuals with severe impairment. Ultimately, this level could offer valuable insights into the overall course of MS. Here, this severity quantification task constitutes the second exploratory objective addressed in this thesis.

The ultimate aim of this hierarchical approach for assessing the condition of PwMS,

from a gait analysis perspective, would consist of making inference about disease progression. This involves both longitudinal monitoring (i.e., tracking changes in gait characteristics over time), as well as longitudinal prognosis, utilising longitudinal data (and possibly some additional information) to predict the likelihood of future gait deterioration or improvement in response to clinical treatment plans. Despite ongoing research in longitudinal gait monitoring, [16, 45, 49, 50, 56], studies specifically targeting longitudinal gait prognosis remain a rarity.

Within this context, before attempting any longitudinal monitoring or prognosis tasks, this thesis will firstly consider the impact of the natural fluctuations in testing conditions during follow-up assessments on the quantification and prediction of gait changes over time. This aspect is rarely discussed in the relevant literature. Some factors contributing to these fluctuations may include marginal variations in sensor attachment locations on body segments, timing of assessments in the presence of medications, among others. As such, these variations may mask subtle gait improvements or declines. Consequently, accurately quantifying longitudinal gait changes while eliminating the influence of confounding factors introduced by these follow-up testing inconsistencies remains a substantial challenge. Because of the variability induced by these confounding factors, in addition to the challenges posed by the heterogeneity of the disease, there is a need for an objective gait consistency measure. This objective measure should serve as a proxy for good motor control and balance during gait assessments, with its absence potentially indicating neurological or musculoskeletal impairments. In addition, it should facilitate an objective assessment of the generalizability of models employed for the prognosis tasks. Only when a particular modelling procedure exhibits sufficient generalizability to account for both the heterogeneity of the disease and the influence of the confounding factors present at follow-up assessments, can it be employed for longitudinal prognosis. Therefore, to address the challenges of longitudinal prognosis, this thesis presents several novel approaches. These approaches focus on, firstly, quantifying the consistency of walking patterns observed during follow-up assessments. Secondly if the consistency measures demonstrate sufficient robustness, the approaches will be leveraged to provide novel insights into longitudinal gait changes in MS.

1.2 Contribution of this thesis

The work enclosed in this thesis aims to introduce, develop and implement some potentially powerful technologies which, having been originally developed in statistics and machine learning communities, can now bring significant added value to the gait analysis community. Before exploring these individual methods, however, it is necessary to present a short introduction to gait analysis from a machine-learning perspective. As such, Chapter 2 introduces the relevant literature.

Chapter 3 targets the first level in the hierarchical framework for assessing the condition of an MS-affected individual. Therefore, this chapter introduces a robust methodology for detection of gait anomalies, robust against outliers in the training data, while maintaining sufficient sensitivity and generality to detect even subtle impairments.

Chapter 4 presents a novel methodology for predicting disease severity using a single wearable sensor and leveraging contrastive learning approaches, to effectively cluster individuals with similar gait patterns. Moreover, this chapter provides additional insights into the features most relevant for the severity prediction, aiding interpretability and building trust into the model predictions.

Chapter 5 presents some fundamental considerations and challenges for longitudinal data analysis. Here, a novel methodology for verifying the consistency of the gait patterns within a timescale of constant disease status is presented to the reader, highlighting the need for more generalizable models.

Chapter 6 approaches gait analysis from a different perspective, introducing a novel Bayesian paradigm on which the author builds probabilistic models of the shank angular velocity for the purposes of longitudinal monitoring and prognosis. In this chapter, a granular analysis of the lower limb distal motion is presented. The chapter further explores comparisons across different scales (i.e., between groups and individual subjects) while incorporating automatic uncertainty quantification.

Chapter 7 firstly validates the modelling procedure proposed in the previous chapter using a period of constant disease status. Then, the chapter extends the novel modelling approach for longitudinal gait monitoring and prognosis, further highlighting the inherent challenges.

Finally, chapter 8 concludes by summarising the key findings of this work and proposes promising directions for future research, based on the presented results.

1.3 Datasets and ethics approval statement

The research presented in this thesis strictly adhered to ethical guidelines, having secured approval from the National Research Ethics Service (NRES) Committee Yorkshire & The Humber-Bradford Leeds (Reference: 15/YH/0300) and the North of Scotland Research Ethics Committee (Reference: 17/NS/0020). It is also important to note that all participants whose data was utilised in this thesis provided written informed consent prior to entering the studies. From this point onwards, all datasets presented in this thesis are underpinned by this ethical approval.

This thesis utilizes data from four distinct datasets, each contributing valuable insights into gait characteristics of PwMS. The first dataset comprises data collected during routine clinical appointments from 32 PwMS participating in an observational study (STH18829; IRAS-183915). Each participant underwent two gait assessments, performed one week apart. Additionally, a comparison group of 24 healthy controls (HCs) was included in this dataset. The second dataset features longitudinal data collected from 31 PwMS who participated in a double-blinded clinical trial of an investigational medicinal product (CTIMP) (STH17249; IRAS-115286). While all 31 PwMS attended the baseline assessment, only 22 participants completed all four subsequent longitudinal assessments conducted at week 24, 48, and 96. The third dataset originates from a double-blinded, intervention-based study not involving a CTIMP (STH19739; IRAS-224422) and includes data from 50 PwMS. Each participant underwent a single baseline assessment followed by a reassessment after a one-hour interval. To establish a reference point for comparison, an additional set of 14 healthy controls (HCs) was collected by the author. Mimicking the assessment procedure of the PwMS group, these HCs also performed baseline and reassessment evaluations separated by a one-hour period. The inclusion criteria for the HCs was strictly defined to include only those individuals with no prior history of musculoskeletal or neurological disorders. In contrast to the HC group, the patient population for this thesis exclusively comprised individuals diagnosed with MS, whose severity was assessed using the EDSS score. Participants diagnosed with relapsing-remitting MS were eligible for inclusion provided they met two key

criteria: 1) no relapses experienced within 30 days prior to the baseline assessment, and 2) maintained stable treatments for the preceding three months.

The data used throughout this thesis consists of IMU recordings from various segments of the human body. Specifically, all assessments utilised OPAL IMUs (APDM Inc, Portland, OR, USA, sampling frequency, $128Hz$, accelerometer range $\pm 6g$). For the reasons explained in the following chapter, a three-sensor modality was adopted. Hence, one sensor was placed on the lower back (L4-L5 segments), while the other two sensors were placed on the anterior aspect of both lower shanks. The sensors were configured for synchronised recording using the manufacturer's provided access point. The axes of the IMUs were approximately aligned along the anatomical vertical (V), medio-lateral (ML), and anterior-posterior (AP) directions. All participants were instructed to walk at their self-selected pace along a 10 or $14m$ corridor, going back and forth for 6 minutes. Resting was allowed, if needed. Additionally, walking assistive devices were permitted, only if used daily.

Chapter 2

Literature review

This chapter provides an overview of the relevant literature regarding the current state-of-the-art in gait analysis, with a particular focus on MS-affected individuals. Following a brief introduction to the preferred sensory modalities used throughout this thesis and some commonly used features extracted by others, a comprehensive review is conducted that spans the levels of the proposed hierarchical framework for gait assessment targeted herein. Particularly, attention will be given to potential methodologies used for gait anomaly detection and disease severity quantification, culminating with longitudinal monitoring and prognosis.

2.1 Sensor configurations and gait features commonly extracted by others

Gait impairment is a biomarker of MS [57] and may be detected even during the prodromal stage of the disease, prior to the appearance of any clinical symptoms [19]. Quantifying gait anomalies has the potential of supporting clinical decision-making strategies, while offering early opportunities for therapeutic interventions. The need of monitoring MS progression and verifying the effectiveness of rehabilitative and pharmacologic treatment plans has driven the demand for periodic and accurate gait assessments. The evaluation of neurological impairment and disability in MS is traditionally done with the aid of the Expanded Disability Status Scale (EDSS) [21], which is a clinical evaluation scale, based on scores attributed following a

neurological exam, as well as the walking range and the level of walking assistance needed. However, clinical scales are not sensitive enough to quantify the small alterations in the gait of PwMS [22], which is crucial for monitoring gait disability which accumulates over the course of a longitudinal clinical assessment.

During the recent years, alternative gait assessment technologies have been developed. These were presented in the introductory chapter and consist of optical motion capture systems [27], force platforms [28], and wearable devices [29]. The first two technologies are hampered by the high investment cost [58] and the inability to assess and monitor gait outside laboratory settings, i.e., in free living conditions [30]. As a result, the condition of PwMS might not be estimated accurately based on generalizing the results from short-period clinical assessments [23]. Due to recent advancements in wearable technology, inertial measurements units (IMUs) have become the preferred method for quantifying gait characteristics, as a result of their low investment cost, robustness and portability outside laboratory settings [9, 29, 30, 43]. Because of these reasons, IMUs are the preferred sensor modality used throughout this thesis, for gait assessment in MS.

As established in the introductory chapter, the condition of a patient can be explored using a hierarchical system whose lowest level entails detection of gait anomalies. This represents a challenging problem, because the gait patterns of MS subjects with little or no signs of disability may not be significantly different from those of healthy individuals [31]. Therefore, the following paragraphs of this chapter are intended to familiarize the reader with the objective measures commonly extracted by field experts, serving as biomarkers of gait anomalies, focusing especially on the MS population.

Human gait is often described by several objective measures such as spatio-temporal parameters, and the accompanying variability, asymmetry and gait quality metrics [25, 59, 60]. These metrics are briefly explained here. The computation of spatio-temporal parameters necessitates the accurate identification of the gait events for each gait cycle. The gait cycle is defined as the forward propulsion of the human body that includes a sequence of events from the first *initial contact* (IC) with the ground to the successive one. The IC is often referred to as the *heel-strike event*. The gait cycle is divided into two main phases: the stance and the swing phase. The stance phase accounts for approximately 60% of the gait cycle [30] and it begins with the initial contact and ends with the *final contact* (FC) of the same foot. The final contact is often referred to as the *toe-off event*. Approximately 40% of the

gait cycle accounts for the single support phase, which occurs during the stance [30]. The stance phase begins and ends with a double support, collectively accounting for approximately 20% of the gait cycle [30]. Finally, the gait cycle ends with the swing phase, which accounts for approximately 40% of the cycle [30]. This is defined as the time period between the final contact and the successive initial contact of the same leg. A visual representation of the gait cycle can be seen in Figure 2.1.

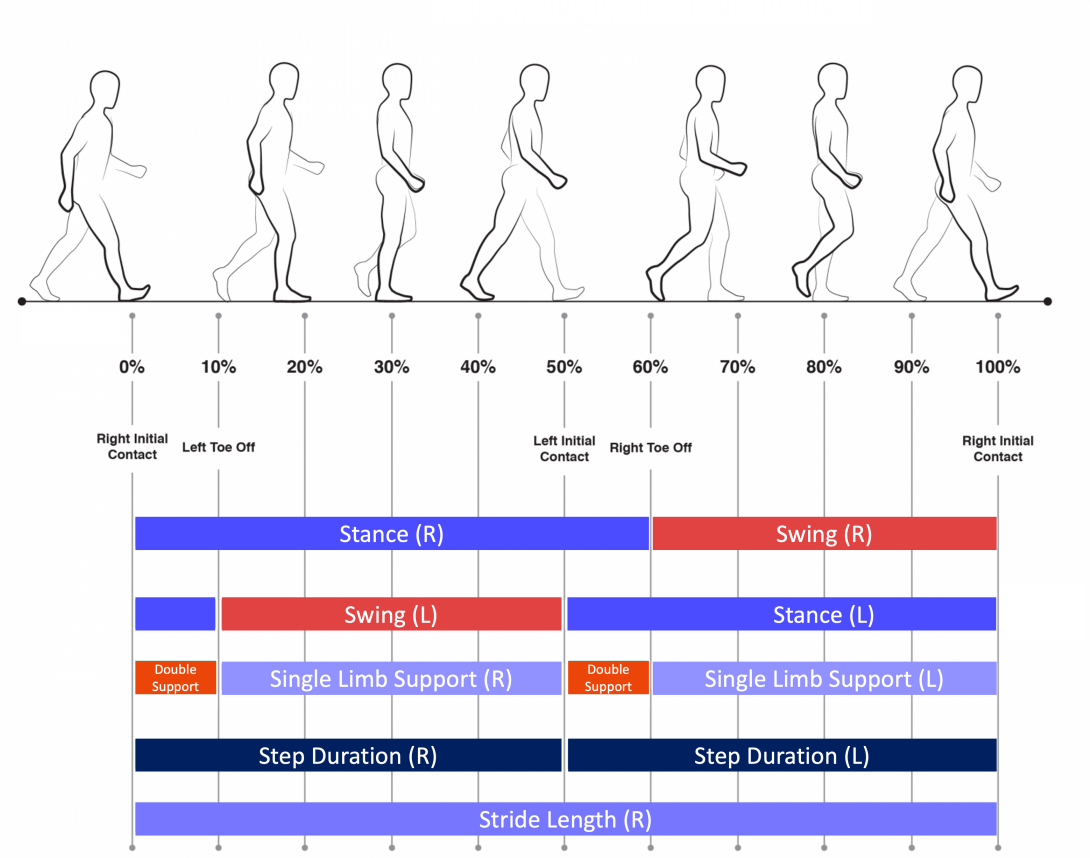


Figure 2.1: Visual representation of the gait cycle. Figure adapted from [61].

Using the gait events as previously described allows for the computation of specific spatio-temporal parameters [62]. The most commonly used temporal gait parameters include the stride, step, stance and swing durations, as well as the single and double support times and cadence, which is defined as the number of steps per unit time. Among spatial parameters, field experts extract the step length and width, as well as the walking speed [63]. Recently included metrics also include the foot pitch angle at heel strikes and toe-off events [64], as well as foot clearance [65]. Variability metrics are often computed as the standard deviation, or the coefficient of variation of the

temporal metrics [17]. Asymmetry metrics are defined as the absolute difference between the mean values of the temporal metrics computed for the right and left lower limbs [66], or the natural logarithm of the absolute ratio between the two mean values [17]. Finally, gait quality metrics (such as the root-mean-squared (RMS) of trunk acceleration [67], the jerk [68], regularity [69], symmetry [22] or the harmonic ratio [70]) are often associated with the ability of individuals to coordinate their motor strategies [17]. Details on the computation of the gait quality metrics mentioned here can be found in the Appendix A section, together with brief explanations of the clinical meaning of the metrics.

Various sensor configurations were found in the literature ([17, 20, 22, 26, 30, 33, 71–79]). Therefore, a vital step towards the accurate detection of gait anomalies requires the identification of specific body segments on which to place the sensors for acquiring reliable signals. Several studies validated the suitability of shank-mounted IMU sensors for the accurate detection of gait events in both healthy and pathological populations [26, 72, 80–83]. Positioning IMU sensors on the shank has a major clinical implication, as the lower limb distal motion can be efficiently captured. In MS, lower limb mobility is often affected as a result of alterations in distal muscle involvement [1, 84, 85]. To this end, Storm et al [80] validated the accuracy of shank-mounted IMU sensors using a pressure insole system. This validation procedure arguably provides a more robust comparison than the motion capture system used as the gold-standard reference in [26]. To test the reliability of IMUs within the context of regular clinical assessments for PwMS, Angelini et al. [22] used two sensors placed on both shanks and one placed on the lower back, overlying the fifth lumbar vertebra (L5). However, considering the optimal placement of sensors for gait analysis, it should be noted that the robustness of signals recorded at the lower back may not match those obtained from other locations, such as the foot or shank [74]. Despite this, lower-back sensor data is still superior to that obtained from wrist-mounted sensors[1] [86]. Panebianco et al. [74] argued that algorithms based on shank or foot-placed sensors offer more accurate and repeatable results, particularly for the precise detection of gait events. This aspect was further reinforced in a recent technical validation study aimed at validating digital mobility outcomes [79]. Moreover, lower-back sensor placement was also found to offer concerning accuracy in terms of asymmetry metrics estimation [30].

[1]Wrists are another common location for placement of wearable sensors, especially due to the emergence of activity monitoring devices [75]. However, this location is not considered appropriate in the case of PwMS, as subjects with EDSS scores higher or equal than 6.0 require assistive walking devices, which may limit their arm movement during walking.

Nonetheless, this location is often considered a clinically advantageous position for a single sensor monitoring approach, due to factors such as its proximity to the body's centre of mass (capturing the overall pattern of human motion [37]), cost-effectiveness of using just one device, ergonomic considerations, but more importantly due to its clinical relevance for assessing fall risk or trunk stability [79], among other gait quality metrics relating to poor balance and coordination [22]. Therefore, while a single sensor placed on the lower back provides a promising balance between accuracy and practicality for remote monitoring, the optimal configuration for accurately and objectively monitoring gait-based biomarkers in MS-affected individuals during clinical longitudinal assessments involves one sensor on the lower back and two sensors on the shanks, in line with the setup employed by Angelini et al. [22]. For clarity, this thesis will adopt specific variations of this optimal sensory configuration, depending on the specific research questions addressed in each chapter.

Typically, objective measures related to a particular motor disease are often grouped together in the form of conceptual gait models [87], which consist of a finite set of features extracted from the gait signals, characteristic of a particular (pathological) population. To deal with the potential high covariance or redundancy of the extracted objective measures, it is common practice to employ dimensionality reduction or feature selection techniques. The relevant techniques will be discussed in 2.3. As a result, the number of features is reduced to a more manageable size, aiding interpretation and providing a disease-specific framework for understanding gait dysfunction, and allowing the end-users to focus on the most relevant features for a particular pathology [87]. For example, conceptual gait models have been developed for older adults [88–90], Parkinson's disease (PD) [87, 91, 92], early-stage neurological gait disorders [93], people with hip fractures [94] and more recently MS [17]. The identification of gait biomarkers grouped in the form of conceptual gait models is crucial for monitoring MS progression and verifying the efficacy of an intervention.

Huisinga et al. [32] studied the variability walking patterns of PwMS, in comparison to healthy controls. In this study, variability metrics reflected the presence of abnormal motor control strategies for PwMS. Additionally, the study also revealed gait anomalies in the speed, stride length, cadence, double support time, swing time, muscle relaxation and contraction times for PwMS. Another study involving PwMS investigated the deterioration of specific aspects of gait during an instrumented 6-minute walking test (6MWT) [95]. The study revealed that cadence, stride time variability, stride and step regularity and the complexity of the gait significantly

altered during the test. A much higher gait variability of PwMS compared to healthy controls was also reported in [34]. Angelini et al. [22] studied gait alterations using IMUs for patients with mild and severe MS, in comparison to healthy controls. The study also incorporated supplementary gait metrics, such as intensity, jerk, regularity, and symmetry, as they can offer additional information regarding the overall gait quality and efficiency [96]. The intensity was calculated as the RMS of the trunk acceleration modulus and is regarded to be an indicative measure of the upper-body dynamic balance, while the jerk was calculated as the first derivative of the trunk acceleration modulus, which can be interpreted as an indicative measure of gait smoothness, reflecting subject's ability to pre-plan motor strategies [22]. The adoption of supplementary metrics seemed to augment the information provided by the spatio-temporal gait parameters [97] and yielded a better indication of the presence of gait anomalies for PwMS. Since these additional metrics were successfully extracted from IMU signals in order to detect gait anomalies in several other pathologies [96–98], it is clear that pre-existing conceptual gait models of a certain pathology can be augmented by additional metrics, which have not been explored before.

Although its integration in conceptual gait models is limited, the short-term Lyapunov exponent (which is an indicator of chaotic dynamics) might be another potential metric which will facilitate detection of gait anomalies, as it offers a good indication of gait stability. This metric has been shown to be able to discriminate between healthy controls and early multiple sclerosis [99] as well as PD [91]. Additionally, the sample entropy was also included in several studies which included PwMS [38, 95]. The latter study identified changes in gait complexity during sustained walking across mild and moderate MS patient groups.

The features discussed above have all demonstrated effectiveness in identifying MS gait impairments, with each feature offering unique insights into different aspects of gait dysfunction. Until recently, however, there was no clear consensus on which combination of gait metrics best characterises MS. This is particularly challenging as many of the metrics introduced previously are not specific to a single pathology. Nevertheless, while spatio-temporal parameters provide fundamental quantitative measures and gait quality metrics offer valuable insights into motor coordination, their individual and collective sufficiency for characterising MS gait patterns remains unclear, particularly in early disease stages. Angelini et al. [17] made important progress by proposing a conceptual gait model for MS using factor analysis to reduce 36 metrics to 20 key features grouped into five domains: rhythm/variability, pace,

symmetry and forward and lateral dynamic balance. However, their initial feature set did not include certain descriptive statistics from acceleration or gyroscopic signals that could potentially highlight clinically relevant gait anomalies in MS. Additionally, while their model demonstrated differences across disability levels, it did not formally implement a classification scheme to distinguish between healthy controls and the varying MS disability levels. This suggests two key research gaps: first, the need to explore whether additional novel features could enhance MS gait characterisation, as the current conceptual model may be incomplete; and second, the requirement for robust classification frameworks that can distinguish between healthy controls and different MS severity levels. These gaps are particularly relevant for early detection in mildly affected patients where gait impairment may not be clinically observable.

2.2 Learning methods

The integration of machine learning, a discipline rooted in statistics and computer science, has emerged as a powerful tool with the potential to significantly revolutionize how data is analysed and interpreted within the gait analysis community [100]. Within this context, before diving into the relevant work relating to either gait anomaly detection, severity quantification, or prognosis, a foundational understanding of the methodologies used to learn from and make inference about the gait data is paramount. These methods rely on a spectrum of algorithms, known as *learning methods*, each tailored to address specific task categories. The choice of the learning method is contingent upon various factors, including the complexity of the task, availability of annotated data, or the size of the dataset. In general, these learning methods can be divided into five main categories [101, 102]:

1. **Supervised learning**: Within this learning framework, the algorithm learns from labelled data, consisting of input-output pairs. The goal is to learn a mapping from inputs to outputs, such that given new inputs, the algorithm can predict corresponding outputs with high accuracy.

2. **Unsupervised learning**: Here, the algorithms learn patterns and structures from unlabelled data, without explicit guidance on desired outputs. Instead, the algorithm identifies inherent relationships or structures within data, such as clusters or associations.

3. **Semi-supervised learning**: This is a learning paradigm that combines elements of both supervised and unsupervised learning. Here, the algorithm is trained on a dataset that contains a small amount of labelled data along with a larger amount of unlabelled data. The algorithm leverages the labelled data to guide the learning process, while also utilising the unlabelled data to extract additional information and improve performance.

4. **Self-supervised learning**: This is a form of learning that uses an unsupervised learning approach for tasks that canonically require supervised learning. This learning category can be broken down into two major components: *pretext tasks* and *downstream tasks*. In a pretext task, the chosen algorithm is trained to learn a meaningful representation of the unstructured input data. Subsequently, these representations within a latent space are used as inputs to a downstream task, such as classification.

5. **Reinforcement learning**: Although not explicitly used in this thesis, this is a machine learning paradigm, where an agent learns to interact with an environment in order to achieve a pre-defined goal or maximise a cumulative reward. The agent learns through trial and error, receiving feedback, in the form of rewards, or penalties based on its own actions. The goal of reinforcement learning is to discover the optimal strategy, or policy, that enables the agent to make the best decisions that lead to the highest possible reward over time.

With these definitions, the subsequent sections within this chapter will expand upon the distinct algorithms and their corresponding learning methodologies applicable to gait analysis framework proposed in this thesis, targeting anomaly detection, severity ranking, and longitudinal progression analysis.

2.3 Data exploration methods

Understanding of IMU data is of paramount importance as a first step towards a thorough gait analysis and detection of gait anomalies, marking the starting point of the patient's condition hierarchical assessment framework. This section focuses on data exploration methods, specifically dimensionality reduction and feature selection techniques, which aim to extract the most informative features from the information-rich data collected via wearable sensors. In the case of MS, early detection of

gait anomalies is necessary especially for mildly affected patients, who often have no clinically observable gait impairment. However, as more and more metrics are computed from IMU data, the high potential covariance among them might limit their clinical interpretation [17]. This highlights the need to reduce the number of features and only retain the most indicative MS biomarkers. This is to say that reducing of the number of features is advantageous, as it will allow clinicians to be presented with a smaller feature set, optimised towards better assessment outcomes and improved interpretability.

In addition to the need to reduce the feature space to a more manageable size in order to increase clinical acceptance and interpretability, the well-known 'curse of dimensionality' presents another imminent challenge in human gait analysis, as with other machine learning classification problems [103]. It was previously noted that beyond a certain point, including additional features can lead to a decrease in the performance of the classification system [104]. This is due to the incapacity to train a certain machine learning model to learn an adequate model due to insufficient observations (which are often related to the small number of participants in a clinical trial) compared to the large number of features extracted from the recorded gait signals. Therefore, mathematical and practical considerations dictate a need for reducing the feature space, either by dimensionality reduction methods, or though feature selection processes. For clarification, dimensionality reduction, in this case, refers to a transformation of data, whereas feature selection methods only extract subsets of the original set of features, without involving data transformations.

To reduce the dimensionality and redundancy in gait data, different methodologies are available. Perhaps the most used method is Principal Component Analysis (PCA), which is a linear and orthogonal transformation, in which the transformed features are sorted in a descending order according to their explained variance [105]. However, as a limitation, PCA is not capable of representing higher order, non-linear data composed of local structures [106]. To mitigate this issue, recently, another approach based on t-Distributed Stochastic Neighbour Embedding (t-SNE) method has been proposed by Mateen and Hilton [107]. This method is suited to deal with complex non-linear gait data in order to reduce dimensionality. Stochastic Neighbour Embedding (SNE) is a dimensionality reduction method which integrates a probabilistic approach [108]. SNE considers every point to be the neighbour of all the other points, by fitting a Gaussian probability distribution on each object in the high-dimensional space and preserving it in the low-dimensional embedded space.

However, this method is hampered by a cost function which is difficult to optimise and tends to crowd points in the center of the low-dimensional space. t-SNE is an improvement over the originally proposed SNE method, for which Mateen and Hilton [107] introduced a simpler cost function and used a Student t-distribution rather than a Gaussian distribution in order to calculate the similarity between two data points in the low-dimensional space [107]. As a result, the cluster structures in the original data are optimally identifiable and neighbourhood identity is preserved. Therefore, t-SNE facilitates understanding of the data by allowing visual representations of typical patterns and clusters in the data [109]. Even though PCA and t-SNE are well established methods within the gait analysis field [110–114], which can reduce dimensionality of the extracted candidate gait features, they alter their original representation, and the original semantics of the variables is not preserved (because of the transformation into the lower-dimensional space). Hence the understanding of the transformed features is hindered and difficult to interpret by domain experts [115, 116].

Reiterating, the process of feature engineering in gait analysis can lead to the accumulation of many features which may not prove to be either needed or significant for the gait impairment detection task. Moreover, their inclusion might also have a negative impact on the complexity and performance of the algorithms which will be used for classification. Since dimensionality reduction techniques do not preserve the original semantics of the gait variables, these are not desirable from a clinical perspective. Fortunately, to address these limitations, feature selection methods are available as an alternative. These methods are preferred, because the original candidate feature set is not altered. Instead, a smaller dimension subset of features can be selected [115]. Three main classes of feature selection methods are identified: filter, wrapper, and embedded methods [115–119]. The following paragraph provides a brief overview of these methods from a general machine learning perspective.

Filter methods perform feature selection as a pre-processing step, evaluating features based solely on the intrinsic characteristics of the training data, without involving any specific classification algorithm [120]. They rank features according to their relevance using statistical measures or distance metrics between classes [115, 121]. These methods are less computationally demanding and are able to achieve better generalization compared to wrapper methods, because they operate independently of any machine learning algorithms. [122]. However, a major disadvantage with filter methods is that they are unable to account for the interaction between features as

well as the interactions with the classifier used [115]. Another notable disadvantage is that filter methods tend to select subsets with a high number of features (in some cases even all the features), thus requiring an objective threshold in order to choose a subset [120]. Wrapper methods, in contrast, are integrated within a specific classification algorithm and evaluate feature subsets by systematically searching through different combinations. For each potential subset, they train and test the classifier to determine performance [115]. Advantages of wrapper methods include the ability to model feature dependencies and the ability to interact with the classifier, meaning that the selected features are evaluated on a specific machine learning classification algorithm. However, a common disadvantage of wrapper-based methods is that they are prone to overfitting and can incur high computational demands, especially when developing a classifier that is computationally intensive [115, 123]. Finally, embedded methods behave similarly to wrapper methods, with the only notable difference that they are built into the classifier construction. As a result, the computational complexity is reduced when compared to wrapper methods [115]. While this survey included filter, wrapper and embedded methods as the three main feature-selection classes, it is worth noting a more generic algorithm recently introduced in the relevant literature, even though it is not specifically used in this thesis. This is a correlation-based algorithm, known as RRCT, which jointly considers feature relevance, redundancy and complementary trade-off [116]. This feature selection method employs a straightforward nonlinear transformation of correlation coefficients, leveraging information theory principles to measure the relationship between features and the response, as well as the shared information among features. It also directly considers feature interactions as an integral element in feature selection. At this point, it becomes clear that the optimal subset of features indicative of the presence of gait impairments in MS might be revealed by utilising feature selection methods. Given the above-mentioned advantages of wrapper-based methods and considering both the preferred anomaly detection technique which will shortly be presented in Section 2.4, as well as the size of the dataset used in Chapter 3, the computational disadvantage of the wrapper-based feature selection methods is not considered to be problematic. As such, this will be the preferred feature selection technique employed in this thesis. This procedure would lead to the creation of a conceptual gait model for MS, as only the most relevant features responsible for the detection of the gait impairment are selected. The clinical meaning of the features is also preserved, which supports the interpretability of the model. Finally, equipped with all the necessary background information, the attention can now shift

towards the specific problems targeted in the previously-established hierarchy for gait condition monitoring.

2.4 Gait anomaly detection

Given that the condition of a patient can be explored following the hierarchy proposed in the introductory chapter, one might start the analysis at the lowest level of the hierarchy and identify features which can give a good indication of the presence of gait impairments for PwMS, to maximise the prospects of early detection. This task is considered crucial for MS, especially for patients who often have no apparent gait impairment during the prodromal stages of the disease, to allow clinicians to devise early optimal treatment plans. Once particular features are computed from the raw measurements, a classification problem arises, as observations need to be categorised according to whether they arose from a HC or an individual affected by MS. To this extent, inspired by the SHM field, an accepted general principle in the form of an axiom was formulated by Worden et al. [124], stating that only anomaly detection can be done in an unsupervised way. In the case of this work, the learning procedure is performed on a dataset containing examples only from the assumed healthy condition of a subject. Here, the 'healthy condition' is defined as the baseline state of individuals who have no prior history of musculoskeletal or neurological disorders. The primary classes of algorithms designed for unsupervised learning applications are those based on outlier detection [124]. Therefore, based on outlier analysis for multivariate data, in order to detect gait anomalies in MS, one simply has to identify if the measured data has deviated from the healthy condition [125].

One popular method of outlier detection relies on a discordancy test which computes the Mahalanobis Squared Distance (MSD) [126] for a possible outlier. At its core, originally proposed by P.C. Mahalanobis in [127], the MSD measures the distance between a point and a distribution in a multivariate space, accounting for correlations among variables. To detect outliers, the MSD is compared against an objective threshold [125]. Worden et al. [124] define a discordant outlier in a dataset as an observation which is found to be inconsistent with the rest of the data. Therefore, the outlier is believed to have been generated by a different mechanism to the 'healthy condition' data. In the current work, it is hypothesised that the discordancy could

have been generated by the presence of gait impairment affecting the MS population. However, detection of outliers using multivariate data can be further complicated due to the presence of inclusive outliers in the healthy control dataset. Here, inclusive outliers refer to inconsistent data points present in the healthy condition data included in the training set. Their presence is not desirable, due to their 'masking effect' [128], which contaminates the training data. In response to this problem, methods have been developed for dealing with inclusive outliers such as the minimum covariance determinant (MCD) estimator [129]. The mathematical considerations for the discordancy test will be given in detail in Chapter 3. To the best of author's knowledge, this approach, although not novel in general [51, 124, 125, 128, 130, 131], has not been applied specifically for detection of gait anomalies or any other biomedical applications. However, its successful application in the field of SHM as a damage detection method, motivates its adoption in the field of gait analysis.

2.5 MS severity quantification

Having introduced the relevant work for gait anomaly detection, this section will introduce the literature related to the subsequent task of quantifying disease severity. Effectively, the previously established binary classification problem is now expanded to a multiclass problem. Considering the overarching goal of remote gait monitoring, outside laboratory conditions, the approach proposed in this thesis seeks to solve this problem using a single IMU positioned on the lower back. The clinical relevance of this approach has been already discussed in Section 2.1. However, recognising the complexity of the data acquired from the lower back sensor, and the concerning accuracy of gait event detection algorithms using the lower back sensor [30, 74, 79], alternative approaches that are not relying on the accurate detection of gait events and circumventing the need of 'expert' feature selection techniques seem favourable. Recently, Creagh et al. [18] has demonstrated that is possible to accurately distinguish across disability levels in MS. However, this study only included mildly and moderately affected individuals, which is not representative of the full spectrum of MS disability levels. Nonetheless, their approach motivated the usage of neural networks within this context. A neural network is can be defined as mapping of the input data to an output, using chain of nonlinear transformations. Over the years, convolutional neural networks (CNNs) have become increasingly used within the gait analysis field, for a variety of tasks, ranging from activity recognition, to biometric authentication

based on kinematic gait data or severity disease quantification [16, 18, 132–139]. At this point, it should be noted that the mathematical introduction to all of the technologies introduced in this chapter is postponed until Chapter 4. Given the substantial inter- and intra- patient variability of MS-related symptoms which can potentially change at different rates across different timescales, the classification task encountered at this level of the hierarchy can be simplified by dividing this task into two components. The first component, aims to find a latent space, where similar patterns are grouped together, whereas dissimilar patterns lie further apart. This idea is motivated by the advancements in *contrastive learning* [140–143]. Next, once a suitable representation space has been found using this technique, the downstream classification task can be simplified [101]. The second task primarily entails linking the output of the optimised network, used for the contrastive learning task, to additional layers. These layers are subsequently employed to predict the disease class [102].

However, although neural network-based techniques yield very promising results in terms of classification accuracy, their major limitation remains their black-box character [144]. Here, the black-box character denotes the inability of neural network-based methods (particularly deep learning methods [102]) to offer any insight into the physical understanding of the data being processed. This drawback of machine learning techniques can be problematic, especially in applications such as medical diagnosis, as it can hinder their clinical acceptance [145]. As such, for responsible clinical adoption of machine learning models for gait analysis, researchers must elucidate how such algorithms arrived at their predictions. This transparency is crucial to verify that the models are learning clinically meaningful gait characteristics, and ultimately agree on the adoption of a specific method.

In a response to the lack of understanding and interpretation of the decision-making process of neural networks, the field of explainable artificial intelligence (XAI) [146] also gained more and more attention in the recent years. In general, the aim of XAI is to demonstrate how complex non-linear machine learning methods operate and why they reach a certain prediction, to justify the reliability of such methods. Even though XAI is still at an early stage of its development, recent applications in medicine have already raised attention [146, 147]. It should be noted that for a thorough classification of gait data, the explanation method chosen should clearly reveal all the clinically relevant regions within the gait patterns, which are associated with a particular gait impairment and might serve as biomarkers for disease progression.

Even though the fundamental mathematical principles in neural network-based methods are understood, why a particular prediction has been made remains often unclear. To this extent, two main categories were identified for explaining the local behaviour of a neural network model: self-explaining models and post-hoc methods.

As a general rule, self-explaining models learn how the relationship between the input and the output relates to a predefined library and generates explanations for the predictions. For example, a self-explaining approach was introduced by Hendricks et al. [148]. The method proposed does not visually highlight relevant regions of the input data, but rather provides textual explanations of the prediction. However, the main limitation of self-explained models is that this approach cannot be used on previously trained models. On the other hand, post-hoc models allow for much greater versatility, as they can be directly applied to already trained models. These models can be further decomposed into 4 distinct categories, namely: i) perturbation-based, ii) function based iii) surrogate- / sampling-based and iv) structure-based, all of which are specific to deep learning methods. Perturbation-based models aim to evaluate the importance of the inputs by measuring how a classifier reacts to changes in the input [149, 150]. Next, function-based methods regard neural networks as a function and provide a functional approach of the explanation (i.e. the gradient or sensitivity values can be computed, or the function can be approximated based on Taylor decompositions) [151]. The surrogate- / sampling-based methods approximate a prediction locally, offering a model-agnostic approach by learning an interpretable model locally around the prediction [152, 153]. Finally, the structure-based approach uses the architecture of machine learning models to explain how the predictions were made [154–156].

In direct connection to the field of clinical gait analysis, the first steps towards model prediction explanation were taken by Hurst et. al. [136] and Slijepcevic et. al. [109]. Both research studies used the so-called layer-wise relevance propagation (LRP) [154] in order to explain a classifier's decision for the identification of healthy subjects from recorded gait data using floor sensors and optical motion capture systems. The proposed approach uses LRP, which works by decomposing the output of a given activation function for any input and attributes relevance scores. [136] successfully employed LRP as an explanation method, in order to distinguish between healthy individuals, based on their gait signatures and verify that deep-learning methods are learning meaningful gait characteristics, in order to highlight which variables at which time frames were used to identify the subjects. However, their approach seems

questionable, due to the body weight normalisation procedure used for the ground reaction forces. This normalisation procedure seemingly disregards the influence of the body weight, an aspect which holds potential significance in the context of biometric identification systems [157], thus overcomplicating the classification problem. Later, Creagh et al. [18] used the same LRP methodology to verify the clinical relevance of the features learnt by a neural network from acceleration signals recorded with a smartphone. In this study, the goal of the neural network was to accurately predict disability levels of MS-affected individuals. While the study highlighted the increased relevance scores corresponding to distinct step inflections regions in the gait signals, additional validation procedures under more controlled environments are necessary. Nonetheless, the explainability method used in their study based on LRP seems to be a very powerful visualisation tool for highlighting the most relevant gait characteristics in the input signals, and is worthy of additional investigations.

2.6 Monitoring longitudinal disease progression

The final stage in the hierarchical framework proposed for evaluating the condition of an individual consists of making inference about the progression of the disease. However, a clear distinction should be made here between longitudinal monitoring and prognosis. The first aspect involves tracking the evolution of the individual gait characteristics over time, while prognosis pertains to predicting the likelihood of future gait patterns. It should be noted the latter is a much more challenging task. This complexity is also mirrored in SHM applications, which have inspired the majority of the methodologies employed in this thesis [158]. While there appears to be a wealth of cross-sectional studies targeting MS gait pattern characterisation [17–19, 24, 25, 31–37, 39, 42, 43, 159], there is a relative scarcity of studies that directly employ longitudinal data [16, 45, 49, 50, 56]. Considering longitudinal monitoring in MS, Spain et al. [56] used gait and balance measures extracted from IMU gait signals to monitor MS progression during an 18-month longitudinal study. Interestingly, no gait parameter was significantly altered over the time frame of the study. It is believed that results might have been obtained either because no changes appear over such a short period of time or because gait of PwMS is susceptible to compensatory strategies [160], which are intrinsic processes used to offset or bypass gait deficits, masking the underlying decline. However, an alternative explanation might be that

the tools used in the analysis are likely to not offer enough sensitivity to capture the small alterations in gait due to progression of the disease over the respective timescale. Kaur et al. [42] focused on examining MS and the disability-related changes in spatio-temporal and kinetic gait features using an instrumented treadmill and leveraging machine learning approaches. The study revealed that temporal, spatial and spatio-temporal gait features may be clinically used to design machine learning-based MS prediction strategies or monitoring its progression. However, the small cohort size included for this study might limit the generalizations of interpretation for the MS community. Additionally, treadmill walking might still be a novel task for some participants, even though accommodation time was accounted for in the study. As [36] also argued, this approach might therefore be not ideal for representing the true real-life walking patterns of MS. Angelini [22] assessed the between-session reliability of several temporal, variability and balance gait metrics using data collected one-week apart from MS-affected individuals in a clinical setting. The study focused on the consistency of the included gait metrics over time, and found good to excellent reliability between the two repeated tests. Within this context, the timescale of this study allowed to test a given gait model within a period in which the disease status of PwMS remained constant. In light of the results presented in [56] and [22], it is believed that the 'expert' feature selection approach utilised in these studies are not offering a good enough resolution to capture any longitudinal changes. Moreover, it can also be argued that this approach can lead to overlooking a significant amount of valuable clinical data, potentially encoding important information about the health condition of specific individuals [91]. More recently, Creagh et al. [16] monitored disability fluctuations in PwMS over a 24-week period. Their approach leveraged raw accelerometer data collected during remotely administered walking tests and employed a previously established neural network [18] for predicting disability levels. Contrary to the results presented above, this study identified significant variations in disability levels throughout the monitoring period. However, limitations were noted. While the unified disability metric demonstrated promise in reflecting gait decline consistent with clinical assessments, participant disability estimates exhibited occasional inaccuracies. Additionally, the methodology lacked a clear clinical interpretability of the results. As acknowledged by the authors, uncontrolled environmental factors or other external influences during remote testing likely contributed to these limitations.

To this end, there is a need for alternative modelling strategies, which should be general enough to account for natural sources of variability present at follow-up

assessments, while also exhibiting sufficient resolution to effectively capture any mobility decline or improvements throughout the heterogeneous disease course of MS. Within this context, a first consideration should be given to autoregressive models, which are a linear representation of a dynamic model in discrete time, predicting current values from past observations. Drawing inspiration from the SHM field, autoregressive models have proven successful for structural damage identification using accelerometer data [161–164]. In SHM, the model residual (the difference between measured data and the model predictions) is often used a a damage-sensitive feature. While the gait analysis community has seen very few applications of such models [165], they seem worthy of investigation as a first step towards establishing an objective gait consistency metic. As will be further detailed in Chapter 5, the idea is to establish a model during a baseline assessment and then deploy the model at a later point, monitoring the residual patterns. As a result, when fundamental characteristics describing the overall body movement occur, then these changes should be reflected in the residual signal.

While autoregressive models provide a valuable baseline, exploring non-linear alternatives may be necessary to fully capture the underlying dynamics of the gait patterns. In this regard, Gaussian Process Regression (GPR) [166] offers a powerful data-driven approach for modelling non-linear gait data, due to its ability to capture complex relationships without making restrictive assumptions about the underlying functional form. In addition, Gaussian processes (GPs) have the added advantage of providing automatic uncertainty estimates. This property is particularly relevant for modelling MS-affected gait patterns, given the intrinsically unpredictable disease progression [16, 46]. Within the gait analysis field, Wang et al. [167] proposed a dynamical GP model for human motion, which was later used in [168, 169] and [170] to generate reference trajectories for robotic gait rehabilitation systems. [171] also used GPR for mapping body parameters to gait kinematics. [172] attempted to model the lower limb joint kinematics of individual subjects and showed that GPs can learn a mapping between patient's gait and therapist-assisted gait. However, limited conclusions can be drawn from this study, as a result of the limited number of subjects included. [173] introduced GP regression for learning the relationship between body parameters and gait features at different walking speeds, as part of a developmental pipeline designed for individualized lower limb exoskeleton robots, while [174] used deep GPs for online gait prediction during human-exoskeleton interaction. In a different context, [175] introduced GPs as a regression tool for efficient sensitivity analysis aimed at reducing the complexity of musculoskeletal modelling, in response to the

computational challenges offered by traditional Monte-Carlo methods. It can be seen that there have been numerous approaches towards modelling gait patterns and that investigation into this problem is an active field of research. Although GPR has found widespread application in modeling healthy gait patterns for robotics applications, its adoption for modeling pathological gait patterns remains limited. It should also be noted the inherent challenges associated with the scalability of GPs are seldom addressed [174]. Moreover, it is also noticed that almost all studies presented here are exclusively using constant-noise modelling approaches. Whereas, in the context of MS gait pattern modelling, heteroscedastic noise modelling approaches [176] might be more suitable. Nonetheless, the full motivation regarding these concepts is postponed until Chapter 6.

2.7 Conclusions

It is clear from the literature that gait analysis in MS is a field of research that is still impacted by several major challenges. The first challenge lies in detecting MS onset, which is essential for providing end users an early opportunity for therapeutic interventions. Although an MS-specific conceptual gait model has been proposed by Angelini et al. [17], integrating additional summary features remains a promising area of investigation and represents a key aspect addressed in this thesis. While gait anomaly detection is only the starting point of the hierarchical framework for assessing the condition of MS-affected individuals, departing from a pre-defined feature set might also prove advantageous. However, this departure introduces new challenges, particularly at the disease quantification level of the hierarchical framework, where the goal is to distinguish across the full severity spectrum of the disease. A critical challenge in this context is ensuring model interpretability to facilitate clinical acceptance, alongside maintaining generalizability to unseen data. Addressing this interpretability challenge represents another key aspect of this thesis, which develops novel approaches that maintain clinical interpretability while advancing disease severity quantification.

Another major challenge concerns the selection of suitable models for longitudinal assessments. Currently, no established best practices exist for longitudinal gait monitoring or prognosis, hindering the understanding of MS progression over time. While data sparsity and over-reliance on cross-sectional studies may partly explain

this gap, another possibility lies in the insufficient resolution offered by the models employed in current studies. This fundamental gap in longitudinal monitoring frameworks represents the third key challenge addressed herein. The final longitudinal monitoring and prognosis task is approached using alternative data-driven methods, proposing a novel modelling methodology capable of personalised longitudinal diagnosis and prognosis that aims to capture subtle changes in gait patterns over time.

Chapter 3

Gait anomaly detection - an outlier detection problem

The main body of this chapter is concerned with the first level of the proposed hierarchical framework for evaluating the condition of a patient, centered around the detection of anomalies in the gait of PwMS. Gait anomaly detection is effectively a one-class classification problem, which, fundamentally, seeks to ascertain whether an observation within a dataset aligns with the typical gait patterns exhibited by healthy individuals. This task necessitates a classification tool capable of evaluating the congruence of the observed gait features with normative patterns. Although it is important to recognize the availability of alternative options [177], given its prominent utilization in the SHM field [51, 124, 128, 130, 131], the Mahalanobis Squared Distance (MSD) emerges as suitable tool for the gait anomaly detection task. At its core, the MSD measures the distance between a point and a distribution in a multivariate space, accounting for correlations among variables.

In the context of gait analysis, the MSD paradigm, which was originally framed as statistical process control method, offers a tailored approach to quantifying deviations from healthy gait patterns observed in MS patients. By modelling gait parameters as multivariate distributions, following careful feature engineering (including feature extraction and feature selection), the MSD facilitates the identification of subtle anomalies that may evade conventional analysis. Moreover, its inherent capacity to adapt to individual variability and account for interdependencies among gait parameters renders it well suited for this task. This offers objective, quantitative

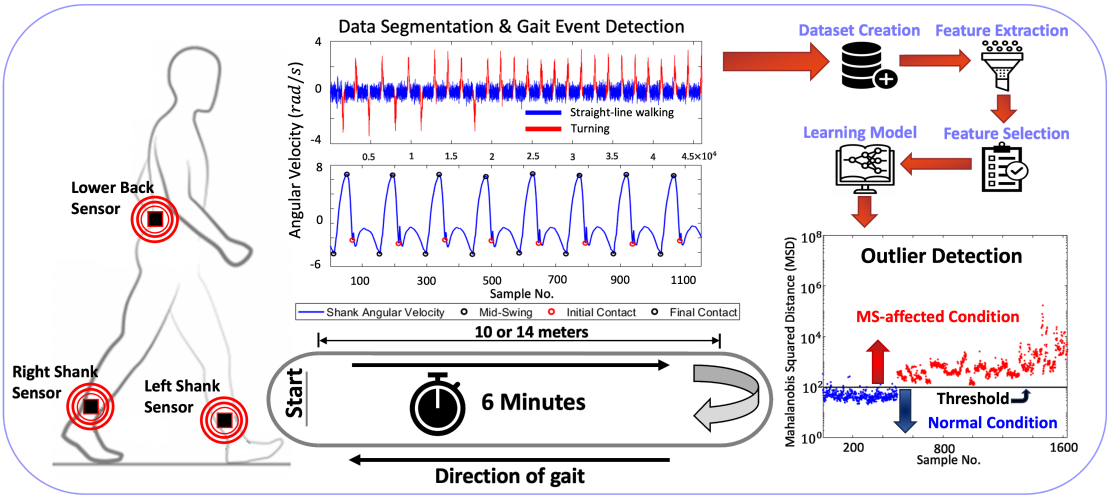


Figure 3.1: Flowchart of the gait anomaly detection approach.

assessments, and it will also be shown that it can even detect subtle gait anomalies even in the prodromal stage of the disease, which traditional methods have previously failed to capture [178]. As a result, its clinical adoption may have the potential of facilitating timely interventions and improve patient outcomes.

This approach for assessing the presence of anomalies in the gait of MS-affected individuals is a multifaceted process involving several stages, as depicted in the flowchart presented in Figure 3.1. The first step involves data segmentation and detection of gait events. Then, the process of feature engineering commences, starting with dataset creation and followed by feature extraction and feature selection though relevance ranking. Finally, the overarching objective is the development of a robust model capable of distinguishing between healthy gait patterns and anomalous ones. In the case of the MSD-based detector, this task is facilitated through a comparative analysis of the MSD against an objective threshold. Here, MSD values exceeding the threshold are considered to be indicative of departures from the healthy condition, highlighting gait abnormalities, whereas values below the threshold are deemed as healthy (i.e., 'normal'). In effect, the MSD is used to assess MS-affected gait patterns, summarising the extent of gait impairment using a single, unified distance metric. However, the gait anomaly detection task can compounded by the presence of inclusive outliers in the training data. It will be demonstrated herein that despite adopting an unsupervised approach and training the classifier using only HC data, the necessity of robust metrics persists.

3.1 Outlier discordancy measure computation for gait impairment detection

In the case of outlier detection, one is attempting to identify whether an observation in a dataset is part of the healthy condition or not. Worden et al. [124] define a discordant outlier in a dataset as an observation which is found to be inconsistent with the rest of the data. Therefore, the outlier is believed to have been generated by a mechanism of some sort and in the case of this work, the discordancy is believed to be generated by the presence of gait impairment in PwMS. The hypothesis introduced here is that if someone has a gait impairment, then the associated datapoint should fall outside the norm, relative to the (assumed) 'healthy condition'.

Analogous to the case of damage detection in SHM, the identification of outliers through an unsupervised methodology necessitates the availability of a training dataset. Following the principles proposed in [125], the training dataset should exclusively comprise samples representing the healthy, unaffected state of HCs. Considering the large accumulation of features extracted following the data acquisition and segmentation procedure depicted in Figure 3.1, another important consideration is the necessity for a multivariate discordancy test.

Here, the Mahalanobis distance, originally proposed in [127], emerges as a suitable tool for outlier detection. It measures how similar an observation is relative to the (assumed) normal distribution of the training data, containing only HC observations. This contrasts the Euclidean distance, which only measures the shortest distance between two points is space, without taking into account any correlations between variables. The Mahalanobis distance scales the contribution of individual variables to the computed distance value, taking into account the variability inherent to each individual feature.

The Mahalanobis distance is a scale invariant metric, and can be interpreted as a covariance-weighted squared-Euclidean distance from the sample mean $\boldsymbol{\mu}$ of the healthy condition data. Its definition, and therefore, the magnitude of discordancy is given in Equation 3.1. It should also be noted that the notation used in this thesis involves representing vectors using bold typography, while matrices are identified by the uppercase letters.

$$MSD(\boldsymbol{x_i})^2 = (\boldsymbol{x_i} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}) \tag{3.1}$$

where $\boldsymbol{x_i} = [x_{i1}, \ldots, x_{iM}]^T$ is a single potentially outlying observation in a multivariate data matrix $X = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}]^T$ which operates in a M dimensional feature space and consists of N observations. Here, Σ denotes the covariance matrix. Notably, if the covariance matrix is equal to the identity matrix, then the Mahalanobis distance becomes synonymous to the Euclidean distance. In summary, the MSD quantifies the likelihood of *new* observations, given the known training data, representing the (assumed) healthy condition.

The main drawback of the classical MSD distance as presented here lies in the detrimental influence of inclusive outliers, which can have a masking effect. If clusters of outliers exist within the training data, they would have a critical influence on the sample mean and covariance. Subsequently, the MSD computation may show minimal distances for new observations or outlying data, thereby rendering the outliers undetectable. The sample mean and covariance matrix are particularly vulnerable to the presence of inclusive outliers. Specifically, when these reside within the training data cloud, they shift the sample mean towards them and can also expand the classical tolerance ellipsoid in their direction [179]. Arguably, the performance of the MSD-based detector is optimal when the training set comprises only 'clean' HC samples. However, in the context of gait analysis, this idealized scenario may not be readily available due to several factors, thus further complicating the detection problem. Practical implementations often encounter inclusive outliers stemming from a variety of sources, such as unaccounted previous injuries, inconsistencies in data pre-processing techniques, which were overseen by the operator or fluctuations in the environmental assessment conditions, among others. The presence of such outliers necessitates a more robust approach, invariant to these limitations, particularly in critical scenarios where the distinction between healthy and abnormal conditions is critical for accurate analysis and decision-making.

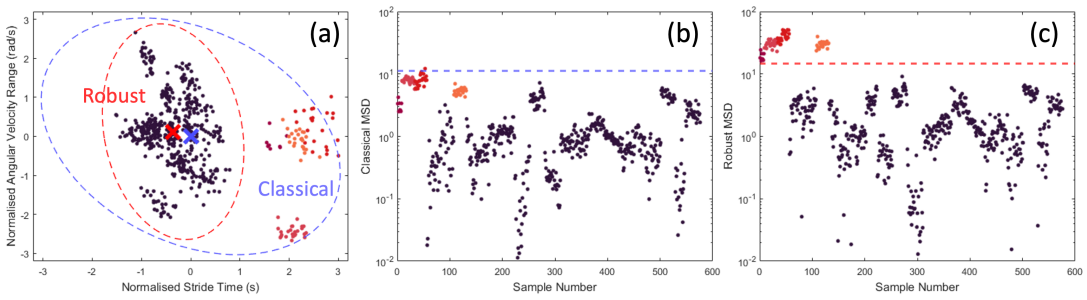


Figure 3.2: (a) - Bivatiate gait data scatter plot, including tolerance ellipses and center locations, represented by crosses; (b) - Classical MSD computation, including the objective threshold; (c) - Robust MSD including the objective threshold.

A typical example of inclusive outliers present in the HC dataset is provided in Figure 3.2. On the left, an example of bivariate gait data is shown. Here, the normalised shank angular velocity range around the medio-lateral axis is plotted against the normalised stride time. The inclusive HC outliers resulted from misclassification of gait events are highlighted in shades of red. The ellipses represent those points whose MSD equals the square root of the 0.99 chi-squared quantile with 2 degrees of freedom, while the crosses represent the center of the data clouds for the classical MSD (blue) and the robust counterpart (red), exclusive of outliers. The influence of the inclusive outliers is immediately visible here. In Figure 3.2b, the classical MSD yields small distances for the inclusive outliers relative to the rest of the control data cloud, as a result of considering all datapoints as being 'normal'. On the other hand, the red ellipse is much smaller and encloses only the 'normal' datapoints, therefore excluding and exposing the inclusive outliers. As a result, the MSD of the inclusive outliers is now much higher and can be clearly separated from the 'normal' points, as seen in Figure 3.2c. Figures 3.2b and 3.2c also show an objective threshold which is used to classify a sample as being an outlier or inliner, based on the magnitude of discordancy. The procedure employed for establishing the thresholds will be presented to the reader in Section 3.1.2.

3.1.1 Computation of robust statistics using the Minimum Covariance Determinant (MCD) estimator

By detecting the inclusive outliers at an early stage, the prospects of achieving good generalisation of the true condition of HCs would significantly increase the classifier accuracy. Therefore, the application of robust statistical metrics for the estimation of the discordancy metric in the MS-affected gait patterns holds significant interest. As such, the method introduced here is based on the estimation of the minimum covariance determinant (MCD) [180, 181]. In essence, the MCD estimator searches for H observations in the dataset whose covariance matrix has the lowest possible determinant. Nonetheless, the estimation of the MCD is not a trivial procedure, requiring exhaustive computations [130]. In practice, efficient implementations of the MCD algorithm have been developed. Therefore, in the case of this work, the FAST-MCD algorithm [129] has been used through a MATLAB library called LIBRA [182]. For the sake of brevity, only an overview of the algorithm will be presented here for the reader. For the extensive details regarding the implementation, users are referred to [129].

Firstly, it is assumed that $X \in \mathbb{R}^{N \times M}$ is the multivariate data matrix, operating in a M dimensional feature space and consisting of N observations. Reiterating, the MCD algorithm begins by searching for H observations whose classical covariance matrix has the lowest possible determinant. The average of these H observations is used to compute the location estimate of the *center*, $\hat{\boldsymbol{\mu}}$ and the corresponding *scatter* matrix, $\hat{\Sigma}$, multiplied by a consistency factor. Based on these estimates, a reweighting step can be added to increase sampling efficiency. The MCD estimator is most robust when $H = (N + M + 1)/2$. Note, that the MCD estimator can only be computed when $H > M$, otherwise the covariance matrix of any H subset has a determinant of 0. When a large proportion of outliers are contaminating the data, it is recommended to set H as $H = \alpha \times N$, where $\alpha = 0.5$ [182]. An interesting property of the MCD estimator is its *affine equivariance*, implying that the robust statistical estimates of the location *center* and *scatter* matrix are consistent even when the data is subjected to certain transformations such as rotation, translation or scaling. The exact computation of the MCD estimator presents significant computational challenges, as it requires the evaluation of all combinations of H subsets from a total of N observations. To this end, [129] proposed an approximate algorithm, whose key component lies in a Concentration step (C-step), as explained in the following

paragraph.

Firstly, let H_1 to be a subset of X of size H. As such, the mean location vector and the corresponding covariance matrix are computed as:

$$\hat{\boldsymbol{\mu}}_{\mathbf{1}} = \frac{1}{h} \sum_{i \in H_1} \boldsymbol{x_i} \tag{3.2a}$$

$$\hat{\Sigma}_1 = \frac{1}{h} \sum_{i \in H_1} (\boldsymbol{x_i} - \hat{\boldsymbol{\mu}}_{\mathbf{1}})(\boldsymbol{x_i} - \hat{\boldsymbol{\mu}}_{\mathbf{1}})^T \tag{3.2b}$$

If $|\hat{\Sigma}_1| \neq 0$, the relative Mahalanobis squared distances are then computed with respect to the centroid $\hat{\boldsymbol{\mu}}_{\mathbf{1}}$ and the scatter $\hat{\Sigma}_1$ estimated from H_1, as follows:

$$MSD(i)^2 = (\boldsymbol{x_i} - \hat{\boldsymbol{\mu}}_{\mathbf{1}})^T \hat{\Sigma}_1^{-} 1 (\boldsymbol{x_i} - \hat{\boldsymbol{\mu}}_{\mathbf{1}}) \quad for \quad i = 1, 2, ..., M \tag{3.3}$$

Next, the concentration step commences by selecting another subset H_2 of size H, corresponding to the smallest distances computed using Equation 3.3. Subsequently, $\hat{\boldsymbol{\mu}}_{\mathbf{2}}$ and $\hat{\Sigma}_2$ are calculated based on the observations contained in H_2. It is expected that $|\hat{\Sigma}_2| < |\hat{\Sigma}_1|$, with equality being achieved if and only if $\hat{\boldsymbol{\mu}}_{\mathbf{1}} = \hat{\boldsymbol{\mu}}_{\mathbf{2}}$ and $\hat{\Sigma}_1 = \hat{\Sigma}_2$. If $|\hat{\Sigma}_1| > 0$, the C-step easily selects a new subset with lower covariance determinant. This concentration step is usually repeated until converge, i.e. $\hat{\Sigma}_{new} = \hat{\Sigma}_{old}$. The succession of determinants obtained through this methodology is bound to converge within a finite number of iterations due to the finite number of subsets available, typically achieving rapid convergence in practical scenarios. Nevertheless, it should be noted that reaching convergence to a global minimum of the MCD objective function is not guaranteed. Consequently, an approximate MCD solution can be derived by initiating the process with numerous choices of H_1 and executing C-steps on each, retaining the solution characterised by the lowest determinant value.

3.1.2 Threshold computation

Setting an appropriate objective threshold based on healthy condition data is not a trivial task. The value of the threshold must be dependent on the number of observations in the training set, as well as the number of features (i.e. the dimension)

of the problem being studied. In the case of this work, the objective threshold values are computed using a Monte-Carlo simulation based on extreme value statistics, following [128]. The procedure that was conducted to set up the threshold is summarised as follows:

- A $N \times M$ (*number of observations* $\times$ *number of features*) matrix is constructed and populated with values drawn from a zero mean, unit standard deviation normal distribution.
- The discordancy value computed using the MSD is evaluated for all matrix entries, where the mean, robust and classic covariance matrices are inclusive.
- The largest (extreme) value of the MSD is then stored.
- The procedure is repeated a number of times (1000 times in the case of this work) in order to generate an array of extreme distance calculations.
- All the extreme values are ordered accordingly, in ascending order of magnitude.
- Finally, the threshold is selected as the value corresponding to a 99% confidence interval.

3.1.3 Discordancy test performance metrics

To quantify the accuracy of the discordancy test, a number of performance metrics (*accuracy*, *precision*, *recall*, *specificity* and *F1 score*) are computed as stated in Equations 3.4a-e, where TP are true positives, TN are the true negatives, FP are the false positives and FN are the false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4a}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.4b}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.4c}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.4d}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.4e}$$

Since single measures are not able to fully determine the classification performance of the outlier test, several performance indicators are necessary. The accuracy is used as a measure which compares the proportion of samples classified correctly to the total number of samples being examined. The precision measures the proportion of MS-affected individuals which have been classified correctly, compared to all individuals taking part in the test set. The recall measure (which is also referred to as sensitivity) is particularly useful to quantify the rate of detection of the gait anomalies in PwMS. The specificity is used as a measure of detecting how many of HC individuals were deemed as outliers by the MSD-based algorithm. Finally, the F1 measure is provided if one must seek a balance between precision and recall.

3.1.4 Sequential feature selection

A final consideration of the gait anomaly detector proposed in this chapter takes into account the selection of relevant features. This aspect is particularly important for the development of conceptual gait models in MS, especially when the candidate features have been selected based on several other successful models proposed for difference pathologies, as well as additional statistical metrics. To this end, identifying MS-specific characteristics becomes a necessity for achieving good predictive power and superior interpretability. This work uses a wrapper-based feature selection method, due tho their ability to consider feature dependencies and their interaction with the classifier [115, 123], as described in Section 2.3.

In this work, a modified version of the sequential forward selection algorithm (SFS) [183] is used. The SFS algorithm follows a greedy approach and stands out for its iterative nature, progressively building the feature set by adding one feature at a time based on its impact on classification performance. Following initialization of the three empty feature sets, namely $F_{selected}$ and $F_{dropped}$, the algorithm is implemented such that it commences by testing the classification performance of pair-wise combinations of features using the MSD-based classifier, as previously detailed, by comparing the discordancy measures of the test set with the objective threshold established during training. The cost function used here is the misclassification error, defined according to Equation 3.5

$$MCE = \frac{FP + FN}{N} \tag{3.5}$$

Subsequent iterations of the SFS algorithm focus on expanding the feature set by identifying additional features that minimize the MCE. Each feature under consideration is evaluated based on its ability to further reduce the MCE when incorporated into the existing feature set. This iterative process continues until a stopping criterion is met, typically defined by reaching a predetermined number of selected features, or when all combinations of features have been visited.

To address the issue of feature redundancy, a critical consideration in feature selection, a specific condition is integrated into the SFS algorithm. This condition stipulates that in cases where multiple features yield the same minimum MCE, only the first feature meeting this criterion is included in the feature set, $F_{selected}$ while the redundant ones are added to the $F_{dropped}$ set. The rationale behind this condition is to minimize the risk of selecting highly correlated features, which could introduce redundancy and potentially compromise classification accuracy. By enforcing this condition, the algorithm ensures that the final feature set comprises diverse and informative features, thereby maximizing its discriminatory power. It is noteworthy that while the MSD-based classifier operates under an unsupervised paradigm, lacking labelled data during training, the wrapper-based feature selection technique introduced herein effectively transitions it to a supervised framework. This is the case, as the feature selection process is guided by the performance of a predictive model (the wrapper) that is trained and evaluated on labelled data.

Finally, upon completion of the iterative process, the optimal feature set, $F_{optimal}$ corresponding to the global minimum MCE is returned:

$$F_{optimal} = \arg \min_{F_{\text{selected}}} (MCE_{global}) \tag{3.6}$$

The pseudocode for the modified SFS algorithm used in this work is given below, in Algorithm 1.

Algorithm 1 Sequential Feature Selection Algorithm for Outlier Detection

Input: X_{train} - Training data (containing only HC observations), X_{test} - Test data (containing both HC and MS observations), y_{test} - True labels of test data
Output: F_{Optimal} - Optimal feature set
Parameters: $k_{\max}$ - Maximum number of features allowed in F_{selected}
Initialize: $F_{\text{selected}} \leftarrow \emptyset$, $F_{\text{dropped}} \leftarrow \emptyset$, $MCE_{\text{global}} \leftarrow \emptyset$
while $|F_{\text{selected}}| < k_{\max}$ or $|F_{\text{selected}}| + |F_{\text{dropped}}| = \text{M}$ **do**
    Evaluate the MCE for each feature combination, and store those in MCE_{vec}
    $MCE_{min} \leftarrow \min(MCE_{vec})$
    Select feature combinations corresponding to the minimum error, MCE_{min}
    if Multiple combinations yield the same MCE_{min} **then**
        Append the first feature to F_{selected}
    else
        Append the redundant features to F_{dropped}
    end if
    $MCE_{\text{global}} \leftarrow MCE_{\text{global}} \cup MCE_{min}$
end while
return $F_{optimal}$, according to Equation 3.6.

3.2 A case study to demonstrate the gait anomaly detection framework.

Having established the methodological details regarding the implementation of the gait anomaly detector, this chapter now transitions to its practical application. The aim of this section is to demonstrate the unique approach for gait anomaly detection, based on the computation of the MSD and its comparison to an objective threshold. Initially, the datasets and the associated pre-processing steps are introduced. This is followed by a presentation of an initial set of 'expert' features (i.e. features describing rhythm/variability, pace, symmetry and dynamic balance domains). These features have been previously used to describe an MS-specific conceptual gait model, as introduced by Angelini et al. [17]. Their clinical relevancy has already been discussed in Chapter 2, Section 2.1. Subsequently, the feature set is then augmented with supplementary statistical metrics, as well as additional temporal metrics. The statistical measures include the mean, variance, standard deviation, range, and extreme values (minimum or maximum), which are calculated across multiple sensory channels, encompassing both acceleration and gyroscopic measurements, in all three anatomical directions. This analysis is applied to data from sensors located on the

shanks as well as the lower back sensor. The classification outcomes, inclusive of feature selection via the SFS method, are then presented for both the initial and expanded feature sets. The chapter ends with the conclusions drawn from this work.

3.2.1 Participants

Remembering that training of the MSD-based anomaly detector requires that the threshold is computed based on the (assumed) healthy condition data, in the case of this work, the training set comprised 70% randomly selected datapoints from the 24 HCs. These HC individuals were included in the first dataset presented in Section 1.3. For the validation set, used to assess the performance of the MSD-based gait anomaly detector, the remaining 30% of the HC datapoints were included, along with data from 72 demographically matched PwMS. Among the PwMS, 30 subjects were selected from the baseline assessment of the second dataset presented in Section 1.3, while 42 additional participants were selected from baseline assessment of the third dataset. In order to rigorously evaluate the effectiveness of the proposed methodology, an additional test set was introduced. This test set included post-intervention data from the latter PwMS subset and an additional subset comprising of 14 completely unseen HC individuals. The complete demographics details, including the clinician-assigned EDSS scores are summarised in Table 3.1.

Table 3.1: Demographics table for gait anomaly detection.

Training and Validation Set

	Age	**Gender**	**MS Subtypes**			**EDSS**	**Walking assistive devices**		
	Mean (SD)	*N male*	*PP*	*RR*	*SP*	*Mean*	*None*	*Unilateral*	*Bilateral*
HC (n = 24)	49.9 (8.3)	8	-	-	-	-	24	0	0
MS (n = 72)	50.5 (12.1)	29	3	33	36	4.68	53	9	10

Test Set

	Age	**Gender**	**MS Subtypes**			**EDSS**	**Walking assistive devices**		
	Mean (SD)	*N male*	*PP*	*RR*	*SP*	*Mean*	*None*	*Unilateral*	*Bilateral*
HC (n = 14)	27.4 (3.7)	8	-	-	-	-	14	0	0
MS (n = 42)	46.2 (12.7)	18	3	33	6	4.19	31	6	5

PP = primary progressive, RR = relapse remitting, SP = secondary progressive
Unilateral = one stick, Bilateral = 2 sticks, walker or rollator

3.2.2 Gait assessment and initial processing

The data was processed using MATLAB 2020b (MathWorks, Inc., Natick, MA, USA). Following re-alignment to a horizontal-vertical coordinate system [184], the raw lower back IMU acceleration data was filtered using a $10Hz$ cut-off, zero phase, low-pass Butterworth filter. Automatic removal of resting breaks and turns was conducted according to [43], ensuring that only segments of steady-state walking were retained for subsequent analysis. In this work, resting breaks were detected by assessing 2-second sliding windows of data and checking if more than 50% of the samples had both the norm of the lumbar IMU angular velocity and acceleration lower than $0.5rad/s$ and within $\pm 10\%$ of $9.81m/s^2$ respectively. Here, the turns at the end of straight-line walking bouts were automatically identified by searching for steep positive and negative gradients in the trunk rotation angle (derived from the integral of the lumbar IMU angular velocity around the vertical axis, and filtered with a $1.5Hz$ cut-off frequency low-pass Butterworth filter), with a threshold of 115°. The gait events were detected from the angular velocity around the ML axis of the shank-mounted sensors, following the peak detection procedure described in [185]. An illustration of turning detection and examples of the gait events landmarks can be seen in Figure 3.1. To mitigate the impact of acceleration and deceleration at the beginning and end of walking bouts, the data signals were trimmed to include only the data points between the first and last initial contacts of straight-line walking bouts. Additionally, for inclusion in the analysis, only those passes comprising a minimum of four strides were considered valid.

3.2.3 Feature set descriptions

For this work, 36 clinically-relevant features are extracted from the straight-line walking bouts, following the work of Angelini et al. [17]. Out of these 36 features, the first 23 describe temporal gait features and are computed following detection of the gait events. The remaining 13 features are extracted from the filtered lumbar acceleration signal and consisted of gait quality metrics computed in the time and frequency domains. For conciseness, only a brief description of the initial feature set is provided here, in Table 3.2. For the corresponding mathematical formulas, the reader is referred to section A of the Appendix of this thesis.

It should be noted that the features presented in Table 3.2 are computed over a single

straight-line walking bout. Since an individual can complete multiple walking-bouts throughout the assessment, multiple observations are generated per subject, with each observation being 36-dimensional. In the case of this work, averaging the pre-selected metrics across the entire walking test was deemed inappropriate, due to both clinical reasons, as well as practical considerations. Firstly, MS-gait may deteriorate throughout a standard walking test, as suggested by [25], due to accumulated fatigue. As such, averaging the pre-selected metrics across the entire walking test may mask this subtle degradation. Moreover, from a practical standpoint, this study is dealing with a limited number of participants, which can be comparable to the number of extracted features. The challenge of identifying outliers arises in this context, as data points may become coplanar [128]. This issue is a well-known challenge in the machine learning community, often referred to as the 'curse of dimensionality' [103]. Here, the stride, step, stance, and swing times are computed as the mean of the left and right lower limb respective times, whereas variability metrics are computed as the pooled standard deviation for both lower limbs. The gait speed was computed as the distance of the straight-line walking divided by the time duration recorded from the first to the last initial contacts within a pass. Moreover, the asymmetry features were defined as the absolute difference between the mean values for the left and right lower limbs [186], and as the natural logarithm of the absolute ratio between the minimum and maximum values of the two lower limbs [187].

Table 3.2: Initial feature set.

Feature no.	***Feature Description***	***Feature no.***	***Feature Description***
1	Stride time (s)	19	Stance time asymmetry (ms)
2	Step time (s)	20	Swing time asymmetry (ms)
3	Stance time (s)	21	Step time asymmetry ln (%)
4	Swing time (s)	22	Stance time asymmetry ln (%)
5	Single support time (s)	23	Swing time asymmetry ln (%)
6	Double support time (s)	24	RMS (m/s^2)
7	Swing phase (%)	25	RMS ratio AP (-)
8	Double support phase (%)	26	RMS ratio ML (-)
9	Gait speed (m/s)	27	RMS ratio V (-)
10	Stride time SD (ms)	28	Jerk (JK) (m/s^3)
11	Step time SD (ms)	29	Jerk ratio AP/V (-)
12	Stance time SD (ms)	30	Jerk ratio ML/V (-)
13	Swing time SD (ms)	31	Step Regularity $(Ad1)$ (-)
14	Stride time CV (ms)	32	Stride Regularity $(Ad2)$ (-)
15	Step time CV (ms)	33	Symmetry (-)
16	Stance time CV (ms)	34	Harmonic Ratio AP (-)
17	Swing time CV (ms)	35	Harmonic Ratio ML (-)
18	Step time asymmetry (ms)	36	Harmonic Ratio V (-)

SD = standard deviation, CV = coefficient of variation,
V = vertical direction, ML = medio-lateral direction, AP = anterior-posterior direction,
$(-)$ = Dimensionless quantity

Unlike [188], the initial feature set proposed in [17] did not include any summary statistics measures computed from the acceleration of gyroscopic signals, which might also have the potential of highlighting clinically relevant discrepancies between HCs and MS-affected individuals. Instead, the initial feature set was derived based on other conceptual gait models proposed in the literature for several other pathological populations. For instance, summary statistical metrics computed from IMU acceleration data has been successfully used in [189, 190] for detection of individuals prone to experiencing falls among older adults, while [191] followed a similar approach for classifying subjects into pathological groups. Moreover, [192] revealed that acceleration summary statistics from the lower back sensor are useful features, allowing to distinguish PD subjects from HCs. The study observed significant statistical differences between groups for the variability statistics in all three anatomical directions. In view of these results, this study explores the hypothesis that augmenting this feature set may improve sensitivity to subtle gait anomalies. As such, the work contained in the upcoming analysis examines whether incorporating additional summary statistical features enhances the detection of MS-induced gait abnormalities.

Based on their successful integration in the above-mentioned studies, summary statistics features (the mean, standard deviation, minimum, maximum and range) extracted from the lower back and shank acceleration and gyroscopic signals across all three anatomical directions are used here to augment the initial feature set proposed by [17]. Moreover, additional measures of periodicity such as the autocorrelation coefficient and time lag of the acceleration signals from the shank sensors were also included. Finally, the augmented feature set also includes mean values and variances for stride, step, stance, and swing times for each lower limb, in addition to the total number of steps and strides taken during the assessment period. Following the creation of the augmented feature set, the statistical variability between the control and MS populations was assessed using the non-parametric Mann-Whitney U test [193], with a minimum significance alpha level of 5%. Additionally, Bonferroni correction was applied to account for multiple comparisons. Through this procedure, 11 features were identified and subsequently eliminated, as they did not exhibit statistically significant differences between MS subjects and HC. The final augmented feature set is provided in Table 3.4, whereas the excluded features are listed in Table 3.3.

Table 3.3: Excluded metrics.

Feature no.	***Feature Description***
1	Mean V Angular Velocity Left Shank (rad/s)
2	Mean ML Angular Velocity Left Shank (rad/s)
3	Range V Acceleration Right Shank (m/s^2)
4	Mean V Angular Velocity Right Shank (rad/s)
5	SD AP Angular Velocity Right Shank (rad/s)
6	Variance AP Angular Velocity Right Shank (rad^2/s^2)
7	Max AP Angular Velocity Right Shank (rad/s)
8	Range V Angular Velocity Right Shank (rad/s)
9	Range AP Angular Velocity Right Shank (rad/s)
10	Max ML Acceleration Back (m/s^2)
11	Harmonic Ratio ML (-)

Table 3.4: Augmented feature set.

Feature no.	*Feature Description*	*Feature no.*	*Feature Description*	*Feature no.*	*Feature Description*	Feature no.	*Feature Description*
1	Mean V Acceleration Left Shank (m/s^2)	**39**	SD ML Angular Velocity Left Shank (rad/s)	**77**	Autocorrelation Time Lag R Acceleration Right Shank (s)	**115**	Range ML Acceleration Back (m/s^2)
2	Mean ML Acceleration Left Shank (m/s^2)	**40**	SD AP Angular Velocity Left Shank (rad/s)	**78**	Mean Stride Time Right Shank (s)	**116**	Range AP Acceleration Back (m/s^2)
3	Mean AP Acceleration Left Shank (m/s^2)	**41**	Variance V Angular Velocity Left Shank (rad^2/s^2)	**79**	Mean Step Time Right Shank (s)	**117**	Stride Time (s)
4	SD V Acceleration Left Shank (m/s^2)	**42**	Variance ML Angular Velocity Left Shank (rad^2/s^2)	**80**	Mean Stance Time Right Shank (s)	**118**	Step Time (s)
5	SD ML Acceleration Left Shank (m/s^2)	**43**	Variance AP Angular Velocity Left Shank (rad^2/s^2)	**81**	Mean Swing Time Right Shank (s)	**119**	Stance Time (s)
6	SD AP Acceleration Left Shank (m/s^2)	**44**	Max V Angular Velocity Left Shank (rad/s)	**82**	Variance Stride Time Right Shank (s^2)	**120**	Swing Time (s)
7	Variance V Acceleration Left Shank (m^2/s^4)	**45**	Max ML Angular Velocity Left Shank (rad/s)	**83**	Variance Step Time Right Shank (s^2)	**121**	Single Support Time (s)
8	Variance ML Acceleration Left Shank (m^2/s^4)	**46**	Max AP Angular Velocity Left Shank (rad/s)	**84**	Variance Stance Time Right Shank (s^2)	**122**	Double Support Time (s)
9	Variance AP Acceleration Left Shank (m^2/s^4)	**47**	Min V Angular Velocity Left Shank (rad/s)	**85**	Variance Swing Time Right Shank (s^2)	**123**	Swing Phase (%)
10	Max V Acceleration Left Shank (m/s^2)	**48**	Min ML Angular Velocity Left Shank (rad/s)	**86**	No Strides Right Shank (-)	**124**	Double Support Phase (%)
11	Max ML Acceleration Left Shank (m/s^2)	**49**	Min AP Angular Velocity Left Shank (rad/s)	**87**	No Steps Right Shank (-)	**125**	Gait Speed (m/s)
12	Max AP Acceleration Left Shank (m/s^2)	**50**	Range V Angular Velocity Left Shank (rad/s)	**88**	Mean ML Angular Velocity Right Shank (rad/s)	**126**	Stride Time SD (ms)
13	Min V Acceleration Left Shank (m/s^2)	**51**	Range ML Angular Velocity Left Shank (rad/s)	**89**	Mean AP Angular Velocity Right Shank (rad/s)	**127**	Step Time SD (ms)
14	Min ML Acceleration Left Shank (m/s^2)	**52**	Range AP Angular Velocity Left Shank (rad/s)	**90**	SD V Angular Velocity Right Shank (rad/s)	**128**	Stance Time SD (ms)
15	Min AP Acceleration Left Shank (m/s^2)	**53**	Mean V Acceleration Right Shank (m/s^2)	**91**	SD ML Angular Velocity Right Shank (rad/s)	**129**	Swing Time SD (ms)
16	Range V Acceleration Left Shank (m/s^2)	**54**	Mean ML Acceleration Right Shank (m/s^2)	**92**	Variance V Angular Velocity Right Shank (rad^2/s^2)	**130**	Stride Time CV (ms)
17	Range ML Acceleration Left Shank (m/s^2)	**55**	Mean AP Acceleration Right Shank (m/s^2)	**93**	Variance ML Angular Velocity Right Shank (rad^2/s^2)	**131**	Step Time CV (ms)
18	Range AP Acceleration Left Shank (m/s^2)	**56**	SD V Acceleration Right Shank (m/s^2)	**94**	Max V Angular Velocity Right Shank (rad/s)	**132**	Stance Time CV (ms)
19	Autocorrelation Coefficient V Acceleration Left Shank (-)	**57**	SD ML Acceleration Right Shank (m/s^2)	**95**	Max ML Angular Velocity Right Shank (rad/s)	**133**	Swing Time CV (ms)
20	Autocorrelation Coefficient ML Acceleration Left Shank (-)	**58**	SD AP Acceleration Right Shank (m/s^2)	**96**	Min V Angular Velocity Right Shank (rad/s)	**134**	Step Time Asymmetry (ms)
21	Autocorrelation Coefficient AP Acceleration Left Shank (-)	**59**	Variance V Acceleration Right Shank (m^2/s^4)	**97**	Min ML Angular Velocity Right Shank (rad/s)	**135**	Stance Time Asymmetry (ms)
22	Autocorrelation Coefficient R Acceleration Left Shank (-)	**60**	Variance ML Acceleration Right Shank (m^2/s^4)	**98**	Min AP Angular Velocity Right Shank (rad/s)	**136**	Swing Time Asymmetry (ms)
23	Autocorrelation Time Lag V Acceleration Left Shank (s)	**61**	Variance AP Acceleration Right Shank (m^2/s^4)	**99**	Range ML Angular Velocity Right Shank (rad/s)	**137**	Step Time Asymmetry ln (%)
24	Autocorrelation Time Lag ML Acceleration Left Shank (s)	**62**	Max V Acceleration Right Shank (m/s^2)	**100**	Mean V Acceleration Back (m/s^2)	**138**	Stance Time Asymmetry ln (%)
25	Autocorrelation Time Lag AP Acceleration Left Shank (s)	**63**	Max ML Acceleration Right Shank (m/s^2)	**101**	Mean ML Acceleration Back (m/s^2)	**139**	Swing Time Asymmetry ln (%)
26	Autocorrelation Time Lag R Acceleration Left Shank (s)	**64**	Max AP Acceleration Right Shank (m/s^2)	**102**	Mean AP Acceleration Back (m/s^2)	**140**	RMS (m/s^2)
27	Mean Stride Time Left Shank (s)	**65**	Min V Acceleration Right Shank (m/s^2)	**103**	SD V Acceleration Back (m/s^2)	**141**	RMS Ratio AP (-)
28	Mean Step Time Left Shank (s)	**66**	Min ML Acceleration Right Shank (m/s^2)	**104**	SD ML Acceleration Back (m/s^2)	**142**	RMS Ratio ML (-)
29	Mean Stance Time Left Shank (s)	**67**	Min AP Acceleration Right Shank (m/s^2)	**105**	SD AP Acceleration Back (m/s^2)	**143**	RMS Ratio V (-)
30	Mean Swing Time Left Shank (s)	**68**	Range ML Acceleration Right Shank (m/s^2)	**106**	Variance V Acceleration Back (m^2/s^4)	**144**	Jerk (JK) (m/s^3)
31	Variance Stride Time Left Shank (s^2)	**69**	Range AP Acceleration Right Shank (m/s^2)	**107**	Variance ML Acceleration Back (m^2/s^4)	**145**	Jerk Ratio AP/V (-)
32	Variance Step Time Left Shank (s^2)	**70**	Autocorrelation Coefficient V Acceleration Right Shank (-)	**108**	Variance AP Acceleration Back (m^2/s^4)	**146**	Jerk Ratio ML/V (-)
33	Variance Stance Time Left Shank (s^2)	**71**	Autocorrelation Coefficient ML Acceleration Right Shank (-)	**109**	Max V Acceleration Back (m/s^2)	**147**	Step Regularity (Ad1) (-)
34	Variance Swing Time Left Shank (s^2)	**72**	Autocorrelation Coefficient AP Acceleration Right Shank (-)	**110**	Max AP Acceleration Back (m/s^2)	**148**	Stride Regularity (Ad2) (-)
35	No Strides Left Shank (-)	**73**	Autocorrelation Coefficient R Acceleration Right Shank (-)	**111**	Min V Acceleration Back (m/s^2)	**149**	Symmetry (-)
36	No Steps Left Shank (-)	**74**	Autocorrelation Time Lag V Acceleration Right Shank (s)	**112**	Min ML Acceleration Back (m/s^2)	**150**	Harmonic Ratio AP (-)
37	Mean AP Angular Velocity Left Shank (rad/s)	**75**	Autocorrelation Time Lag ML Acceleration Right Shank (s)	**113**	Min AP Acceleration Back (m/s^2)	**151**	Harmonic Ratio V (-)
38	SD V Angular Velocity Left Shank (rad/s)	**76**	Autocorrelation Time Lag AP Acceleration Right Shank (s)	**114**	Range V Acceleration Back (m/s^2)		

SD = standard deviation, CV = coefficient of variation, V = vertical direction, ML = medio-lateral direction, AP = anterior-posterior direction, (−) = Dimensionless quantity

3.3 Results and discussions

After establishing the technical details of the MSD-based gait anomaly detector and introducing both an initial feature set, as proposed in [17], together with an augmented feature set, the first half of this section will focus on detecting gait anomalies using the initial feature set. Subsequently, the latter half will explore the impact of incorporating additional summary statistical features on the detection of MS-induced gait anomalies. Because these latter features can also hold important information describing the health status of a particular individual, it is expected that incorporating additional summary metrics would positively impact the sensitivity of the gait anomaly detector proposed here. To reiterate, here, it is expected that the relative MSD will surpass the objective threshold when an individual exhibits an impaired gait pattern, with this threshold being determined based on the observed healthy condition data. In addition, it should also be noted that prior to computing the MSD, all feature sets underwent a normalisation procedure, ensuring a zero-mean and unit-standard deviation, with the training set (containing only HC individuals) serving as a reference baseline. This procedural step was deemed essential to mitigate potential scaling discrepancies that could arise when computing the covariance matrix in the MSD formulation. To aid visualisation, the PwMS were sorted in ascending order of increased MS severity levels. The MSD was computed among all straight-line walking bouts, across all individuals and was compared against the objective threshold. In this work, the threshold was computed according to the procedure described in Section 3.1.2, using a 99% confidence interval.

3.3.1 Outlier detection using the initial feature set

This section commences with the examination of the gait anomaly detection outcomes using the SFS algorithm, as depicted in Figure 3.3. On the left side of the figure, the results using the classical MSD computation are displayed, while the outcomes using the robust MSD, utilising the MCD estimator are shown on the right. The upper figures illustrate the comparison between MSD values and the objective threshold using the optimal feature set minimising the MCE, as determined by the SFS algorithm. Concurrently, the lower figures present the MCE observed during the iterative process of the SFS algorithm. The training set contained 576 samples, while the validation set contained 1535 samples. Here, datapoints corresponding to the MS

individuals have been coloured according to the MS severity level, with colder hues signifying a less severe condition, whereas the hotter hues are denoting increased levels of mobility impairment. The number of observations for the MCD subset has been set following the recommendations of [182], where the value of α was set to 0.5, thus, the number of observation whose classical covariance matrix has the lowest possible determinant was set equal to 288. These recommendations were deemed suitable, as the ratio between the cardinality of the training set and the number of features is higher than 5. The thresholds were set at 69.7 and 56.9 for the classical MSD approach and the robust one, respectively.

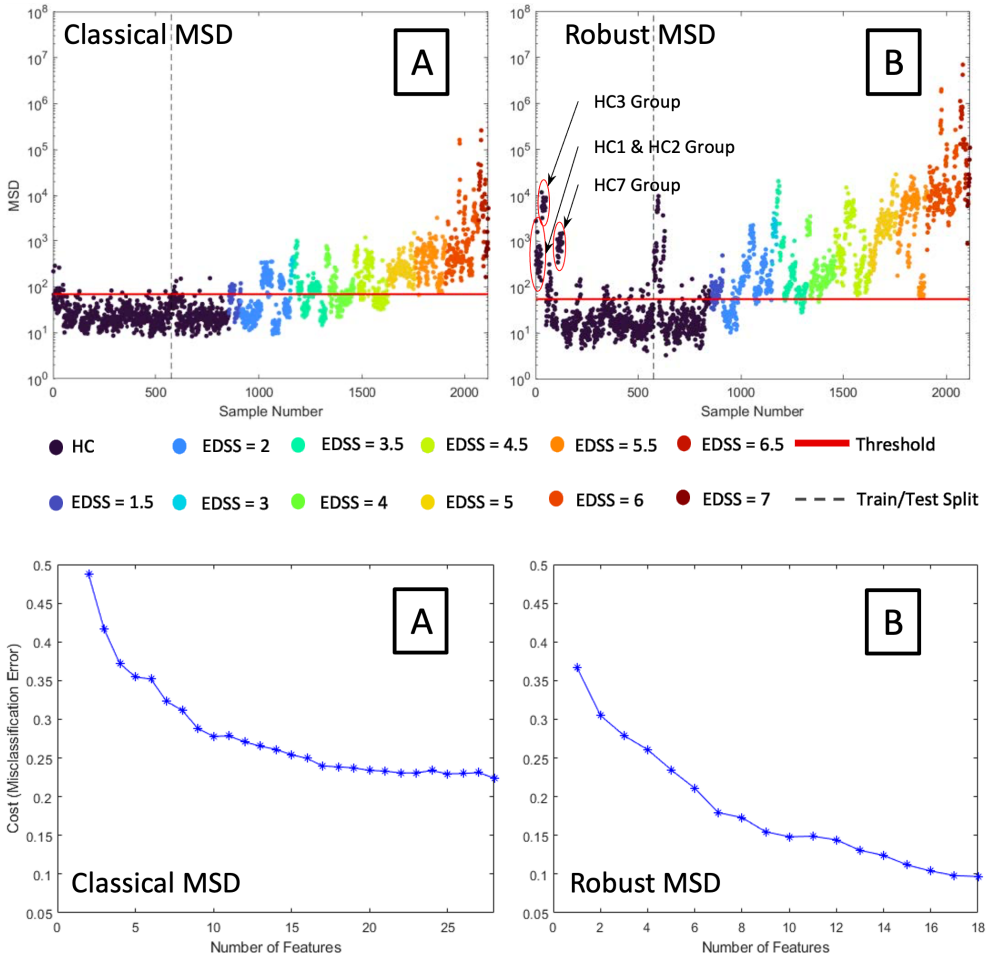


Figure 3.3: Summary of the gait anomaly detection results using the initial feature set, as derived by [17]. (A) - classical MSD computation, (B) - robust MSD computation, using the MCD estimator. The top figures represent the MSD comparison against the objective threshold using the optimal feature set found through the SFS algorithm. These features correspond to those that minimize the misclassification error (MCE). The bottom figures show the MCE obtained during the iterative process of the SFS algorithm.

By analysing Figure 3.3, three main observations are immediately evident. The first observation suggests that the underlying data distribution may not conform to a Gaussian distribution. Despite the assumption of a Gaussian distribution for the training data by MSD-based anomaly detector, the approach employed here still appears effective in distinguishing between HCs and MS classes. This effectiveness is attributed to the MS data points being significantly distanced from those representing healthy conditions. Moreover, the second observation is that by using the robust MSD formulation, the anomaly detector now offers a trade-off between sensitivity and specificity. Thirdly, it is evident that the robust approach immediately reveals the inclusive outliers present in the HC dataset. As depicted in Figure 3.3B, four groups of inclusive outliers are highlighted by the red ellipses and are labelled HC1, HC2, HC3 and HC7. To clarify, each label corresponds to a single HC individual, and each datapoint corresponds to the MSD value computed across a single straight-line walking bout. The additional groups of outlying datapoints detected in the validation set on the right-hand side of the train/test split line correspond to the same individuals identified as outlying in the train set (i.e., HC1, HC2, HC3 and HC7). The reason these HC individuals were deemed as outlying was not readily apparent. However, on closer inspection, it was found that the temporal metrics corresponding to these subjects exhibit dissimilarities relative to the rest of the HCs, due to misclassification of the heel strike event during gait event detection. Consequently, this result underscores the necessity for employing robust methodologies that inherently handle the presence of inclusive outliers within the training data. Here, without employing the MCD-based MSD computation, these outliers may not have been detected.

In addition, it is also observed that detection of gait anomalies in MS patients with EDSS scores below 5 remains challenging using the classical MSD approach, as seen in Figure 3.3A - top. Conversely, the robust MSD approach demonstrates improved gait anomaly detection, as quantified by the increase the recall metrics from 0.771 to 0.936. Expectedly, the specificity value has decreased from 0.971 to 0.773, as a result of the inclusive outliers being now detected in the test set. The comparison of the classification performance for the optimal feature sets of both the classical and robust approaches is presented in Table 3.5. Except for a small number of observations, the robust approach correctly identifies MS-affected individuals with EDSS scores higher than 3. However, the slightest gait anomalies of mildly affected PwMS with EDSS scores of 2 remain undetectable. Comparing the classification outcomes, the classical MSD approach achieved a minimum MCE of 0.22 with 28 features, while the robust

MSD method attained a minimum MCE of 0.09 with 18 features. This highlights the effectiveness of robust methodologies addressing inclusive outliers and mitigating their detrimental impact on outlier detection.

Table 3.5: Classification performance of the two approaches using SFS.

Metric Description	***Metric Value***	
	A - Classical MSD	***B - Robust MSD***
Accuracy	0.758	0.906
Precision	0.991	0.949
Recall	0.711	0.936
Specificity	0.971	0.773
F1	0.828	0.942

Table 3.6 exclusively presents the optimal feature set obtained through this approach. The reduction in the feature set holds a particular significance in this context, as it equips end-users with a more streamlined feature set, thereby enhancing assessment outcomes and interpretability. Here, it is observed that only several temporal metrics were included in the reduced set, thus decreasing the covariance imposed by the highly correlated temporal features, which may compromise clinical acceptance. Only the step time coefficient of variation was included among the gait variability metrics. Its inclusion is further supported by [194]. The same behaviour was observed for the regularity measures, as only the step regularity metric (Ad1) was included. The smoothness metrics represented by the harmonic ratios are not included. This result is supported by the findings of [17, 31], which also found the gait smoothness metrics to be irrelevant for the detection of the slightest gait anomalies in MS. Another notable remark is the fact RMS ratio along the ML direction, (which is a measure of gait instability) was not included in the reduced feature set. This is result is counterintuitive, as Sekine et al. [67] emphasised that the RMS ratio along the ML direction is a potential quantitative feature for measuring gait abnormality. Interestingly, when the reduced feature set is used, one MS patient with EDSS score of 5.5 appears to be closer to the healthy condition. This result is exhibited as a result of not including temporal variability metrics (which were the most discriminative features for this subject) in the reduced feature set and highlights the uniqueness of individual gait patterns, reinforcing the importance of selecting pathology-relevant features to enhance interpretability.

Table 3.6: Optimal feature set for the robust MSD using SFS.

Feature no.	*Feature Description*	*Feature no.*	*Feature Description*
1	Gait speed (m/s)	10	Step time asymmetry ln (%)
2	Stance time (s)	11	Step Time CV (s)
3	Step regularity (Ad1) (-)	12	Jerk ratio ML/V (-)
4	RMS Ratio AP (-)	13	Jerk ratio AP/V (-)
5	RMS (m/s^2)	14	Swing phase (%)
6	Swing time asymmetry (ms)	15	Single support time
7	Symmetry (-)	16	Double support phase (%)
8	RMS ratio V (-)	17	Double support time (s)
9	Step time asymmetry (ms)	18	Jerk (m/s^3)

Despite employing the robust MCD approach, the detection of the slightest gait anomalies in mildly affected MS individuals remains challenging. Therefore, the following analysis will demonstrate how additional metrics can improve gait anomaly detection prospects, even for mildly affected MS individuals.

3.3.2 Outlier detection using the augmented feature set

The subsequent results, obtained using the augmented feature set, exclude the outlying HC individuals identified in Section 3.3.1, namely, HC1, HC2, HC3 and HC7. To clarify, although the robust MSD using the MCD estimator safeguards against the presence of inclusive outliers, their early detection allows for their removal from the dataset, facilitating the use of classical MSD for gait anomaly detection. The outlying data points from the HC individuals comprised 183 samples. Consequently, their removal reduced the total number of observations in the training and validation set to 1973. While the previous section illustrated the initial feature set's capability to identify gait impairments in MS-affected individuals, the challenge of detecting slight gait anomalies among mildly affected subjects has also been highlighted. Hence, the subsequent results offer insight into whether enhancing the initial feature set enhances the sensitivity of the MSD-based detection algorithm.

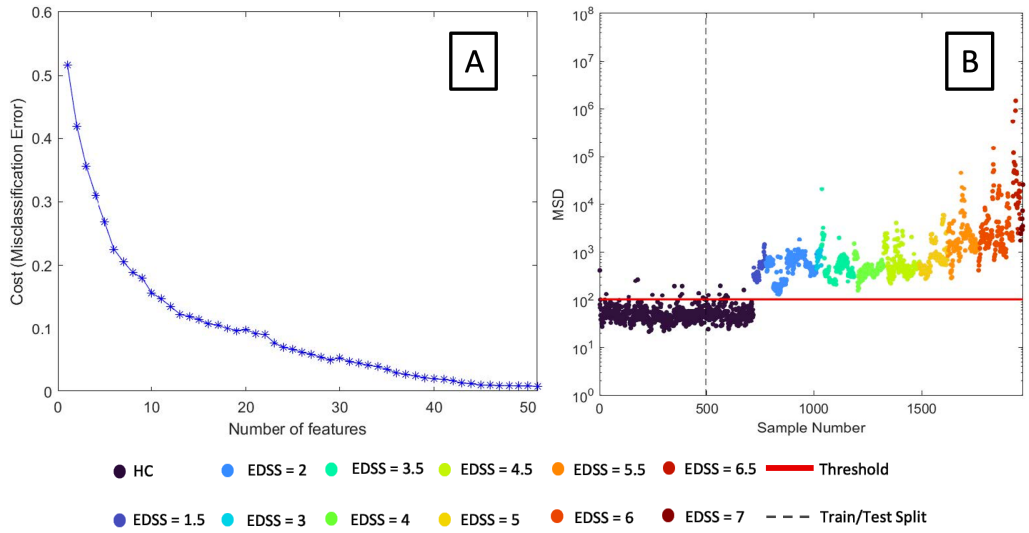


Figure 3.4: Summary of the gait anomaly detection results using the augmented feature set. (A) - History of the MCE obtained during the iterative process of the SFS (B) - MSD comparison against the objective threshold using the optimal feature set found through the SFS algorithm. These features correspond to those that minimize the misclassification error (MCE)

The results of gait anomaly detection utilising the SFS algorithm on the augmented feature set are illustrated in Figure 3.4. The left plot presents the MCE observed during the iterative process of the SFS algorithm. It can be inferred that the minimum error was achieved using a combination of 51 features. The right plot shows the corresponding Mahalanobis distances, together with the objective threshold, which, in the case of this analysis was set to 100.1. The performance metrics achieved using the optimal features set are presented in Table 3.7.

Table 3.7: Performance metrics for the gait anomaly detector using the optimised feature set.

Metric Description	***Metric Value***
Accuracy	0.993
Precision	0.992
Recall	1.000
Specificity	0.954
F1	0.996

Examining Figure 3.4B reveals that all the MS datapoints are regarded as outliers, and even the slight gait alterations of the mildly affected MS subjects are detected.

This is an improvement over the detection prospects using the initial features set, where only the severely affected MS patients were consistently classified as outliers. As the EDSS score is increased, the distances from the threshold are gradually increasing too, which might indicate progressive worsening of the gait parameters in PwMS. Several HC observations can be seen to lay above the threshold. This behaviour is reflected in the specificity value of 0.954 recorded in Table 3.7. One might expect a specificity value of approximately 0.99, since the threshold was set up based on extreme values corresponding to a 99% confidence interval following the Monte-Carlo method presented in Section 3.1.2. However, despite the assumption that the threshold is computed based on normally distributed data, which may not accurately reflect the true distribution of the data, the specificity value in this instance is deemed acceptable. The features selected by the SFS algorithm are concisely listed below, in Table 3.8, according to the order in which they have been selected. Except for the gait speed, visually inspecting the feature space with the aid of 2-dimensional scatter plots did not show a clear separation between the two subject classes, let alone across different MS disability levels. Therefore, using a multivariate analysis allows the clear separation of the MS individuals from the HC population. In order words, the separation is only possible when the gait data is holistically analysed. For brevity, subsequent paragraphs will focus solely on discussing the initial 20 metrics of the optimal feature set.

Table 3.8: The optimised feature set.

Feature no.	Feature description	Feature no.	Feature Description
1	Gait Speed (m/s)	**27**	Variance Stride Time Right Shank (s^2)
2	Max. V Acceleration Right Shank (m/s^2)	**28**	Variance Swing Time Right Shank (s^2)
3	Swing Time asymmetry (ms)	**29**	Variance ML Acceleration Back (m^2/s^4)
4	Mean AP Acceleration Back (m/s^2)	**30**	No. Strides Right (-)
5	Autocorrelation Coefficient V Acceleration Left Shank (-)	**31**	Range V Acceleration Back (m/s^2)
6	Max. V Angular Velocity Right Shank (rad/s)	**32**	Autocorrelation Time Lag ML Acceleration Left Shank (s)
7	No. Steps Left Shank (-)	**33**	Min. AP Angular Velocity Left Shank (rad/s)
8	Swing Time (s)	**34**	Max. V Acceleration Left Shank (m/s^2)
9	Step Time Asymmetry ln (%)	**35**	Range AP Acceleration Right Shank (m/s^2)
10	Jerk Ratio AP/V (-)	**36**	SD V Acceleration Back (m/s^2)
11	Max. ML Angular Velocity Left Shank (rad/s)	**37**	Symmetry (-)
12	Range ML Acceleration Left Shank (m/s^2)	**38**	Min. V Acceleration Right Shank (m/s^2)
13	Max. AP Acceleration Right Shank (m/s^2)	**39**	Harmonic Ratio AP (-)
14	Autocorrelation Coefficient AP Acceleration Left Leg (-)	**40**	Mean ML Acceleration Left Shank (m/s^2)
15	Max. ML Angular Velocity Right Shank (rad/s)	**41**	Autocorrelation Time Lag R Acceleration Right Shank (s)
16	RMS Ratio AP (-)	**42**	Mean Swing Time Right Shank (s)
17	Variance V Angular Velocity Right Shank (rad^2/s^2)	**43**	Range AP Acceleration Left Shank (m/s^2)
18	Max. ML Acceleration Left Shank (m/s^2)	**44**	Range ML Acceleration Back (m/s^2)
19	Autocorrelation Coefficient AP Acceleration Right Shank (-)	**45**	Mean V Acceleration Left Shank (m/s^2)
20	Step Regularity (Ad1) (-)	**46**	Range V Angular Velocity Left Shank (rad/s)
21	Min. ML Acceleration Left Shank (m/s^2)	**47**	Mean Step Time Left Shank (s)
22	Range ML Angular Velocity Left Shank (rad/s)	**48**	Autocorrelation Time Lag AP Acceleration Right Shank (s)
23	Min. ML Angular Velocity Right (rad/s)	**49**	SD Swing Time (s)
24	SD V Angular Velocity Left Shank (rad/s)	**50**	Variance AP Acceleration Back ($m2/s^4$)
25	Mean ML Acceleration Right Shank (m/s^2)	**51**	SD Step Time (s)
26	Autocorrelation Coefficient ML Acceleration Left Shank (-)		

V = vertical direction, ML = medio-lateral direction, AP = anterior-posterior direction, R = Resultant direction

While the gait speed is said to reflect changes in lower limb weakness and spasticity

in MS, and is the primary variable measured in clinician-administrated walking tests [195–197], when used standalone, it might not reflect the real world gait impairment for the mildly affected PwMS, as reported by [198]. In this work, the gait speed also emerged as one of the most significant features for assessing gait impairment. The maximum vertical acceleration recorded for the right shank was the second feature added to the optimal feature set. For this metric, a higher variability was noticed within the MS population, which might reflect uncontrolled lower limb movements, possibly during the swing phase, as a result of muscle weakness, spasticity, fatigue or balance impairments [159, 199]. The swing time asymmetry was ranked as the third most discriminative metric, demonstrating a noticeable increase with higher levels of MS severity. It was included in a number of studies involving pathological subjects [17, 43, 87, 200]. The mean AP acceleration computed from the lower back sensor, which was selected as the fourth metric, reflects the postural tendency of MS subjects to lean forward or backwards while walking. In contrast, the HC individuals tend to better control this movement. This phenomenon is evident at lower disability levels but intensifies with increasing disability. The fifth metric added to the feature set is the height of the first peak of the normalised autocorrelation signal in the V direction for the left shank. This metric has been identified as being useful for the identification of tremor in PD [201]. In this context, this feature reflects the non-random movements of the shank along the vertical direction. While HC individuals display coefficients closer to 1, indicative of nearly perfect periodicity, moderately and severely affected MS individual, display notably lower values for this metric. The maximum value of the angular velocity around the vertical axis for the right shank is selected as the sixth metric of the optimal feature set. While PwMS with EDSS scores higher than 6 had a more restricted shank movement, a visual inspection of 2-dimensional scatter plots revealed that this metric is particularly useful at discriminating between several MS patients with EDSS scores of 2 and 3.5. The number of steps are highly correlated with the gait speed and are traditionally extracted in clinical gait assessments using IMUs [20]. HC subjects typically require fewer steps to traverse a given distance of the corridor used for the walking test, whereas MS-affected individuals necessitate additional steps due to reduced walking speed. Notably, in this study, the number of steps recorded for the right shank was automatically excluded following the SFS procedure. The eighth feature extracted was the swing time. [202] revealed that the swing time is not statistically different for the MS group compared to healthy controls when the subjects walk at a preferred walking speed, as in the case of the 6MWT performed here. Differences only arise

when the walking speed in imposed. However, the swing time was useful in this case, as it revealed two additional outlier groups of PwMS with EDSS scores of 5.5 and 6.5. The ninth metric extracted was the natural logarithm of the step time asymmetry. Its inclusion into the optimal feature set is further supported by its appearance in a multifactorial model for MS proposed by Angelini et al. [17]. Similarly, the tenth and twentieth metrics are also included in the model. The maximum angular velocity around the ML direction for the left and right shanks (which were the eleventh and fifteenth metrics included in the optimised feature set) reveal a restricted range of motion for the MS population. Notably, an MS-affected individual with an EDSS score of 5.5 exhibited a distinct walking pattern solely on the right foot, contrasting with the absence of such pattern on the left foot. The twelfth feature selected by the SFS algorithm is the range of the ML acceleration for the left shank. However, the selection of this seems to be counterintuitive, as no evident separation between subjects is visible in a 2D space. Therefore, it might only be useful when combined with additional features in a multi-dimensional feature space. The same behaviour is observed for the maximum AP acceleration recorded with the IMU placed on the right shank. The fourteenth selected feature is the first peak of the normalised autocorrelation signal in the AP direction for the left shank, as a measure of periodicity. The same feature was recorded in the nineteenth position, this time for the right shank. The sixteenth feature added to the feature set was the ratio of the trunk acceleration RMS in the AP direction to the RMS vector magnitude, as a measure of AP stability, as described by [43]. However the inclusion of this metric contrasts the findings of [67], who emphasised that the RMS ratio in the ML direction is a potential quantitative feature for measuring gait abnormality, as opposed to the AP direction selected by the SFS algorithm. Inclusion of the angular velocity variance along the vertical axis of the IMU placed on the right shank revealed additional groups of PwMS outliers, with EDSS scores of 1.5, 2 and 4. Finally, the maximum acceleration in the ML direction for the left shank was selected as the eighteenth metric and further revealed outlying datapoints corresponding to mildly affected PwMS.

The discriminative power of the feature set presented in Table 3.8 was further validated on the additional test presented in Table 3.1. Here, the additional test set consisted of 404 datapoints from HCs and 897 datapoints from PwMS. The MSD comparison against the threshold value of 100.1 (as previously computed during the implementation of the SFS algorithm) can be seen in Figure 3.5. The corresponding classification performance metrics are provided in Table 3.9.

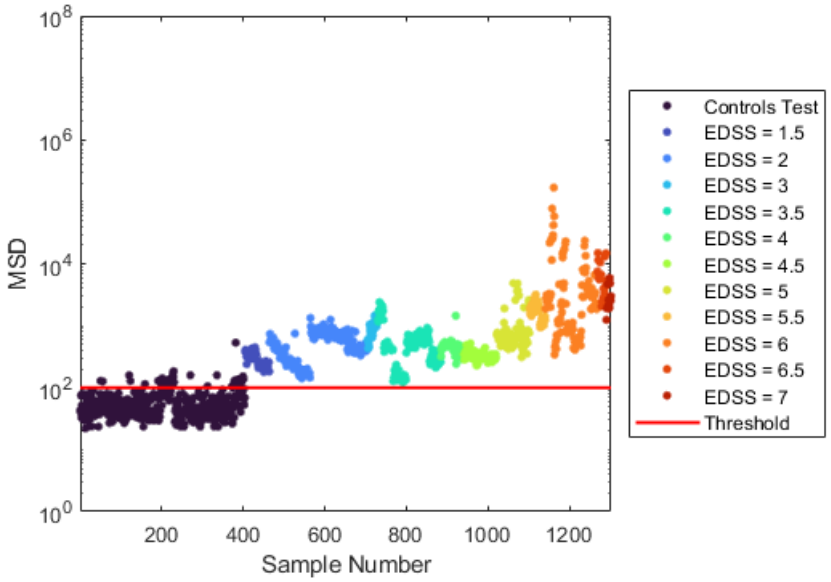


Figure 3.5: Gait anomaly detection results computed on the test set.

Finally, the performance metrics presented in Table 3.9 suggest that the proposed model generalizes well to new data, reinforcing its robustness and reliability. The high precision, recall and F1 score indicate minimal false positives and false negatives, emphasizing the model's effectiveness in accurately detecting gait abnormalities in MS-affected individuals.

Table 3.9: Performance metrics for the gait anomaly detector using the optimised feature set.

Metric Description	***Metric Value***
Accuracy	0.972
Precision	0.961
Recall	1.000
Specificity	0.911
F1	0.980

In summary, the results presented in this section illustrate that enhancing the detection of gait impairment in MS can be achieved by incorporating additional summary statistical metrics and rhythmicity measures alongside conventional features extracted by expert researchers in gait analysis field. The variability within the HC group was reduced in the second part of this analysis through the exclusion of inclusive outliers. In turn, this led to an increased MSD for PwMS relative to the HC distribution. By employing the SFS algorithm, feature redundancy in the optimised

feature set was minimized. This approach enhances the likelihood of detecting gait anomalies in MS-affected individuals, even those with mild impairment, statement supported through the recall value of 1.00.

3.3.3 Further discussion

This chapter focuses on the first level of the hierarchical framework proposed in Chapter 1 for assessing the condition of PwMS, aiming to detect gait impairment at an early stage. Here, a robust methodology for detection of gait anomalies has been introduced using the MSD as a statistical process control method. The key novelty of the methodology presented here consisted in the introduction of inclusive outlier detection methods, by employing the MCD estimator. As such, 'malignant' inclusive outliers, masking the true healthy condition have been identified and removed at an early stage, allowing the gait impairment detector to maintain sufficient sensitivity and generality for detection of subtle impairments. This stage represents the first step towards a robust data-driven gait assessment, aiming to augment clinical practice.

In this work, an initial feature set which consisted of gait metrics previously extracted by other experts in the field was initially considered. It was demonstrated that when combined with additional features, such as descriptive statistics, and periodicity indicators extracted from the acceleration signals with the aid of the autocorrelation function, the classification prospects can be marginally improved. The optimal feature set was derived using the SFS algorithm. This allows for detection of gait abnormalities, even in individuals with mild MS disability levels, and can lead to an improved assessment of gait impairments in MS.

The development of conceptual gait models was previously driven by clinical acceptance, due to a need to reduce the size of features indicative of a certain pathology and retain as much of the original information as possible [17]. Here, the initial feature set, which consisted of 36 metrics was not considered adequate to distinguish gait anomalies even at the lowest severity levels, which lead to the integration of additional metrics. The feature set augmentation procedure resulted in 151 variables, which were then reduced to 51 through the SFS algorithm.

The work contained in this chapter addresses a useful research question, namely whether someone has a gait impairment or not. This is because early detection of gait anomalies holds significant importance for MS-affected individuals, as it could

inform optimal treatment planning strategies. Considering the subjectivity of the EDSS scoring system, PwMS with scores below the threshold value of 3.5 seem to have no apparent gait impairment. However, this might only hold true when the gait assessment is performed without instrumentation (i.e. motion capture systems or IMUs). Furthermore, the scale is not linear and shows variable sensitivity to change across the disability spectrum [75]. Therefore, when objective tools are utilised, such as assessments using IMUs, alterations in gait are noticeable even for the mildly affected patients (EDSS < 3.5) [17, 31, 57, 203]. The results presented in this analysis further demonstrated that gait anomalies are detectable using an objective approach, even for PwMS with no clinically observable walking disability.

It can also be concluded that while the MS subjects seem to walk in a unique manner, the HC group seems to be less varied, particularly after accounting for inclusive outliers. Apart from gait speed, which emerges as a highly discriminative feature, no single metric alone demonstrates a clear differentiation between the two groups under investigation, nor across varying levels of MS disability. Here, it is perhaps unsurprising the walking speed was selected as the primary distinguishing feature, given that this metric is the most widely explored digital mobility outcome [204]. The results presented in this chapter suggest that the biomechanical control and coordination of gait is a multifaceted process, likely influenced by a combination of deficits rather than isolated mechanisms, as suggested by [194]. This complexity is reflected in the results presented above, where a clear differentiation between MS individuals and HC only emerges upon a comprehensive analysis of gait metrics.

The methodology used in this analysis for gait impairment detection is novel in the field of clinical gait analysis. A new approach for gait anomaly detection was proposed based on the calculation of the Mahalanobis squared distance, whereas robust improvements for early detection of inclusive outliers present in the HC training set were introduced though the computation of the minimum covariance determinant estimator. However, the idea of using outlier analysis for anomaly detection problems is not new, as it was previously introduced by others in the field of SHM [124, 125, 128]. Interpretability and traceability are the additional benefits of these methods proposed here, due to the relative simplicity of the algorithms used. For this reason, the methodology has the potential of being clinically adopted, although further validation is required.

Having presented the advantages and potential uses for the newly proposed methodology for gait anomaly detection, some thought must be also given to the possible limitations.

Firstly, it is perhaps worthy to acknowledge that the MS individuals incorporated in the test set are not entirely novel. Specifically, post-intervention data has been utilised from the participants enrolled in the intervention-based clinical trial, which were also present in the validation set. For clarifications, please refer to the dataset description provided in Section 3.2.1. Post intervention data has been used here in order to have access to wide spectrum of EDSS scores, which was not possible otherwise. Hence, further validation is warranted for the future work. In addition, some attention must be also paid to the marginally reduced specificity values observed in the test set, relative to the validation set. Nonetheless, this marginal reduction does not underscore significant concern here. Next, one might argue that the number of features included in the optimal feature set is relatively high and might impact clinical adoption, considering that most conceptual models proposed in the literature consist of a reduced number of features. As such, alternative regularisation approaches are also worthy of investigation.

3.4 Conclusions

The work presented in this chapter completed the first level of the hierarchical framework for evaluating the condition of a patient, focused on detection of gait anomalies in the gait of MS-affected individuals. Inspired by the SHM field, the Mahalanobis squared distance (MSD) has been used as novel distance metric that not only allows for detection of the MS impaired gait, but also suggests how the dimensionality of the dataset can be reduced without losing efficiency. Furthermore, a robust implementation of the Mahalanobis distance has been facilitated by the adoption of the minimum covariance determinant estimator. Consequently, utilising robust statistical measures led to the early identification and removal of inclusive outlying data points contaminating the healthy control dataset, thereby enhancing detection prospects.

Using the MSD-based anomaly detector, an optimal feature set containing 51 metrics was found through the implementation of the sequential forward selection algorithm. In this context, feature selection ensures the exclusion of irrelevant features that may impede clinical interpretation. This optimal feature set derived here encompasses not only traditionally extracted gait metrics, but also descriptive statistics and periodicity measures extracted from the IMU gait signals. Notably, aside from gait

speed, which is recognized as one of the most discriminative features, no individual gait feature exhibited a clear differentiation between HC and MS, nor across various levels of MS disability. This result highlights the complex nature of human gait and emphasizes that distinction between the MS population and healthy controls requires a comprehensive multivariate exploration of gait data.

The novel methodology proposed in this chapter has the potential to detect even the slightest gait anomalies of the mildly affected MS patients, who often lack clinically observable walking disabilities. This approach not only enhances the detection of gait anomalies in MS but also showcases the potential of novel machine-learning approaches as complementary tools to support clinical practice.

Having detected MS-related gait deficits, in the form of gait impairment, it is then necessary to delve deeper into the assessment and accurately evaluate the severity of the condition and monitor the progression of the disease. This chapter primarily addressed a binary classification problem. While it is evident that individuals with a severe condition deviate significantly from the healthy distribution, the assessment of disease severity for those with mild to moderate symptoms remains somewhat ambiguous due to the considerable overlap. As such, building on this starting point, the next chapter will investigate the severity quantification problem in more detail.

Chapter 4

Contrastive learning approaches for MS severity assessment

This chapter delves into the nuanced task of assessing the severity of Multiple Sclerosis (MS), focusing specifically on the third level of the hierarchical system proposed for assessing an individual's condition. As a step forward towards the overarching goal of real-life gait monitoring, the work presented in this chapter attempts this task using a single wearable sensor. However, the data captured using this single sensor approach demands a heightened appreciation for complexity. The intricate biomechanical characteristics captured by this sensor surpass the simplicity of the data acquired using sensors placed on the lower body segments. Consequently, traditional handcrafted feature extraction methods fall short in fully capturing the richness of this data. To this end, this chapter adopts a novel methodology for assessing severity of the disease using neural networks and contrastive learning - a method that learns optimal data representations by maximizing agreement between gait patterns from the same severity class while discriminating between patterns from different severity classes, enabling automatic discovery of intrinsic gait characteristics without relying on predefined feature engineering. Rather than imposing explicit criteria for pattern discrimination, this approach allows the underlying relationships to emerge naturally from the data structure. The gap between computational predictions and clinical understanding is addressed through the Layer-wise Relevance Propagation (LRP) technique, which decomposes the network's output by propagating relevance scores backwards through its layers, quantifying the contribution of each temporal

region of the input gait signals to the final severity class prediction. This systematic attribution of relevance not only provides a more nuanced understanding of the underlying biomechanical characteristics that inform severity classification, but also establishes a robust methodology for accurate MS severity assessment.

4.1 Introduction

In light of the overarching goal of remote gait analysis, this chapter presents a novel approach for quantifying MS disability levels using a single IMU sensor positioned on the lower back. This is because current mobility endpoints, which are based on functional performance, physical assessments or patient self-reported outcomes offer restricted sensitivity. This constraint diminishes their utility in a clinical setting [9]. In contrast, IMUs can overcome these constraints in both supervised structured assessments and real-world scenarios. In a trade-off between usability and accuracy, a single wearable device situated on the lower back—an ergonomically convenient position near the centre of mass—has been well received by participants [205, 206] and has been thoroughly validated for the output digital mobility outcomes in pathological populations, including MS [9, 79, 204, 207]. Moreover, this location can provide insightful perspectives into the dynamic balance, stability and smoothness of gait, aspects which are often regarded as hallmarks of MS [31, 99]. Building on this, recently, Creagh et al. [16] have provided valuable insights into the fluctuations of disability levels among individuals affected by MS over a 24-week period of continuous remote monitoring, using accelerometer data from a waist-worn smartphone, further validating the utility of a single sensor monitoring approach.

To this end, IMU-based summary features can be extracted from the data acquired using a single sensor approach, facilitating the development of machine-learning based models capable of distinguishing MS disability from healthy participants [18]. However, this approach is predicated on prior assumptions and often fails short to fully capture the complexities and richness of the lower-back IMU data. In this chapter, it is proposed that instead of relying on 'expert' selection of representative features, it may be more effective to allow an algorithm to learn its own features, a process known as *representation learning*. In this regard, neural networks have demonstrated significant success in numerous time-series related tasks, such as activity recognition [208], person identification [136, 139] or even remote characterisation of

MS of ambulation [16].

Although [16] has demonstrated that it is possible to effectively stratify MS subjects across disability levels, their study only included mildly and moderately affected individuals, which is not representative of the full spectrum of MS disability levels. To address this limitation and provide a more comprehensive representation, the study presented here also includes individuals with severe MS disability, thereby offering an enhanced quantification of the full spectrum of MS disability. This effectively extends the gait anomaly detection task presented in the previous chapter to a four-class classification problem. However, given the high heterogeneity of MS as a disease [14] and, considering the findings presented in Chapter 3, where a clear overlap across disability levels was observed, it is expected that the task of accurately quantifying MS severity will pose a significant challenge.

To overcome these challenges, this chapter introduces *self-supervised learning* as a viable solution to this problem. This learning paradigm uses *contrastive learning*, which acts as a *pretext task* (i.e., an interim objective preceding final classification), aiming to obtain a model which can produce a good representation of the IMU data in a latent space. Although the formal introduction is postponed until Section 4.4, briefly, a contrastive learning framework aims to cluster together examples exhibiting semantic similarity. Over the recent years, this approach has emerged as a promising technique, attaining excellent performance with instance discrimination serving as its pretext task. Notably, it has demonstrated the capability to surpass *supervised learning* alternatives in subsequent classification problems, particularly in terms of accuracy [143, 209]. The key idea behind contrastive learning approaches is that similar instances (presumably belonging to the same class) are grouped together in the latent space, whereas dissimilar instances are scattered further apart. In the case of this work, the *triplet-loss* framework is presented as a contrastive learning paradigm, which incorporates triplet instances [141]. A triplet consists of an *anchor*, a *positive* instance (of the same class as the anchor) and a *negative* instance (dissimilar to the anchor). Within this framework, the objective is to maximise the distance between the anchor and the negative pair, while simultaneously minimising the distance between the anchor and the positive pair using a hard margin parameter. Having successfully obtained a good representation of the data in the latent space, the final severity classification task will be significantly simplified.

In addition, it is also well known that neural networks, and especially deep neural networks, can exhibit high non-linearity and complexity. This often creates an

inherent challenge in interpreting the decisions that contribute to a prediction [136, 210]. As such, a number of techniques have been developed over the years to explain neural network's decisions [146, 149–156]. In the context in this work, inspired by its success in the work of Horst et al. [136], the layer-wise relevance propagation (LRP) technique [154] emerges as a suitable approach for attributing, explaining, and interpreting the lower back sensor patterns that are relevant for accurately assessing the severity of the disease.

4.2 Neural networks

In the recent years, neural networks have evolved to become, by far, the most important machine learning technology for practical use cases [102]. Briefly, neural networks, particularly multi-layer ones, are comprised of interconnected elements known as nodes, organised in layers. These networks function by transmitting signals from the input layers through intermediate hidden layers, ultimately culminating in the output layers. This hierarchical structure allows neural networks to process complex information, extract meaningful features, and make predictions or classifications based on the learnt patterns. The fundamental concept behind neural networks is to find a set of nonlinear transformations of the inputs, facilitated by the use of basis functions, denoted as $\phi_j(x)$, which possess parameters that can be learnt, and subsequently adjusting these parameters alongside coefficients w_j during the training process. Then, the entire model is optimised by minimising an error function through gradient-based optimisation techniques. An example of such a technique is represented by the stochastic gradient descend method, wherein the error function is jointly defined across all model parameters [211].

In order to construct nonlinear basis functions, one essential requirement is their differentiability in relation to their learnable parameters, in order to facilitate gradient-based optimisation. Adopting basis functions that resemble a specific form has proven particularly effective, meaning that they can facilitate identification of complex patterns in the data, leading to better model performance [102]. In this approach, each basis function represents a nonlinear transformation of a linear combination of input variables, with the coefficients therein being learnable parameters. Notably, this construction can be recursively extended to yield a hierarchical model with many layers, forming the foundation of deep neural networks.

To illustrate the mathematical representation of neural networks, a simple case of a neural network model with two layers is considered here. Initially, a series of M linear combinations of the input variables $x_1, x_2, \ldots, x_D$ is constructed in the following form:

$$a_j^{(1)} = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{4.1}$$

where $j = 1, 2, \ldots, M$, and the superscript (1) denotes the parameters corresponding to the first layer of the network. The parameters $w_{ji}^{(1)}$ are referred to as *weights*, while $w_{j0}^{(1)}$ are identified as the *biases*. The quantities $a_j^{(1)}$ are termed *pre-activations*. Subsequently, each a_j undergoes a transformation using a nonlinear differentiable *activation function*, $h(\cdot)$, as seen in Equation 4.2. Here, these basis function transformations are referred to as *hidden units* with examples of typical nonlinear activation functions $h(\cdot)$ being provided in Figure 4.2.

$$z_j^{(1)} = h(a_j^{(1)}), \tag{4.2}$$

Notably, provided that $h'(\cdot)$ (the derivative of $h(\cdot)$) can be evaluated, then, the overall network function also becomes fully differentiable. Consequently, it follows

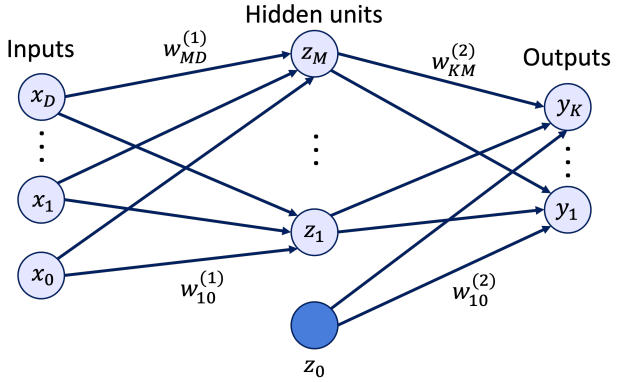


Figure 4.1: Two-layer neural network diagram. Nodes represent the input, hidden, and output variables, while the connections between these nodes signify the weight parameters. The bias parameters are represented by links originating from additional input and hidden variables x_0 and z_0, which are also depicted as solid nodes. The forward propagation information flow direction is indicated by the arrows. Figure adapted from [102].

that these values are once again combined linearly to yield:

$$a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(2)} z_j^{(1)} + w_{k0}^{(2)} \tag{4.3}$$

where $k = 1, 2, \ldots K$. Here, K represents the total number of outputs. This transformation corresponds to the network's second layer, with $w_{k0}^{(2)}$ representing bias parameters. Subsequently, $a_k^{(2)}$ also undergoes transformation using a suitable activation function $f(\cdot)$, resulting in a set of network outputs y_k. The choice of activation functions is dictated by the characteristics of the input data and the presumed distribution of target variables. It is worth nothing that in the context of regression problems, the activation function is typically the identity function, meaning that $y_k = a_k$. Similarly, for binary classification problems, each output unit activation function undergoes transformation using logistic regression functions, resulting in $y_k = \sigma(a_k)$. Here, $\sigma(a)$ denotes the *sigmoid* activation function. Finally, for multiclass problems, the typical activation function being employed is the *softmax* function, defined as follows:

$$\sigma(a) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{4.4}$$

These various stages of computation can be integrated to yield the overall network function, which, for sigmoid output activation functions, assumes the following form:

$$y_k(\boldsymbol{x}, \boldsymbol{w}) = \sigma \left(\sum_{j=1}^{M} w_{kj}^{(2)} h \left(\sum_{i}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} \right) \tag{4.5}$$

Within this framework, the concatenation of all weight and biases parameters is

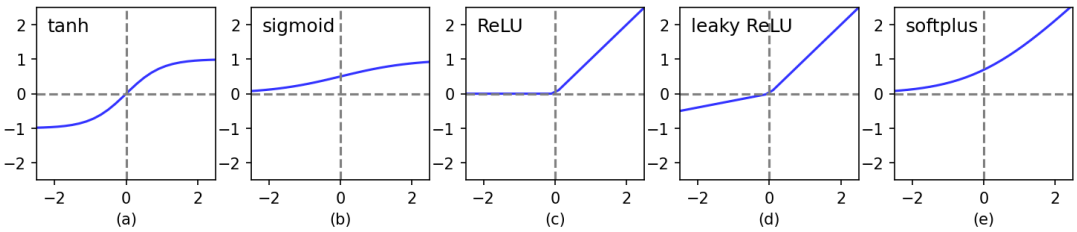


Figure 4.2: Example of typical activation functions used in neural networks.

performed, resulting in vector $\boldsymbol{w}$. It can be seen that the neural network model effectively becomes an overall nonlinear function which maps the input variables x_i to a corresponding set of output variables y_k, controlled by a vector of adjustable parameters $\boldsymbol{w}$. Visually, this function is represented in the form of a network diagram and can be seen in Figure 4.1. The evaluation of Equation 4.5 is referred to as the *forward propagation* of information through the network. In this process, the input data is traversed across the network layers, undergoing nonlinear transformations, and ultimately generating output predictions.

To determine the parameters of the network that allow an accurate mapping from the vector of inputs $\boldsymbol{x}$ to the outputs $\boldsymbol{y}$, a common approach is to minimize the difference between predicted and actual outputs. This process is typically performed using the technique of *error backpropagation*, which is sometimes also referred to as *backprop*. Briefly, error backpropagation relies on the evaluation of the gradient of an error function, denoted as $E(\boldsymbol{w})$, using a local message passing scheme. This scheme involves the alternating forward and backward propagation of information through the network [212]. The simplest such technique is the one originally proposed by Rumelhart et al. [213], involving *gradient descent*. Here, it is important to recognise the two distinct stages involved in the learning process. As such, the first stage, namely the propagation of errors backwards through the network, is used in order to compute the derivatives of the error function with respect to the parameters (i.e., weights and biases) of the network, expressed as:

$$\nabla E(\boldsymbol{w}) = \left(\frac{\partial E}{\partial w_{ji}}, \frac{\partial E}{\partial w_{j0}} \right) \tag{4.6}$$

Secondly, the parameter adjustment occurs using the previously computed derivatives. This update process is represented by the following equations:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}} \tag{4.7a}$$

$$\Delta w_{j0} = -\eta \frac{\partial E}{\partial w_{j0}} \tag{4.7b}$$

where Δw_{ji} and Δw_{j0} denote the changes in weights w_{ji} and biases w_{j0}, respectively,

and the parameter $\eta > 0$ represents the learning rate, which is a small step in the direction of the negative gradient.

Following each update, the gradient is recalculated for the newly updated weight vector. Subsequently, the entire process is then repeated for a number of iterations. It should be noted that the error function is defined in relation to a training set, and the processing of the entire training set. Consequently, the evaluation of $\nabla E(\boldsymbol{w})$ necessitates the processing of the entire training set at each step. Methods that utilize the entire dataset at once are termed *batch* methods. Effectively, at each iteration, the weight vector is adjusted in the direction corresponding to the steepest descent of the error function, hence the name *gradient descent*. Although such an approach seems intuitive, it is demonstrably an ineffective algorithm due to reasons elaborated upon in [214]. Specifically, apart from the increased memory requirements and slow convergence, this approach completely ignores second-order information, as no curvature information is being used. Over the years, several improvements have been proposed in the literature, including, but not limited to the momentum addition [215], Adagrad, [216], RMSProp [217], ADAM [218] or NADAM [219]. For the sake of brevity, the complete details regarding these optimisation algorithms are omitted here. Instead, the reader is referred to the original papers. It should also be noted that the information contained in this section is only intended to showcase the key ideas used in this chapter. For a comprehensive introduction to neural networks, the reader is referred to [102].

4.3 Convolutional neural networks

Having introduced the basic operation principles behind neural networks, attention can be now given to the preferred network architecture used in this study, the convolutional neural network (CNN). While 2D CNNs have become the state-of-the-art in computer vision tasks following their success in numerous applications, such as object detection or semantic segmentation tasks [220, 221], over the recent years, their 1D counterparts have also gained prominence in signal analysis, and more specifically in gait analysis [16, 18, 133, 136, 138, 139]. Similar to their two-dimensional counterparts, 1D CNNs consist of layers of learnable filters that slide across the input data, capturing local patterns and features. These filters, which are

also referred to as kernels[1] in the relevant literature, are convolved with the input data, resulting in the creation of feature maps, that encode the information about the local patterns and structures present in the input space [102]. The convolution operation involves sliding each filter along the input data, computing the dot product between the filter weights and the values within its receptive field across each position. Subsequently, this operation produces an output value for each position, reflecting the degree of similarity between the filter and the corresponding segment of the input data. This operations can be intuitively thought as a cross-correlation operation between the time-series data and the sliding filter. Formally, the *discrete* convolution operation for a time series x of length m, with filter f, of length l is defined as:

$$(x * f)(k) = \sum_{m=0}^{l-1} x_{k+m} \cdot f_m \tag{4.8}$$

Figure 4.3, illustrates a convolutional layer employing a 1D filter. The dashed neurons at both extremities of the time series represent explicitly added zero-valued elements. This operation is often denoted as *padding*. In this specific example, the padding size is equal to one on either side. This padding strategy is employed to exert control over the output shape of the convolution operation.

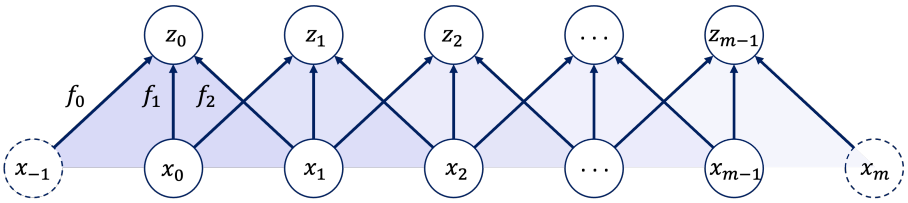


Figure 4.3: Visualisation of a 1D convolutional layer with a filter length 3 and stride 1. The arrows and blue triangles demonstrate the convolution of filter f across multiple time series values x_t. The dashed nodes at both ends represent 'padded zeros', serving as placeholder neurons to regulate the shape of the convolution output. The blue triangles further illustrate the stride of the same filter f across the time series. Figure adapted from [54].

Following the computation of the hidden unit, z_0, the filter is shifted one position to the right. This movement is referred to as the convolution's *stride*, which is a configurable hyperparameter. Stride values exceeding one indicate that the filter is not performing a contiguous slide across the entire time series, but rather a jumping

[1]For clarifications, to prevent any potential confusion with its later usage in dot product computations, this terminology will not be used in this chapter.

motion where it skips certain positions. The output length of a convolution, given an input time series of length m, a convolution involving a filter of length l, padding of size p, and stride s can be mathematically expressed as $((m + 2p - l)/s) + 1$.

In practice, convolutional layers are frequently stacked to create a deeper architecture. Each layer within this architecture employs a set of multiple filters. As such, each individual filter generates a corresponding feature map, highlighting specific characteristics within the input data. This hierarchical stacking of convolutional layers is a key advantage of CNNs. It enables them to learn increasingly complex and abstract representations of the input data through a process of automatic feature extraction. For classification tasks, CNNs are typically constructed to include a series of convolutional and pooling layers. These are subsequently followed by a sequence of fully-connected layers that serve the purpose of mapping the extracted feature maps to a final output layer [102].

4.4 Contrastive learning

One of the most common and effective *representation learning* methods is known as *contrastive learning* [140, 142, 143]. This approach aims to construct a representation space, where pairs of similar inputs, termed *positive pairs*, are positioned close together in the embedding space. Conversely, dissimilar input pairs, termed *negative pairs*, are mapped far apart. The underlying intuition of this approach is that when input pairs are selected such that they exhibit semantic similarity and negative pairs are chosen to demonstrate semantic dissimilarity, a representation space is leant wherein similar inputs tend to be clustered together. This clustering significantly simplifies *downstream tasks* such as classification. Unlike conventional machine learning paradigms, the outputs of contrastive learning are not directly used. Instead, it is far more common to select the activations at some earlier layers to construct the embedding space [102]. Furthermore, unlike methods relying on individual labels or target outputs, the error function in contrastive learning is defined solely with respect to the input data itself.

The efficacy of a particular contrastive learning algorithm is dictated by the methodology employed for selection positive and negative pairs, thereby leveraging prior knowledge to specify good representation characteristics. Commonly, positive pairs are generated by perturbing the inputs in a manner that preserves their semantic

content, while inducing significant alterations in their feature space representations [143]. These perturbations are closely related to *data augmentation*. For time-series data, some examples of data augmentation techniques can be found in [222–225]. Nonetheless, generating physiologically plausible representations is a rather challenging task, although some recent advancements, such as stable diffusion techniques have started to show promising results [226, 227]. However, this aspect is beyond the scope of this thesis. In the case of this work, the labels (i.e., the EDSS score signifying disease severity) were available, meaning that signals of the same class can be used as positive pairs and signals of different classes can be used as negative pairs. In turn, this approach relaxes the reliance on advanced data augmentation techniques for achieving invariance in the representation space and also mitigates the risk of treating two semantically similar signals as negative pairs. Within this context, this is referred to as *supervised contrastive learning* [228] and it was shown to yield superior outcomes compared to conventional cross-entropy based learning [102].

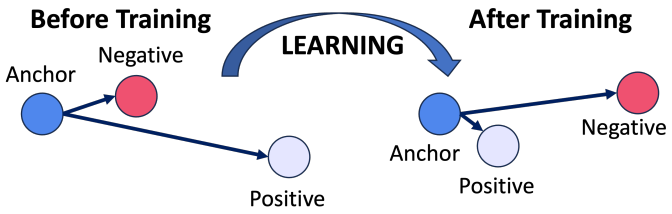


Figure 4.4: Visualisation of the triplet loss framework. The framework learns to transform the input space such that samples from the same MS severity class (*anchor* and *positive* examples) are pulled closer together, while samples from different severity classes (*anchor* and *negative* examples) are pushed further apart. The left panel shows the initial state, while the right panel shows the desired state after training.

An emerging subclass of contrastive learning methods is represented by the *triplet loss* framework. Unlike conventional contrastive learning, which operates on pairs of instances, the triplet loss framework extends this paradigm to incorporate triplet instances. As explained in the introductory section of this chapter, each triplet consists of an *anchor* instance, a *positive* instance (having the same class as the anchor) and a *negative* instance (being dissimilar to the anchor class). The objective within this framework is to maximise the distance between the anchor and the negative pair, while concurrently minimising the distance between the anchor and the positive pair [141]. However, it is not desirable for the training embedding to collapse into very small clusters. The sole requirement is that, given an anchor and a positive

instance, the negative instance should be further away than the positive example, by a certain margin [229]. This concept is akin to the margin used in support vector machines (SVMs) [230, 231] , with the goal of ensuring that the clusters of each class are at least separated by this margin.

Formally, the embedding is represented by $f(x) \in \mathbb{R}^d$ that maps points from an input space $\boldsymbol{x}$ to a d-dimensional Euclidean representation space. Additionally, to force the embeddings to live on a d-dimensional hypersphere, an L2 normalisation is applied to the embedding vector, i.e., $||f(x)||_2 = 1$. Here, it is desired that the *anchor* signal x_i^a of a specific class is closer to all other *positive* signals, x_i^p than to any *negative* signals of a different class, x_i^n. This concept can be visualised in Figure 4.4. Formally, it is desired that:

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2, \tag{4.9a}$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \tau \tag{4.9b}$$

where α is the enforced margin between positive and negative pairs and τ is the set of all possible triplets in the training set, whose cardinality is N. As such, the loss L that is being minimised is defined in the following equation, where the $[\cdot]_+$ denotes the hinge function, which ensures that the loss is non-negative by taking the maximum of the expression and zero:

$$L = \sum_i^N \Big[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha \Big]_+ \tag{4.10}$$

It can be seen that generating of all possible triplets could lead to a multitude of triples that readily satisfy the constraint in Equation 4.9. Such triples, while not contributing to the training, would still be processed by the network, resulting in slow convergence. It is therefore a requirement to select 'hard' triplets, which are active and can therefore contribute to improving the model. These triplets refer to those examples in which the negative sample is closer to the anchor than the positive sample.

To this end, an *online* triplet selection strategy will be presented here, although

an *offline* version is also available (i.e., established before the beginning of each training epoch). The key idea behind online triplet selection is to compute useful triplets during the training process. This method allows for the selection of the most challenging triplets in real-time, based on the current state of the model. It dynamically adapts to the model's learning progress, potentially leading to more efficient and effective training. Formally, given x_i^a, it is desirable to select a *hard positive* instance, x_i^p, such that $argmax_{x_i^p}||f(x_i^a) - f(x_i^p)||_2^2$. Similarly, the *hard negative* is selected such that $argmin_{x_i^n}||f(x_i^a) - f(x_i^n)||_2^2$. It should be also noted that selecting the hardest negative instances can, in practice, lead to collapsed models. It is often more desirable to select x_i^n such that $||f(x_i^a) - f(x_i^p)||_2^2 < ||f(x_i^a) - f(x_i^n)||_2^2$. These triplets are called *semi-hard*, since they are further away from the anchor than the positive instance, yet still presenting a challenge, since the squared distance is close to the distance between the anchor and the positive instance. Such negative instances lie within the margin α.

Finally, having selected a suitable training strategy by mining hard and semi-hard triplets within a batch of training data, and obtaining a good representation of input gait signals, the network used within the triplet-loss framework can then serve as backbone network for the final classification task.

4.5 Layer-wise relevance propagation

Layer-wise relevance propagation (LRP) is an interpretable machine learning explanation technique, which operates by backpropagating the final prediction $f(x)$ backwards through the neural network, using locally defined propagation rules [154]. This technique has already been successfully employed in the context of gait analysis to uncover unique characteristics in the gait patterns [136], but also for the remote characterisation of ambulation in multiple sclerosis [18], motivating its suitability in this work. From an engineering perspective, LRP functions as a systematic method for determining feature relevance by tracing the network's decision back to its inputs, enabling the identification of the most influential temporal regions within the gait signals that contributed to a particular classification decision. The method operates in two stages: first performing a forward pass through the network to gather activation values, weights, and biases of each neuron, followed by a backward pass where relevance scores are assigned to individual nodes and processed layer-by-

layer. The propagation of these relevance scores adheres to a conservation principle analogous to Kirchhoff's current law in electrical circuits - the total relevance received by a neuron in one layer is entirely redistributed to the neurons in the preceding layer, in proportions reflecting their contribution. To illustrate this concept, consider neurons j and k situated in consecutive layers of the neural network, as depicted in Figure 4.5. The propagation of relevance scores R_k from neuron k to neuron j, positioned one layer lower in the network architecture, is governed by the following conservation rule:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \tag{4.11}$$

Within this context, the term z_{jk} represents the degree to which neuron j has contributed to the relevance score of neuron k. The denominator of the equation enforces the LRP conservation rule, ensuring that the total relevance received by neuron k is entirely distributed to its predecessor neurons in the previous layer. This propagation process continues until the input features are reached. By applying this rule to all neurons within the network, it can be readily demonstrated that the layer-wise conservation property holds true: $\sum_j R_j = \sum_k R_k$. Consequently, the global conservation property, $\sum_i R_i = f(x)$ can also be verified. In the following paragraphs, the application of LRP to deep neural networks - particularly those utilising rectifier (ReLU) nonlinearities - will be considered. Here, deep rectifier networks are comprised by neurons of the following type:

$$a_k = \max(0, \sum_{0,j} a_j w_{jk}) \tag{4.12}$$

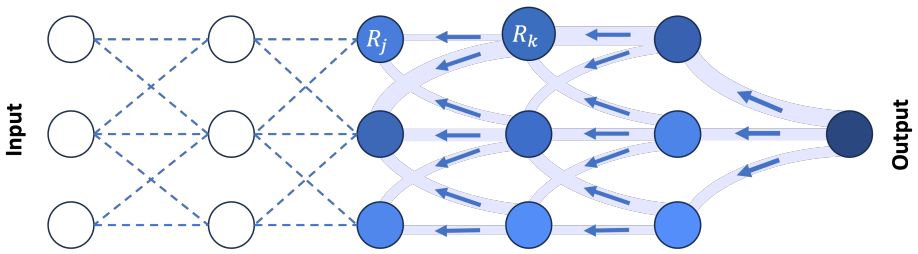


Figure 4.5: Illustration of the LRP procedure. Each neuron redistributes to the lower layers an amount equivalent to what it has received from the higher layer. Figure adapted from [232].

where the sum $\sum_{0,j}$ encompasses all activations $(a_j)_j$ across all lower layers, plus an extra neuron that represents the bias. Specifically, a_0 is set to 1, and w_{0k} is defined as the neuron bias. To begin with, the basic LRP rule is denoted by the $LRP-0$ rule [154]. This rule proportionally redistributes relevances based on each input's contribution to the neuron activations, as expressed in Equation 4.13.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \tag{4.13}$$

This rule satisfies fundamental properties, as it insures that if the activation a_j of a neuron is zero, or if the weight $w_{j:}$ of a connection is zero, then the relevance R_j attributed to that neuron or connection is also zero. This aligns with the understanding that a zero weight, deactivation, or absence of connection would not contribute to the output, and hence, should not be assigned any relevance. While this rule may appear intuitive, it can be demonstrated that its uniform application across the entire neural network yields an explanation equivalent to Gradient $\times$ Input [151]. However, the gradient of a deep neural network is typically noisy, necessitating the design of more robust propagation rules. To counteract this limitation, the $LRP-\epsilon$ rule seems to be viable solution, which has been shown to work well in shallow networks [232], such as the one used later in this work. The enhancement proposed by this rule is the addition of a small positive term, ϵ to the denominator of the $LEP-0$ rule:

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k \tag{4.14}$$

The term ϵ in the above equation represents a stabilization term. This term has been shown to effectively filter noisy relevance maps by absorbing a small amount of relevance when the contributions to the activation of neuron k are either weak or exhibit conflicting influences [232, 233]. As the value of ϵ increases, only the most significant explanatory factors are retained, resulting in sparser and less noisy explanations. Although it is important to acknowledge that other rules exist, such as the $\alpha\beta$ rule, the γ rule, or other compound rules, for the sake of brevity, these are not presented here. Instead, the reader is referred to [232, 233]. In this study, individual LRP heatmaps were generated for the out-of-sample test signals, by employing the iNNvestigate toolbox [234].

4.6 A case study for demonstrating contrastive learning

4.6.1 Participants and data processing

Having established the mathematical details of the technology used in this chapter, the attention can be now given to the dataset used in this study. Given the data-intensive requirements of deep learning methodologies, this study included 125 participants, combining the baseline assessment data from datasets presented in Section 1.3. Out of those, 35 were HCs, while the remaining 90 were MS-affected individuals. Following the classification employed in [17], those PwMS who were able to walk independently and had assigned EDSS scores below 3.5 were classified as mild (*MSmild*). Those who were able to walk independently, but only for limited distances, given EDSS scores ranging between 4.0 to 5.5 were classified as moderate (*MSmod*). Finally, those individuals who were only able to walk using a unilateral or bilateral assistive walking device, and had assigned EDSS scores of 6 and above were classified as severe (*MSsev*). It is also important to note that, given the heterogeneity of the disease and the well-known subjectivity problems associated with the EDSS score [235], a more granular classification, for example into predicting the actual EDSS score was not possible. This aspect was also highlighted previously in Chapter 3, where a clear overlap was seen in the MSD values, across all the MS-affected individuals. The demographics details for the subjects included in this study are given in Table 4.1.

Table 4.1: Demographics table.

	Age	**Gender**	**MS Subtypes**			**Walking Assistive Devices**		
	Mean (SD)	*N male*	*PP*	*RR*	*SP*	*None*	*Unilateral*	*Bilateral*
HC (n = 35)	41.11 (12.75)	14	-	-	-	35	-	-
MSmild (n=23) $\overline{EDSS} = 2.5$	52.57 (11.46)	8	0	5	18	11	8	4
MSmod (n=39) $\overline{EDSS} = 4.76$	46.13 (14.6)	14	3	30	6	32	4	3
MSsev (n=28) $\overline{EDSS} = 6.18$	56.64 (7.91)	11	0	5	23	22	1	5

PP = primary progressive, RR = relapse remitting, SP = secondary progressive
Unilateral = one stick, Bilateral = 2 sticks, walker or rollator

Here, the lower back data was collected using a single three tri-axial IMU, attached to

the body using elastic straps and overlaying the L4-L5 lumbar segments. Following re-alignment to a vertical-horizontal coordinate system, [184], the raw signals were then filtered using a $10Hz$ cut-off zero phase, low-pass Butterworth filter. Resting breaks and turns were automatically removed, according to [17], ensuring that that only segments of steady-state walking were retained for subsequent analysis. Following this pre-processing step, the straight-line walking bouts were further segmented into 2-second segments, using a sliding-window segmentation method, with a stride of 16 samples. This effectively ensured that all segments used as inputs to the neural network were of size 256×6 (where the second dimension corresponds to the number of channels, i.e. vertical, medio-lateral and anterior-posterior acceleration and angular velocity signals). Here, it should be noted that it was decided not to scale the signals, for two main reasons. Firstly, while feature scaling is typically employed to enhance the efficiency of gradient descent training [102], it was deemed unnecessary in this context due to the relatively similar scales of the signals. Secondly, this decision was informed by the findings presented in the previous chapter, which highlighted the significance of the statistical moments characterising the acceleration and angular velocity data. These moments play a important role in enabling the classifier to effectively differentiate between the classes. Therefore, to preserve these characteristics, the filtered sensor data was used in its original scale for the classification task.

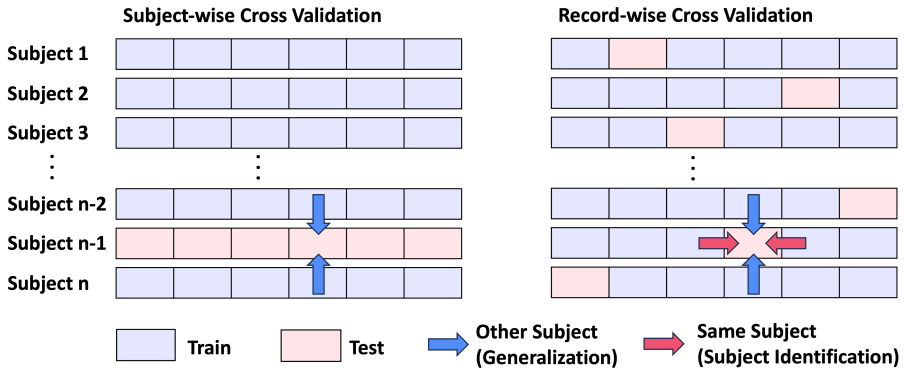


Figure 4.6: Schematic representation of the cross-validation strategies. This figure, adapted from [47], illustrates the subject-wise cross-validation strategy on the left, while the record-wise counterpart is presented on the right.

To validate the proposed methodology, a *k-fold* validation strategy is employed, incorporating both record-wise and subject-wise cross-validation (CV) methods. These CV strategies are visually represented in Figure 4.6. The subject-wise method

(see Figure 4.6 - left), which aligns with the clinically relevant scenario of diagnosing newly recruited subjects (unseen by the model during the training process), was contrasted with the record-wise strategy (see Figure 4.6 - right), which lacks such a clinical interpretation. In a diagnostic context, the aim is to develop a model that can generalise to new subjects, necessitating the use of subject-wise CV. However, record-wise CV, which indiscriminately divides data into training and test sets irrespective of subject affiliation, can also be employed. This method allows for the presence of data from the same subject in both training and test sets, enabling the machine learning algorithm to associate unique subject features with their clinical state, thereby enhancing prediction accuracy on their test data. Consequently, record-wise CV may overestimate the algorithm's prediction accuracy. A stratified *k-fold* CV approach was adopted in this study, dividing the dataset into five folds while preserving the proportion of samples from each class in each fold, thereby mitigating the risk of overfitting or biased evaluations due to class imbalance. Both CV strategies are presented here, since the problem of reliable validation approaches is still an active issue in clinical machine learning applications [47], underscoring the importance of demonstrating the superiority of subject-wise CV.

4.6.2 Network architecture and evaluation metrics

The network architecture used in this study is presented in Table 4.2.

Table 4.2: Final network architecture.

Layer Name	No. Filters	Filter Size	Stride	Feature Map	No. Params
Conv1D	32	8×1	1	256×32	1568
Batch Normalisation	-	-	-	256×32	128
Max Pooling 1D	-	4×1	2	127×32	0
Conv1D	32	4×1	2	64×32	4128
Batch Normalisation	-	-	-	64×32	128
Max Pooling 1D	-	4×1	2	31×32	0
Conv1D	64	3×1	1	31×64	6208
Batch Normalisation	-	-	-	31×64	256
Global Average Pooling1D	-	-	-	1×64	0
Flatten	-	-	-	1×64	0
Dense	-	-	-	1×4	260
Batch Normalisation	-	-	-	1×4	16
L2 Normalisation	-	-	-	1×4	0
					Total params: 12692

It can be seen that the embedding space was effectively selected to be 4-dimensional. To accomplish this, a sequence of three convolutional layers was constructed, each

followed by a batch normalisation and pooling layer. The filter sizes used in these convolutional layers decreased progressively through the network, inspired by the network architecture proposed in [139] for person authentication using IMU data. This network architecture is presented in Figure 4.7, within the triplet-loss contrastive learning framework.

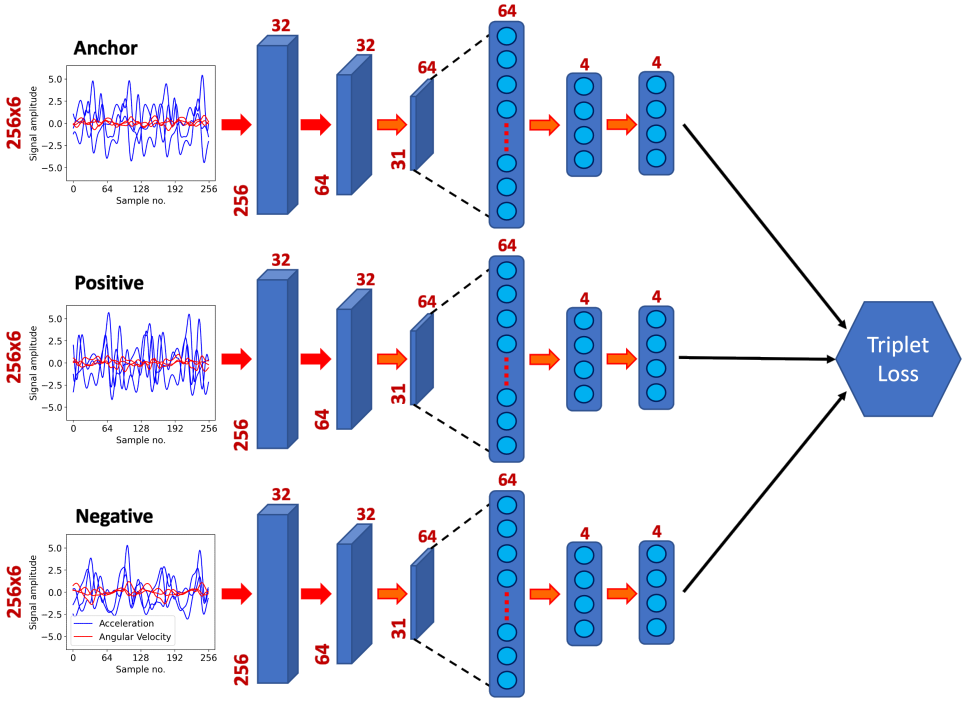


Figure 4.7: Triplet-loss framework visualisation. Here, the loss is minimised by computing the distance between the embeddings of the *anchor* and *positive* pairs and *anchor* and *negative* pairs respectively.

The metrics used to evaluate the quality of the embedding space are presented next. In a traditional classifier, the performance assessment is typically performed using a confusion matrix, and assessing performance metrics such as accuracy, precision or recall. However, within the triplet loss framework, this is not possible, since the model produces embeddings that are in turn used to compute distances between samples. Here, if two gait signals belong to the same class, it is expected that the distance between then will be 'low'. Conversely, when the two signals are from different classes, a 'high' distance is anticipated. However, determining whether these distances are sufficiently 'low' or 'high' can only be achieved by using a threshold. If the computed distance falls below the threshold, it can be inferred that the samples originate from the same class. On the contrary, if the distance between the embeddings surpasses the threshold, the two samples are deemed to belong to different classes. Therefore, careful consideration is required for selecting this threshold. Setting it too low can

result in high precision, but also increases the occurrence of false negatives, whereas setting it too high leads to an increased amount of false positives. Clearly, this problem effectively translates into a receiver operating characteristic (ROC) problem, where the Area Under the ROC Curve (AUC) emerges as a suitable metric for evaluating the quality of the embedding space. Consequently, to classify a pair of gait signals as either 'same' or 'different', a threshold d is compared to their squared L_2 distance, denoted as $D(x_i, x_j)$. Here, x_i and x_j represent the two gait signals being compared. All signal pairs (i, j) belonging to the same class are collectively denoted as $\mathcal{P}_{same}$. Conversely, all signal pairs originating from different classes are denoted as $\mathcal{P}_{diff}$. Following this classification step, the *true accept* is then defined as:

$$TA(d) = \{(i, j) \in \mathcal{P}_{same} \,|\, D(x_i, x_j) \leq d\} \tag{4.15}$$

This set signifies the signal pairs that were correctly classified as belonging to the same class, at a threshold value d. Similarly, the set of all pairs that were incorrectly classified as *same* (*false accept*) is defined as:

$$FA(d) = \{(i, j) \in \mathcal{P}_{diff} \,|\, D(x_i, x_j) \leq d\} \tag{4.16}$$

With these definitions, the true positive rate (TPR) and false positive rate (FPR), for a given threshold d are defined as:

$$TPR(d) = \frac{|TA(d)|}{|P_{same}|}, \quad FPR(d) = \frac{|FA(d)|}{|P_{diff}|} \tag{4.17}$$

Once a good representation is established, this network can be used as the backbone network in the final classification task, simply by adding another *softmax* layer, which computes the probabilities given to each class in the classification problem, i.e. *HC*, *MSmild*, *MSmod* or *MSsev*. Importantly, training of the parameters corresponding to the backbone network is suspended, and only the weights of the newly added layers are optimised using a categorical cross-entropy loss function [102]. As a final implementation note, the margin within the contrastive-loss framework was set to a rather conservative value of 2, while all optimisation tasks during training were performed using the NADAM optimizer [219] with a learning rate set to 0.005, $\beta_1 = 0.89$ and $\beta_2 = 0.989$.

4.7 Results

This section presents the outcomes of the innovative methodology employed for the classification of MS severity, a task which can be framed as a four-class classification problem. The initial step involved the computation of an appropriate embedding space using the triplet-loss framework. Within this latent space, the distance between *anchor* and *positive* examples was minimised, while the distance between *anchor* and *negative* examples was concurrently maximised. To enhance training efficiency, batches of 512 signals were randomly sampled from the dataset. From these, the first 128 hard triplets were selected for the minimization of the triplet loss. Following the recommendations of outlined in [141], 16 additional semi-hard triplets were incorporated into the training set used for each batch. Next, the backbone CNN, previously utilised within the triplet-loss framework, was repurposed for the final severity assessment. The section concludes with a comprehensive analysis of the gait characteristics and their relevance to the model's predictions, conducted using the Layer-wise Relevance Propagation (LRP) technique.

To begin with, the quality of the embedding space is evaluated using both the record-wise and subject-wise CV methodologies. For each evaluation, 20000 signals were used, with 5000 samples drawn from each class, i.e. HC, MSmild, MSmod, MSsev respectively. Table 4.3 presents the AUC metrics achieved across all folds. Here, the highest AUC scores have been highlighted in bold. Moreover, the corresponding ROC curves can be seen in Figure 4.8.

Fold	Subject-wise CV		Record-wise CV	
	Train	Validation	Train	Validation
1	0.9963	0.9748	0.9991	0.9994
2	0.9996	0.9713	0.9995	**0.9996**
3	0.9990	**0.9982**	0.9995	**0.9996**
4	0.9991	0.9981	0.9981	0.9977
5	0.9991	0.9912	0.9988	0.9991

Table 4.3: Cross-validation AUC metrics.

Analysing the results in Table 4.3, it can be seen that the record-wise CV exhibited slightly superior performance on the validation sets compared to the subject-wise approach. This results was perhaps anticipated, since windows of gait signals from the same subject can be present in both training and validation sets. This

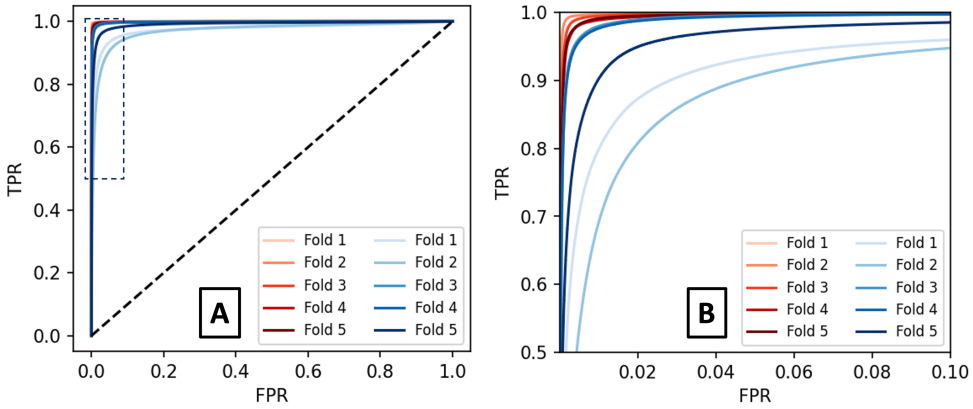


Figure 4.8: ROC curve comparison across all 5 cross-validation (CV) splits. The record-wise ROC curves are shown using red shades, while the subject-wise ROC curves are shown using blue shades. (A) - Full ROC curve plot, the dashed line represents the performance of a random classifier. (B) Zoom into the area enclosed by the blue rectangle found in the top left corned of figure A.

problem is often referred to as *data leakage* and in this case, the model might learn subject-specific characteristics instead of generalizable patterns. Consequently, the model might perform well on validation data due to subject recognition rather than true generalization. The average validation AUC metrics were computed to be 0.9991 and 0.9867 for the record-wise and subject-wise CV respectively. Nonetheless, heuristically, the subject-wise CV achieved excellent results. Visually, this behaviour is also presented in Figure 4.8 B, where the overly optimistic performance of the record-wise CV is immediately apparent. Therefore, the remainder of this work will employ subject-wise CV to ensure robust evaluation and generalizability of the findings.

Next, considering that the backbone network within the triplet-loss framework generates a 4-dimensional embedding vector, Principal Component Analysis (PCA) was utilized to facilitate visualization of the embedding space. As a result, Figure 4.9 illustrates the first three principal components of the embedding space. These components were derived from the network that demonstrated the highest AUC metric on the subject-wise validation set. The hypersphere projection, achieved through the L2 normalisation layer, is readily apparent in this figure. The four latent variable clusters (i.e, *HC*, *MSmild*, *MSmod* and *MSsev*) seem to be arranged in a non-isometric tetrahedral formation on the unit sphere, highlighting the efficiency of the embedding space.

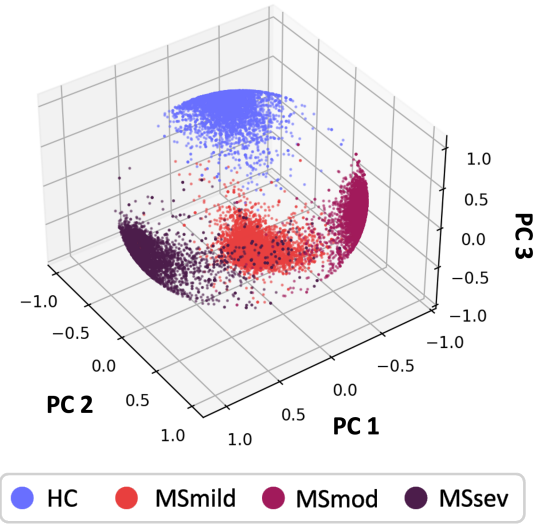


Figure 4.9: Visualisation of the 3-dimensional projection of the embedding space. The datapoints are coloured according to MS severity. For clarification, the marginal *MSsev* datapoints that appear on the center of the sphere are positioned on the sphere's wall, and are not overlapping *MSmild* cluster.

In addition, to further assess the quality of embedding space, pairwise squared Euclidean distances between embeddings were also computed (see Figure 4.10). Here, following the same procedure as presented above, 5000 samples per class were randomly sampled form the entire validation dataset, then squared Euclidean distances were computed between each embedding in the anchor class and all embeddings from the other classes. Across all four plots, there is a clear evidence of class separation, given that the distance between the anchor class embeddings and those from other classes are generally larger than the intra-class distance. This finding aligns with the objective of the triplet-loss framework, which aims to maximise inter-class distances, while minimising intra-class distances. However, these seems to be a difference between the degree of separation between classes. Figure 4.10 A (HC as anchors) shows a clearer distinction between HC and the other classes when compared to plots B, C, and D. This indicates that the embeddings for *MSmild*, *MSmod*, and *MSsev* are more similar to each other in the latent space compared to their similarity with HC embeddings.

Following the establishment of a suitable embedding space, a softmax layer was added to allow an end-to-end neural network classification. The average accuracy obtained across all 5 folds using the subject-wise CV method with the addition of the softmax layer was 98.83%. Heuristically, this represents a very good result. For

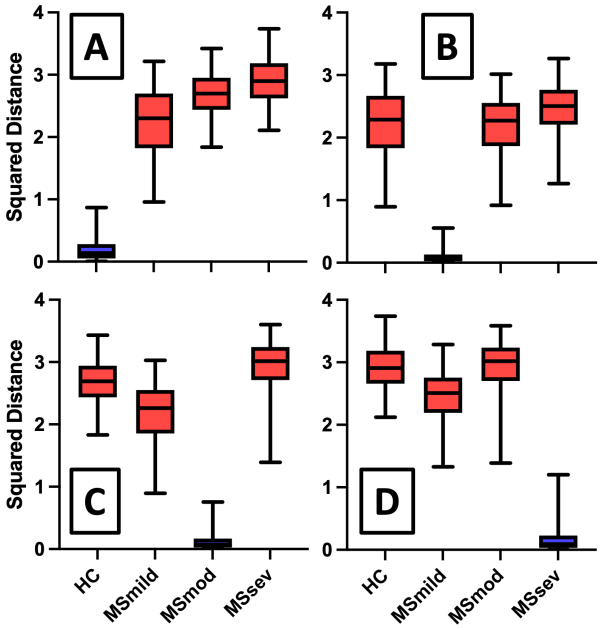


Figure 4.10: Summary of the pairwise squared distances. (A) - All *HC* samples are used as anchors (B) - All *MSmild* samples are used as anchors, (C) - All *MSmod* samples are used as anchors, (D) - All *MSsev* samples are used as anchors. The horizontal line inside the boxes represents the median value, while the box is showing the interquartile range. The whiskers indicate the 2.7 standard deviations range, considering a Gaussian distribution

comparison, the closest resembling study is the one performed by Creagh et al. [18], where the classification accuracy between *HC*, *MSmild* and *MSmod* 84.1%, wherease binary classification tasks, such as *HC* vs. *MSmild*, *Msmild* vs. *MSmod* and *HC* vs. *MSmod* achieved accuracies of 77.6%, 91.8% and 91.1% respectively. It should be noted that the results presented in [18] achieved these accuracy metrics using *transfer learning*. This technique effetively uses knowledge gained from a pre-trained model on a related problem, in order to imporve the performance of the current task [236, 237]. It has been noticed that this approach improved training accuracy by upwards of 8% to 15%, in comparison to traditional training approaches, starting from random network weight initialization. Here, the increased performance of methodology used in this study, is believed to be related to the triplet mining strategy utilised within the contrastive learning framework.

The results described from this point onwards aim to interpret the lower back IMU data using attribution techniques. As such, correctly classified predictions were decoded using the LRP method. Here, it is proposed that this framework allows

end-users to gain at least a partial understanding of the classification decisions. In this work, the LRP framework was employed to decompose the output, $f(x)$, of a learnt function f, given an input vector $\boldsymbol{x}$. This process attributed relevance values R_i to individual samples x_i. Since x_i represents discrete sensor samples from the time domain, the relevance scores, R_i, are directly embedded within the time domain.

Examples of patterns and characteristics across a 2-second signal window are depicted in Figures 4.11, 4.12, 4.13, and 4.14. These figures present relevance scores for representative examples of correctly classified *HC*, *MSmild*, *MSmod*, and *MSsev* subjects respectively. Here, the relevances have been normalised with respect to the maximum value achieved across all channels, in order to highlight the most relevant characteristics for a particular individual. The colorbars displayed at the bottom of the figures represent the color spectrum corresponding to the visualisation of the relevance scores. As such, black regions within the signal window represent regions that are irrelevant to the model's prediction. Red and hot hues represent positive relevance scores, whereas blue and cold hues signify features within the gait signal that are contrdicting the model prediction. For clarity, acceleration signals across the vertical (V), medio-lateral (ML) and anterior-posterior (AP) axes are denoted by acc_x, acc_y and acc_z (measured in m/s^2) while angular velocity recordings are denoted by gyr_x, gyr_y and gyr_z (measured in rad/s) respectively. To facilitate interpretation, the gait events have been overlaid over the sensor data. These events are represented by the initial contacts (IC) and final contacts (FC). Additionally, the probabilistic output of the network is provided in the caption of the figures.

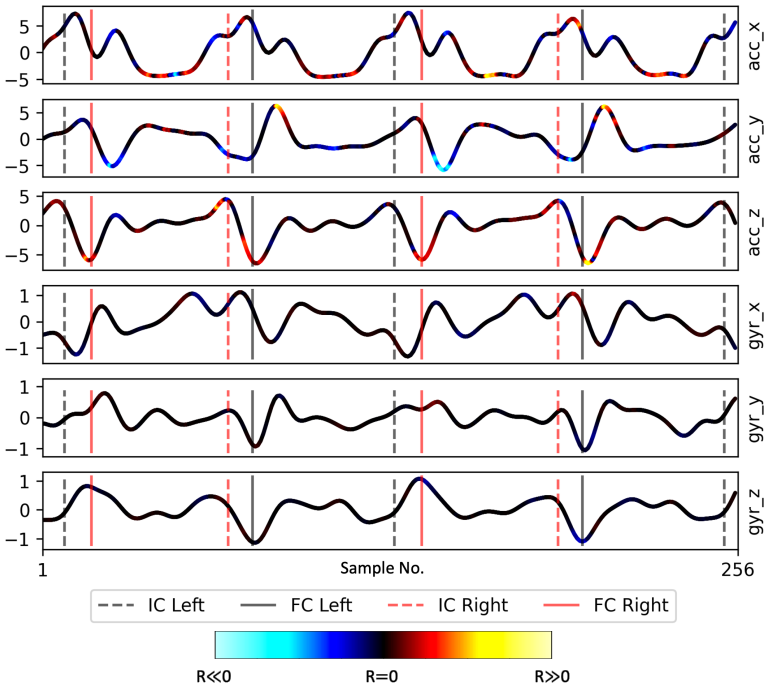


Figure 4.11: Example of a correctly classified *HC* signal window. Probability: 0.8266. Relevance scores are coloured according to the colormap presented at the bottom of the figure.

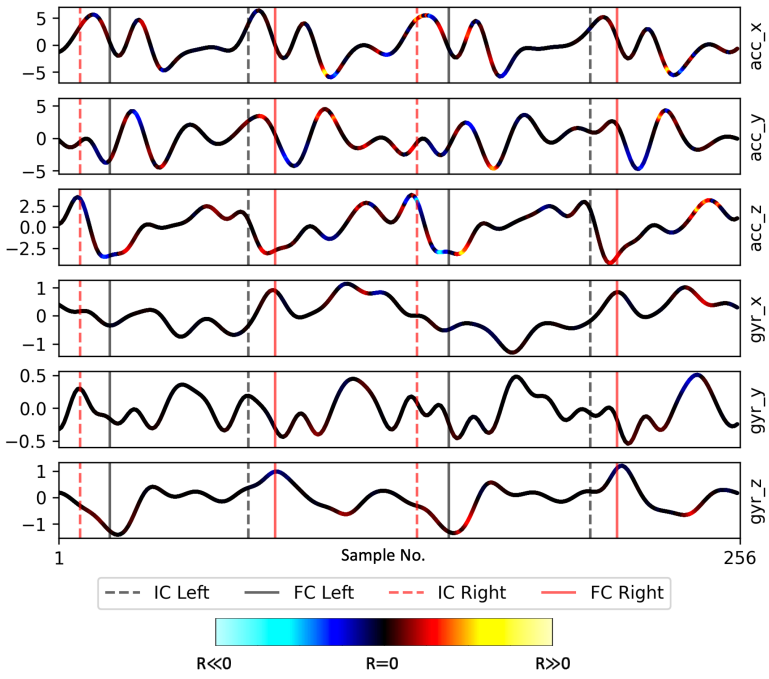


Figure 4.12: Example of a correctly classified *MSmild* signal window. Probability: 0.7027. Relevance scores are coloured according to the colormap presented at the bottom of the figure.

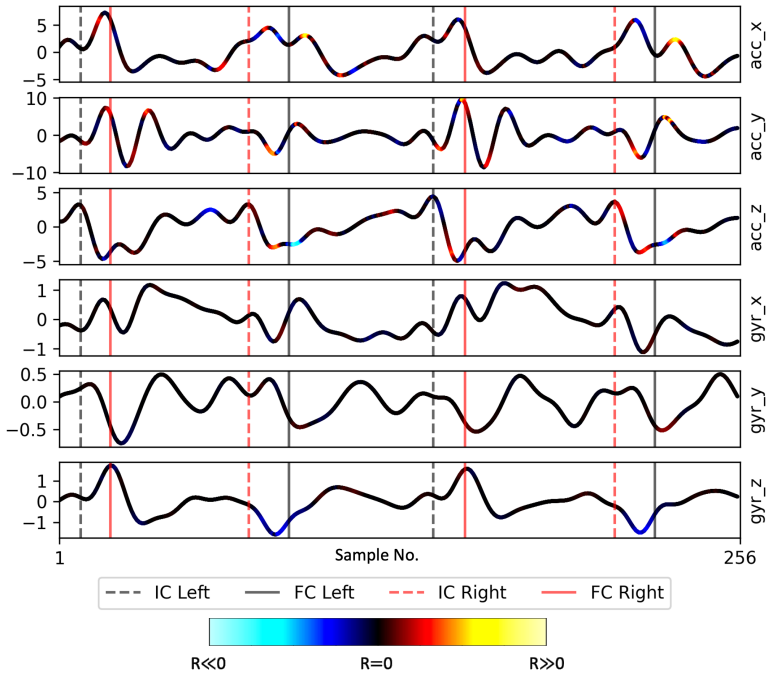


Figure 4.13: Example of a correctly classified *MSmod* signal window. Probability: 0.7732. Relevance scores are coloured according to the colormap presented at the bottom of the figure.

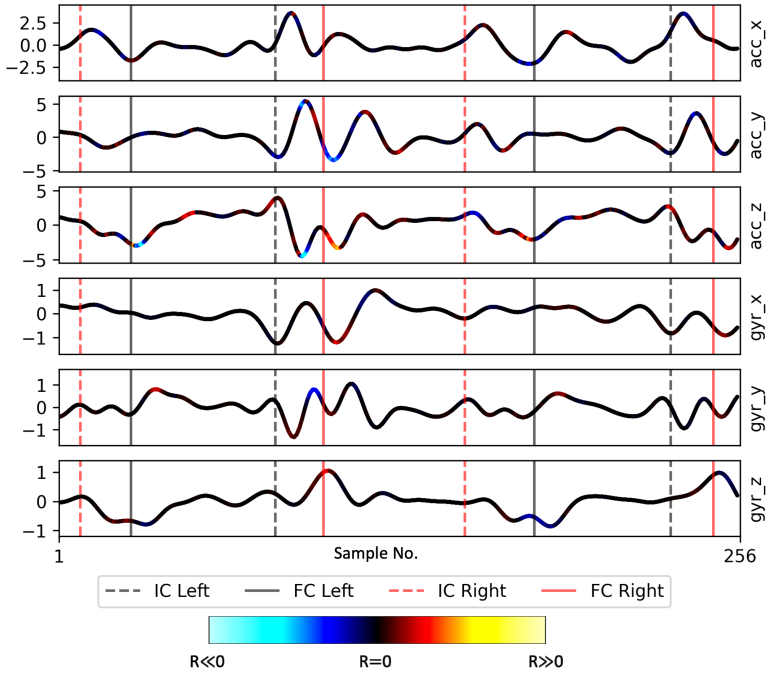


Figure 4.14: Example of a correctly classified *MSsev* signal window. Probability: 0.7905. Relevance scores are coloured according to the colormap presented at the bottom of the figure.

Figure 4.11 depicts an exemplar signal window from a correctly classified *HC* individual. Here, it can be seen that LRP consistently attributed positive relevance scores R_i during the double support phase (denoted as the time between the IC and FC on the contralateral limb), which can be clearly seen in the AP acceleration signal. Moreover the left foot swing phase seems to be a primary characteristic of this particular subject, as denoted by the positive relevances attributed across the ML acceleration signal. Figure 4.12 depicted an example from a representative correctly classified *MSmild* subject. In this case, a clear jerk movement is immediately apparent, which is primarily apparent in the ML acceleration and angular velocity signals. Once again, the positive relevance scores were found during the double support phase, visible in the AP acceleration signal. Interestingly, the presence of asymmetry is also evident here, as indicated by the different relevance scores attributed during the swing phases on contralateral feet. Moreover, positive relevance scores are also given to the AP angular velocity signals. Moving on, Figure 4.13 presents the relevance scores for a correctly classified *MSmod* individual. Similarly to the previous example, the asymmetry is immediately evident, with the highest relevance scores consistently attributed to movements on a single side of the body. Finally, Figure 4.14, depicts the results for a correctly classified *MSsev* individual. Here, the increased duration of steps are immediately recognised. Moreover, the double support phase is again highlighted as the most relevant region in the gait cycle.

4.8 Discussion

This study introduced a self-supervised learning approach for classifying the severity of MS, using wearable sensor data recoded with a single IMUs worn on the lower back. By leveraging the triplet-loss framework as a *pretext* task, an excellent discrimination has been achieved between clusters of *HC*, *MSmild*, *MSmod* and *MSsev* datapoints. Within the scope of this framework, the complexity and variability of the lower back data necessitated a departure from canonical approaches that rely on a set of 'expert' handcrafted features. These features are often viewed as a constrained transformation or approximation of the original signal [18], and can potentially omit some useful information encoding the health status of a particular individual. Given these considerations, a data-driven approach was deemed more suitable. Consequently, a CNN-based approach was employed to extract unconstrained features that are

representative of the four classes considered in this study. Following the successful classification task, the LRP technique has been employed to provide insights into the features most relevant to the model predictions, aiding interpretability of the model, as well as helping in building trust in the model's predictions.

The major component of the novel methodology proposed here is the triple-loss framework. The key motivation of using this self-supervised versus a canonical supervised methodology lies in the establishment of the embedding representation. This allows learning similarities and differences between pairs of examples, being more robust against class imbalance [238]. The triplet-loss framework allows the users to perform end-to-end learning between the input signals and the desired embedding space. This means that the network can be directly optimised for the final task, which renders an additional metric learning step obsolete [229]. Instead, the signals can be simply compared by computing the squared Euclidean distance of their embeddings. Part of the success of this approach is the integration of the hard margin. In the case of this work, a rather conservative value of 2 was selected for this parameter. Given the L2 normalisation layer of the backbone network, the maximum possible squared Euclidean distance between two L2 normalised vectors is 4, which happens when the two vectors are diametrically opposite points on the unit hypersphere. This conservative margin setting might explain the observed non-isometric tetrahedral cluster distribution visualized in Figure 4.9. Whereas the optimal value for the margin parameter is task-dependent, future work may benefit from a sensitivity analysis. Nonetheless, the conservative value used in this study demonstrated exceptional performance on the validation sets, affirming its suitability for this task.

Another important aspect of the triplet-loss framework is the selection of training examples. In this study, an online batch selection strategy was employed, which involves selecting the hardest triplets within each batch of training [141]. However, relying solely on the most challenging examples could disproportionately select outliers in the data, potentially hindering the network's ability to learn meaningful associations [229]. For this reason, additional *semi-hard* triplets were also included within each training batch. This triplet mining strategy, combining hard and *semi-hard* triplets proved to be crucial for fast model convergence. In addition, this study also considered two different CV techniques and highlighted the importance of subject-wise CV techniques, an often overlooked aspect in many machine learning applications within the healthcare sector [47]. In this case, subject-wise CV emerged

as the most appropriate method, aligning with the clinical scenario of diagnosing new patients unseen by the model. This approach, therefore, ensures generalizability and avoids performance overestimation, which is often observed in record-wise CV.

Next, it is perhaps important to revisit and further discuss the implications of the results presented in Figures 4.11, 4.12, 4.13, and 4.14. Here, the LRP holds potential clinical utility in visualising and interpreting the decisions of the neural network used in this study. However, albeit motivated by a clinical hypothesis, the relevance scores were only qualitatively evaluated. For instance, LRP values corresponding to the double support phase were consistency highlighted across all individuals. Moreover, swing phase onset regions were also highlighted to be uniquely describing the gait patterns of PwMS, in line with the results presented in [18, 159, 203, 239–241]. The unique characteristics could be explained by several factors. Firstly, individuals with MS exhibit distinct gait patterns compared to HCs. These patterns are characterised by reduced ankle flexion, particularly in terms of lower limb distal motion and plantar flexor torque. This deficit is most pronounced during the terminal stance and pre-swing phases, which collectively correspond to the double support phase of gait. Consequently, the reduced plantar flexor torque hinders forward propulsion of the trunk and disrupts proper initiation of the swing phase, leading to gait abnormalities during this subsequent phase [239, 240]. Then, considering MS individuals, LRP analysis also revealed that the network assigned significantly different relevance scores to events occurring on opposite sides of the body, reflecting the asymmetric patterns characteristic of MS gait [17]. Moreover, apart for evidencing distinct step inflections, which could represent cadence-based features, and are well-known differentiating factors of MS ambulation, [20, 242], the positive LRP scores attributed to single support regions may also indicate postural instability [203]. For example, the trunk sway can be immediately seen Figure 4.11, for an *MSmild* subject, in line with the results presented in [39]. It can be seen that while verifying the specific clinical meaning of LRP scores presents some challenges, this study offers initial evidence that the proposed network learned clinically meaningful features, further validating the utility of this approach for correctly classifying disability levels in MS.

The proposed methodology offers a number of advantages. First of all, considering that only a single IMU is utilised and the rather conservative number of parameters used by the CNN model, it may have a potential future usage for online embedded systems applications. As such, this methodology may be directly applied in remotely administrated tests, similarly to the study proposed by [16]. Given the excellent

classification results achieved here, it might be also suitable to use this methodology for detecting transition phases from one severity class to another, making it useful for early detection of gait decline. Moreover, another intrinsic advantage of this methodology lies in the avoidance of gait event detection algorithms, which may be prone to inaccuracies. This aspect was certainly visible in the results presented in Chapter 3. Instead, the method presented here only requires unlabelled signal windows of 2-second durations. However, having stressed its advantages, some though must be also given to the limitations of the proposed methodology. As such, it is also noted that although only the $LRP - \epsilon$ propagation scheme was utilised in this study, additional propagation schemes are worthy of investigation in the future, alongside different network architectures, such as modifications of readily available networks used for activity recognition [137, 208]. Moreover, while these results exceed the accuracy of previously reported studies, such as the one proposed in [18], before this approach can be clinically adopted, additional validation procedures are necessary. As such, before deploying this methodology within longitudinal studies, an intermediary validation procedure should investigate the consistency of these result within a time frame of constant disease status, as a first initial step, aiming to validate the robustness of the methodology proposed here.

4.9 Conclusions

In conclusion, the work presented here aimed to explore the ability of deep neural networks to detect impairment in PwMS and correctly classify disease severity using a single inertial sensor mounted on the lower back. Excellent results were achieved using contrastive learning techniques, which effectively established a latent representation space, enabling accurate disease severity classification. Moreover, feature relevances were also investigated and visually represented using the layer-wise relevance propagation technique. This approach provided some evidence that the proposed CNN model learned clinically relevant features.

Enhanced visual representations of deep learning outputs might have the potential to foster closer collaboration between machine learning practitioners and clinical experts. By interpreting these visualizations, clinicians may gain a deeper understanding of how the sensor data captures disease-related gait features, including the most prominent characteristics associated with specific gait patterns. This improved

understanding may ultimately lead to more informed assessments for PwMS. This study particularly highlighted significant differences during the double support phase, but also some potential balance and coordination deficits throughout the swing phase, aligning with previously reported findings. However, while this study showcased the promise of objective and interpretable cross-sectional results, the utility of this proposed framework for longitudinal assessment remains to be further validated.

In summary, the methodologies presented in this thesis thus far offer promising solutions for addressing the first two levels of the proposed hierarchy for exploring the condition of MS individuals, namely gait anomaly detection and severity classification. However, predicting disease progression (gait prognosis) remains a critical challenge. This problem will be the central focus of the following chapters, where the complexities posed by this task will be uncovered and discussed in detail.

Chapter 5

Quantification of gait pattern consistency using autoregressive residual modelling and kernel two-sample testing

Having established robust methodologies for both detecting gait anomalies and quantifying disease severity in the preceding chapters, attention now shifts towards the challenge of longitudinal disease monitoring - the final level of the hierarchical framework proposed for evaluating the condition of MS-affected individuals. However, a key aspect in longitudinal gait assessments lies in the ability to accurately quantify gait consistency, in order to distinguish the inherent multifactorial variability of the gait patterns from disease progression or treatment effects. While the ideal gait pattern would be identical across all steps from an energetic standpoint [243], it is well established that, in reality, gait only demonstrates approximate periodicity [244] and is subject to alterations over time [245]. Here, it is proposed that the longitudinal quantification of gait may be significantly influenced by multifactorial variability encountered at follow-up clinical assessments. Factors contributing to this variability may include marginal discrepancies in sensor attachment locations on body segments or the timing of assessments in the presence of medications, among others. These might mask the subtle degradation or rehabilitation in pathological population, rendering the accurate quantification of longitudinal gait changes, while mitigating

the influence of confounding factors, a challenging task. As such, there is a need for an objective gait consistency measure, capable of assessing an individual's ability to consistency replicate the same walking pattern, regardless of environmental or task-related variations. In turn, this consistency measure can serve as an indicator of good motor control and balance, with its absence potentially indicating an underlying neurological or musculoskeletal deficit. In view of the observed gait pattern variability, owing to both environmental and pathological influences, this chapter introduces a novel objective measure for gait consistency, through the examination of the dynamic link between the lower limbs and the upper body movements during walking tests. This measure utilises two novel components. Firstly, it employs the residuals of a dynamic AutoRegressive with eXogenous input (ARX) model [246] between both shanks and the lower back as a sensitive feature. Secondly, it introduces the maximum mean discrepancy (MMD) [247] to quantify differences in the distribution of the residuals. This approach offers a sensitive and informative method for quantifying and evaluating gait pattern consistency.

5.1 Introduction

In the field of gait analysis, the term "gait consistency" has been subject to various interpretations. Some studies define it as the precision of the measurement tools used for data acquisition [22, 38, 248, 249], assessed through the comparison of specific gait features across different assessments. Conversely, other researchers characterise consistency as the regularity of recurring gait patterns within each gait cycle [244, 245]. It is important to note that consistency pertains to the similarity of gait patterns over a period of time [250]. Therefore, a comprehensive understanding of gait consistency is imperative for discerning between natural variability[1], disease progression, or treatment effects.

Ensuring effective coordination between the lower limbs and upper body is a key requirement for maintaining balance and stability during walking, leading to consistent walking patterns [97, 251, 252]. Similarly, optimal coordination of the hip, knee, and ankle joints is critical for facilitating proper weight distribution and forward propulsion during gait [253, 254]. Conversely, a lack of coordination can

[1]For clarification, even though the term "natural variability" is usually employed to denote the intrinsic variability necessary to maintain balance and adapt to environmental changes, here it is used solely to denote inherent fluctuations in the walking patterns recorded over a period of time.

manifest through inconsistent gait patterns, often associated with specific pathological conditions [255] or an increased risk of falls [256, 257].

Recent advancements in wearable technology have facilitated the use of IMUs to effectively monitor the dynamic relationship between the lower limbs and upper body [17, 253]. While many studies in the literature have focused on extracting spatio-temporal metrics from gait signals [17, 20, 22, 37, 64, 97, 248, 258], this approach may overlook valuable information crucial for quantifying gait consistency [97]. Alternative models proposed in the literature, such as the inverted pendulum model by [259] and the dynamic walking perspective by [243], aim to describe basic acceleration patterns of the pelvis using physics-informed frameworks. However, these simplified modelling approaches may not accurately capture the complexities of pathological gait dynamics. Although effective for estimating spatio-temporal parameters in healthy populations [258], they often fall short when applied to more complex pathological gait patterns, leading to inaccuracies in estimating gait periodicity and symmetry [260].

To counteract these limitations, this chapter explores the application of a data-driven modelling approach as a potential alternative for disease progression monitoring. Leveraging acceleration measurements acquired at the beginning of a baseline walking test, it is proposed that a well-established model might detect changes in gait patterns either throughout the walking test, as well as longitudinally, as the patient's health status evolves. Drawing inspiration from the SHM field, this approach uses an ARX model, which is a linear representation of a dynamic system in discrete time. In SHM, autoregressive models have already been successfully deployed for structural damage identification using accelerometer data in [161–164], relying on analysing the residual error —the difference between measured data and model predictions— as a damage-sensitive feature. Inspired by the concept of monitoring the progress of the residual error as an indicator of change, this chapter presents a similar approach for gait analysis. Within this context, the objective is to detect gait anomalies present in the distribution of the residuals, potentially caused by gait disabilities due to the presence of some disease. Consequently, determining whether there has been a significant change in the residual patterns necessitates conducting statistical hypothesis testing. Fortunately, the MMD two-sample hypothesis test, which will be later explained in Section 5.3, provides a suitable objective hypothesis testing method. This methodology aims to provide a better understanding of gait consistency in both healthy and pathological subjects, while also quantifying the impact of varying

environmental testing conditions.

5.2 Overview of the novel approach for assessing gait consistency

In this section, a high-level overview of the novel methodology for quantifying gait consistency will be presented to the reader. The full details of the modelling procedure are held until Section 5.3.

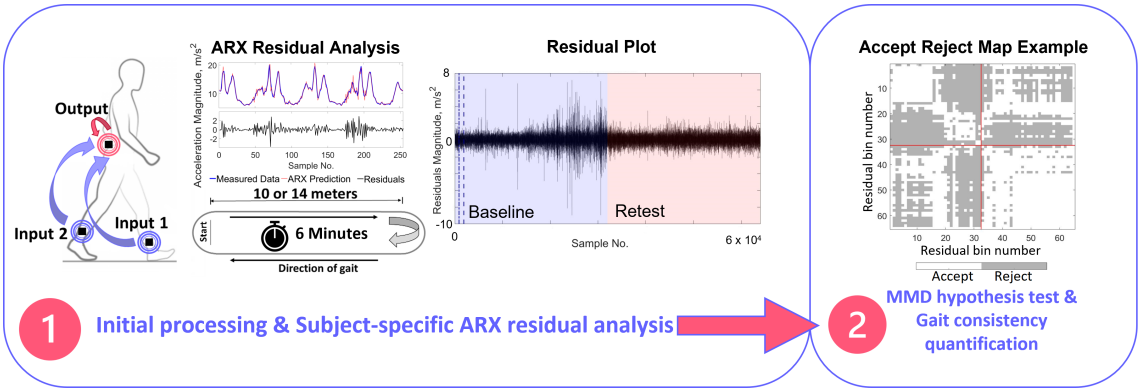


Figure 5.1: Flowchart of the modelling approach.

It is well known that the trunk acceleration signal displays a pseudo-periodic pattern, which closely mimics the repetition of the gait cycle [244]. This pseudo-periodicity induces autocorrelation in the acceleration measurements, preventing the use of any statistical methodology that ignores correlation, as highlighted in [261]. Failure to address this correlation may result in anomaly detectors generating false alarms and failing to identify anomalies, such as irregular walking patterns indicative of neurological or musculoskeletal diseases. However, inspired by the successful adoption of data-driven inference methods in the SHM field, this work addresses this challenge by modelling the dynamic relationship between the upper body and lower limb movements using ARX-type models [246]. In such model, the output is a linear function of previous lagged instances of the output and instances of the inputs. Here, the model output is represented by the resultant acceleration measured at the lower back, while the inputs are the resultant of acceleration signals measured at the shanks. If the ARX-type model accurately reflects the underlying system dynamics,

then, the model residuals, computed as the difference between the actual measured data and the model prediction, should exhibit minimal or no correlation and should appear to be white noise, lacking any discernable systematic patterns.

The philosophy introduced by this modelling approach is straightforward: when a suitable time-series model is identified for a particular individual with a stable and consistent gait, then, the model should yield accurate predictions. Consequently, the residuals are expected to exhibit low variance and be centered around zero. However, if there are changes in the system response, such as those stemming from balance or coordination deficits, the previously identified model may no longer provide accurate predictions, leading to an increase in the variance of the residual sequence compared to the stable condition. Detection of inconsistencies within the residual patterns indicates a change in the system being modelled. These changes could arise from various sources, including alterations in health status or differences in testing conditions, etc.

Once the modelling strategy is established, the next important step is to determine an objective strategy for assessing the similarity of the residuals. Consequently, employing an objective statistical measure, along with a corresponding hypothesis test becomes a necessity. In this work, the employed statistical test revolves around the kernel-based [212] computation of the Maximum Mean Discrepancy [247]. Here, a kernel-based implementation only requires the user to specify a similarity function, formulated as an inner product in a feature space, which is infinite dimensional[2] [212]. The full motivation for adopting the MMD as the preferred statistical measure is deferred until Section 5.3.

Once the statistical metric is established, the final step of the novel methodology introduced here is to convert the associated hypothesis test into an objective measure, enabling the quantitative assessment of gait consistency. Among the other reasons outlined in Section 5.3, one can simply cross-compare smaller segments of the residual patterns and assess their similarity by setting up a hypothesis test. This test utilises the MMD metric to gauge the dissimilarity between distributions, comparing samples drawn from each distribution. The MMD test statistic is computed as the difference between Hilbert space embeddings [262] of the two sets of samples under comparison. A substantial difference suggests different residual distributions,

[2]Although there are a large number of kernels available, practical considerations often lead to the adoption of specific options, such as the Gaussian or Radial Basis Function (RBF) kernel, as utilised in this work

indicating an inconsistency in gait. The number of comparisons equals the square of the total number of smaller residual segments, as illustrated graphically in Figure 5.3. Implementing the hypothesis test across all combinations of residual segments allows the creation of *accept-reject maps*, as depicted in Figure 5.4. The results presented in the flowchart diagram (see Figure 5.1) are primarily qualitative; a comprehensive explanation and interpretation will be provided at a later stage in this chapter. Finally, the consistency of gait can then be computed as the percentage of the number of times the hypothesis test identifies differences in the distribution of residual segments, relative to the total number of comparisons.

5.3 Measuring gait consistency

5.3.1 ARX time series residual modelling

Time series analysis embodies a statistical framework aimed at extracting significant statistics or characteristics from sequences of observations, serving various purposes such as model identification and forecasting future values based on past and present data [263]. As discussed in the previous sections, the entire philosophy introduced in this initial part of this study revolves around monitoring the residual sequences resulting from fitting ARX-type models to the gait data. It is also worth noting that the upcoming section is a focused introduction solely to the modelling strategy employed in this study. For a more comprehensive overview regarding time-series modelling approaches, readers are directed to [246]. In the context of this work, the ARX model takes the following form, where the output y at time t, is given by:

$$y(t) = \sum_{i=1}^{n_a} a_i y(t-1) + \sum_{j=1}^{2} \sum_{k=1}^{n_b(j)+1} b_j(k-1) u_j(t-k+1) + e(t) \tag{5.1}$$

where n_a is the number of lags for the output (in this case the lower back resultant acceleration), $n_b(j)$ is the number lags for the corresponding input (the left or right shank resultant acceleration), a_i is the i-th output coefficient, u_j is the j-th system input and its corresponding coefficient is b_j. Finally, the noise is represented by $e(t)$. Note that the inputs also contain the static regression, as the lower limb and upper body movements occur simultaneously. Additionally, Equation 5.1 is only valid for

one-step ahead predictions. Importantly, this modelling procedure is valid only if the sampling rate of the data acquisition system is maintained constant between consecutive assessments. In the case of this work, the sampling rate has been set to a constant rate of 128Hz across all measurements.

The parameters of the ARX model were estimated by minimising the one-step-ahead prediction error through ordinary least squares - methodology which can be found in [246] or in any good text book on time series analysis. Remembering that the dataset used in this work comprises individuals performing a walking test, going back and forth along a straight corridor, the coefficients were derived from gait acceleration signals recorded during the first straight-line walking bout for all combinations of model orders, with n_a and n_b ranging from 1 to 15. Subsequently, model order selection can be conducted using the Bayesian Information Criterion [264] applied to a validation set, which, in the case of this work, consisted of gait data measured during the second straight-line walking bout. The optimal model was then selected as the one corresponding to the minimum BIC value, defined as:

$$BIC = -2\ln(\hat{L}) + p\ln(N) \tag{5.2}$$

where $\hat{L}$ is the maximum value of the likelihood estimate of the model tested, given the data, p denotes the number of parameters used by the model, and N denotes the total number of observations. The BIC facilitates the comparison of different model structures and serves as a suitable model selection criterion, as it introduces a penalty term for the total number of parameters used in the model, thus mitigating the risk of overfitting [265].

Following the ARX model fitting process, the next step entailed calculating the model residuals vector, denoted by $\mathbf{R}$. This vector represents the difference between the measured data and the model's predictions, as indicated by Equation 5.3. Here, $\mathbf{Y}$ signifies the output vector, X represents the input matrix and $\hat{\theta}$ denotes the estimated ARX coefficients vector.

$$\mathbf{R} = \mathbf{Y} - X\hat{\theta} \tag{5.3}$$

Upon selecting an appropriate model, the residuals are computed across all remaining straight-line walking bouts of the baseline test, as well as throughout all straight-line

walking bouts of the retest. This procedure highlights any alterations in system dynamics during the retest and enables the quantification of gait pattern consistency. Leveraging the availability of test-retest data, this approach offers a key advantage: once a suitable ARX model is established at the baseline assessment, it does not have to be determined again at a later stage. As it will be discussed in forthcoming sections of this chapter, this feature will prove to be extremely useful at highlighting the influence of the confounding factors. Hence, the extraction of model residuals serves as the fundamental basis of the workflow outlined in this chapter. The subsequent phase involves monitoring these residuals, under the assumption that the residual pattern remains consistent across walking tests, provided that the individual's gait remains stable and controlled. Determining whether two sequences of residuals are consistent across different time points involves statistical hypothesis testing, which provides a means of quantifying the degree of gait consistency.

5.3.2 Introduction to the Maximum Mean Discrepancy (MMD) as the preferred statistical metric

Assumptions regarding the form of the residual distributions should not be made without scrutiny. Thus, there is a need to explore flexible methodologies for accurately quantifying the observed discrepancies between repeated measurements and facilitate an objective comparison between test and retest residual patterns. For this purpose, a statistical metric is required, along with a corresponding hypothesis test, which should:

1. quantify the differences between residual patterns using an objective approach, while providing end-users with consistent and interpretable results;
2. account for the complete form of the distribution, rather than a subset of statistical moments;
3. provide non-parametric estimations with convergence guarantees for the density estimations, enabling its application to any given distributions.

While the initial requirement on this list is obvious and is targeting operator bias, the following two necessitate further introduction. Due to the natural variability in the gait patterns in both healthy and pathological populations, no two residual

distributions must be assumed to be the same. This poses a challenge as no *a-priori* knowledge about the form of the residual distributions should be assumed. The kernel trick presents a viable solution to this problem by effectively enabling the assessment of an infinite array of statistical moments via inner products through a reproducing kernel Hilbert space (RKHS) [212, 266].

Formally, a RKHS, $\mathcal{H}$, of functions over an input space $\mathcal{X}$ with kernel k, is a specialised type of Hilbert space of functions $\mathcal{X} \to \mathbb{R}$ with dot the product $\langle \cdot, \cdot \rangle$, satisfying the reproducing property [266]:

$$\langle f(\cdot), k(x, \cdot) \rangle = f(x) \tag{5.4a}$$

$$\text{and consequently } \langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x') \tag{5.4b}$$

This means that the linear mapping from a function f on $\mathcal{X}$ to its value x can be viewed as an inner product. The functional evaluation is given by $k(x, \cdot)$, i.e. the kernel function. An alternative view, is that of a *feature map*: $x \to \phi(x)$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$. This mapping is made possible due to an *isomorphism* between $\mathcal{X}$ and the Hilbert space $\mathcal{H}$. An isomorphism is characterised by being *injective*, signifying one-to-one mappings. The 'trick' arises from the fact that individual mappings $\phi(x)$ do not need to be explicitly computed or formulated; only the kernel function $k(x, x') = \langle \phi(x), \phi(x') \rangle$ is required. As a result of its inherent flexibility, this technique allows non-linear transformations to be applied to the data, improving separation separation between classes, in part by leveraging a greater number of basis functions (dimensions) in the RKHS. As such, the comparison of residual distributions can be extended beyond predefined features, therefore enhancing analysis capabilities. For additional details regarding RKHS, [247, 266] are recommended as helpful references.

Using the reproducing property of RKHS, Smola et al. [266] show that the comparison between two probability distributions, $\mathbb{X}$ and $\mathbb{Y}$, can be effectively computed through a function class that maximises the difference in expectations between probability distributions, called the maximum mean discrepancy (MMD) [267], defined as:

$$MMD[\mathcal{F}, \mathbb{X}, \mathbb{Y}] := \sup_{f \in \mathcal{F}} \left(\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)] \right) \tag{5.5}$$

where $\mathcal{F}$ is a class of functions $f : \mathcal{X} \to \mathbb{R}$, and the shorthand notation $\mathbb{E}_x[f(x)] \coloneqq \mathbb{E}_{x\sim\mathbb{X}}[f(x)]$ and $\mathbb{E}_y[f(y)] \coloneqq \mathbb{E}_{y\sim\mathbb{Y}}[f(y)]$ is used to denote expectation with respect to $\mathbb{X}$ and $\mathbb{Y}$ respectively, $x \sim \mathbb{X}$ indicating that x has distribution $\mathbb{X}$. This formulation facilitates the comparison of distributions by examining their mean embeddings in the RKHS, thereby offering a versatile tool for a range of statistical and machine learning applications.

The MMD is a metric which fulfils all the requirements specified in the above list of requirements [247]. In addition, the MMD can be empirically estimated using both biased or unbiased formulations (this work utilising the latter), depending on whether the V-statistics or U-statistics [268] are used to calculate the sample means:

$$MMD_b^2(\mathbb{X}, \mathbb{Y}) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) \tag{5.6}$$

$$\begin{aligned} MMD_u^2(\mathbb{X}, \mathbb{Y}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i, y_j) \\ - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, y_j) \end{aligned} \tag{5.7}$$

where $\mathbb{X}$ and $\mathbb{Y}$ are the two residual distributions to be compared, x_i and y_i are samples drawn from these distributions, and m and n are the corresponding sample sizes of $\mathbb{X}$ and $\mathbb{Y}$ respectively. To clarify, $\mathbb{X}$, and $\mathbb{Y}$ are stated as probability measures, but generally, will be utilised in the form of a probability density function (PDF).

Finally, although there are many popular kernel functions that can be selected for the computation of the MMD, one of the most popular choices is represented by the radial basis function (RBF) kernel [247, 269], defined as:

$$k(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right) \tag{5.8}$$

where σ is the parameter controlling the of bandwidth of the kernel. It should be noted that typically, this kernel also contains a variance parameter. However, as suggested in [270], this parameter was empirically set as 1. While Gretton et al. [271]

suggest setting up the kernel bandwidth using the median distance between points in the aggregate sample (i.e., concatenating the two datasets to be compared into a single sample), this is only a heuristic, i.e., there is no theoretical understanding of when this is a good choice, and in some cases, it might not be the optimal solution [272, 273]. Therefore, an optimisation procedure would be better suited for this application, which is to be discussed in the sections to follow.

5.3.3 MMD hypothesis test

To quantify gait consistency, one could pose the question of whether the two residual distributions to be compared are similar, and set up a hypothesis test. When provided with a set of independent observations drawn from two distributions, $\mathbb{X}$ and $\mathbb{Y}$, this hypothesis test is used to discern between the null hypothesis and the alternative hypothesis by evaluating the test statistic against a specific threshold. Here, the null hypothesis was set up as H_0: $\mathbb{X} = \mathbb{Y}$, whereas the alternative hypothesis is denoted as H_1: $\mathbb{X} \neq \mathbb{Y}$. For a comprehensive understanding, the procedure involving the MMD-based hypothesis test is detailed in Algorithm 2.

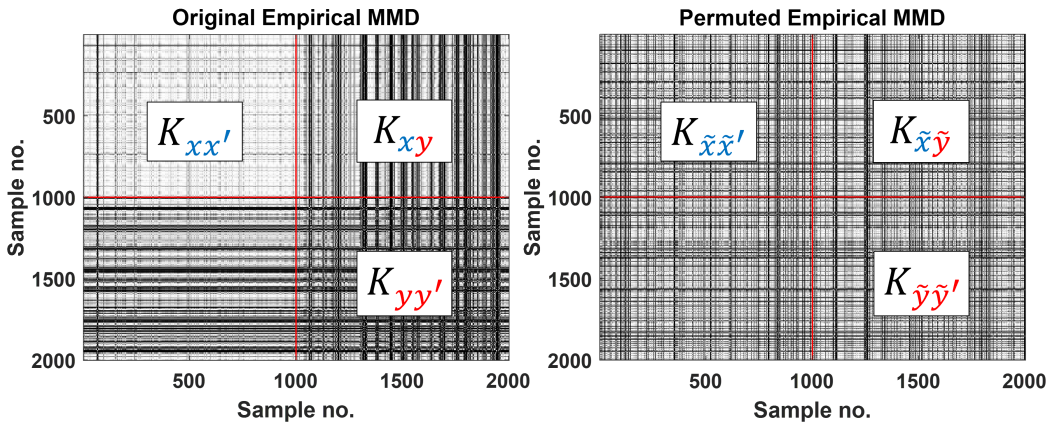


Figure 5.2: Visualisation of the aggregate sample matrix KZ and the permuted matrix, KZ_{perm}. The initial 1000 samples correspond to a segment of 1000 datapoints within a residual signal, while the subsequent 1000 samples correspond to a different sequence of 1000 datapoints within a distinct residual signal. Here, x and x‘ are independent variables with distribution $\mathbb{X}$, y and y‘ are also independent variables with distributions $\mathbb{Y}$, where x‘ and y‘ are independent copies of x and y within the same distributions. $\tilde{x}$ and $\tilde{y}$ are independent variables drawn from the permuted distributions $\tilde{\mathbb{X}}$ and $\tilde{\mathbb{Y}}$. The red lines are used just for delimitation purposes.

Setting up the hypothesis test starts by computing the kernel embedding matrices $K_{xx'}, K_{yy'}$ and K_{xy}, followed by the computation of the MMD test statistic, using either Equation 5.6 or 5.7, depending on whether the biased or unbiased formulations are required. In the case of this work, the unbiased formulation has been used. Subsequently, the aggregate sample matrix, KZ, is formed, as outlined in Line 3 of Algorithm 2. This pre-computation avoids the quadratic-time computational cost in the forthcoming permutation loop. To induce artificially symmetric distributions, a bootstrapping procedure is utilised to establish the objective threshold, as depicted in Figure 5.2 [247]. This procedure is performed a specified number of times, indicated by the variable "*no permutations*". During each iteration, the MMD is recalculated for the permuted distributions. Finally, distances are sorted in ascending order of magnitude, and the threshold is identified as the distance corresponding to the desired confidence level. If the test statistic surpasses this threshold, the null hypothesis is rejected due to insufficient evidence supporting the notion that samples x and y originate from the same distribution. Otherwise, the null hypothesis is accepted, signifying similarity between the two distributions.

Algorithm 2 MMD Hypothesis test

1: Compute $K_{xx'} = k(x, x')$, $K_{yy'} = k(y, y')$ and $K_{xy} = k(x, y)$
2: Compute the MMD test statistic as:
$testStat = \mathbb{E}[K_{xx'}] + \mathbb{E}[K_{yy'}] - 2\mathbb{E}[K_{xy}]$
3: Store $KZ = \begin{pmatrix} K_{xx'} & K_{xy} \\ K_{yx} & K_{yy'} \end{pmatrix}$
4: **for** $i = 1 : no\,permutations$ **do**
5: Permute elemets of KZ and construct:
$KZ_{perm} = \begin{pmatrix} K_{\tilde{x}\tilde{x}'} & K_{\tilde{x}\tilde{y}} \\ K_{\tilde{y}\tilde{x}} & K_{\tilde{y}\tilde{y}'} \end{pmatrix}$
6: Compute the permuted MMD as:
$MMD_{perm} = \mathbb{E}[K_{\tilde{x}\tilde{x}'}] + \mathbb{E}[K_{\tilde{y}\tilde{y}'}] - 2\mathbb{E}[K_{\tilde{x}\tilde{y}}]$
7: Store MMD_{perm} in MMD_{perm_array}
8: **end for**
9: Sort MMD_{perm_array}
10: Compute $threshold$ as the distance corresponding to the desired confidence level
11: **if** $testStat > threshold$ **then**
12: Reject null hypothesis: H_1: $\mathbb{X} \neq \mathbb{Y}$
13: **else**
14: Accept null hypothesis: H_0: $\mathbb{X} = \mathbb{Y}$
15: **end if**

5.3.4 MMD kernel bandwidth optimisation

From Equation 5.8, it is noted that the bandwidth hyperparameter (σ), governing the width of the kernel, needs to be objectively established. Despite heuristic methods being proposed for setting up this hyperparameter, as discussed previously, these methods are deemed unsuitable in the case of this work, due to their propensity for significant type-II errors, particularly in the case of large datasets, as illustrated by Gretton et al. [247]. In a different study, Gretton et al. [273] proposed an optimisation approach for large sample sets, aiming to minimise type-II errors by selecting linear combinations of kernels, thereby enhancing the MMD's robustness to false negatives when used as a test-statistic for two-sample hypothesis testing. Moreover, the RBF kernel embedding used in this work allows for an increased resolution for characterising any given distribution. This effectively translates into the embedding of an infinite dimensional vector of statistical moments, resulting in an asymptotic guarantee that the hypothesis test will capture any distribution, rendering it overly sensitive to infinitely small differences, which is is undesirable.

Noting the above observations and acknowledging that, in practice, the MMD is rarely applied datasets containing more than a few thousand observations, the residual signals were segmented into smaller, more manageable units. These segments, each comprising 1000 data points, will be referred to as *data bins* throughout the remainder of this chapter. Within the context of this work, which utilizes both baseline and retest data (detailed in Section 5.4.1), the MMD hypothesis test was employed to compare all possible residual bin combinations. These comparisons included baseline test vs. baseline test (T-T), baseline test vs. retest (T-R), retest vs. baseline test (R-T), and retest vs. retest (R-R). This approach, visually represented in Figure 5.3, facilitated the generation of an *accept-reject map*, which can be visually interpreted using the examples provided in Figure 5.4.

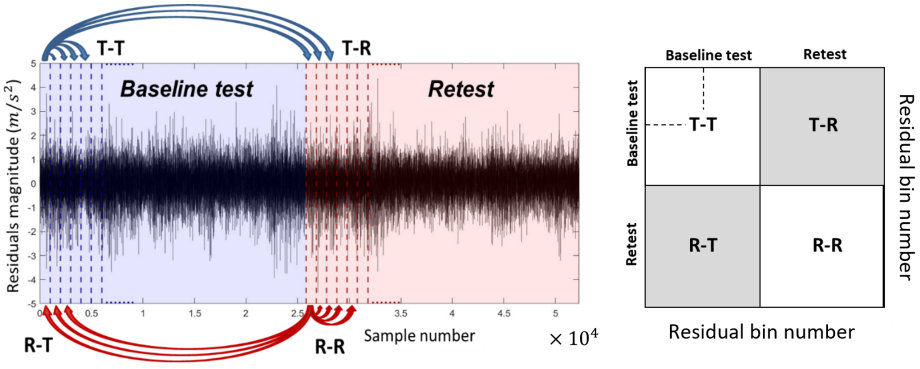


Figure 5.3: Illustration of the data bin comparisons. On the left, the dotted lines represent the divisions of the residuals bins to be compared against the remainders. The arrows represent the two-sample comparisons. On the right, the comparison map is shown. This will then allow the user to visualise the location of the inconsistencies in the residual signals.

Upon visual examination of the residual signals associated with the MS group, it was observed that, in several instances, the residual variance increases towards the end of a walking test, suggesting a potential fatiguing behaviour over prolonged walking periods. Consequently, to assess the consistency of gait patterns between tests and capture the seemingly fatiguing behaviour described previously, it was decided to optimise the kernel bandwidth by minimising the test accuracy in the area enclosed by dotted lines in the T-T quadrant (i.e., utilising the first half of baseline residuals), assuming that a person can walk relatively consistently during half of the baseline test. To clarify, to ensure uniformity, this optimisation procedure was employed for both the HC and MS groups. Formally, the optimal bandwidth with L2 regularisation imposed was determined as:

$$\sigma_{optim=} \arg\min_{\sigma}(accuracy(\sigma) + \sigma^2) \tag{5.9}$$

To arrive at the optimal kernel bandwidth for each subject, it is important to ensure the robustness of the solution in a couple of ways. As conventional optimisation schemes involving gradient descent are impractical due to the bootstrapping procedure within the hypothesis testing, a gradient-free optimisation approach was adopted. To ensure convergence to a global minimum, multiple runs of MATLAB-built-in *fmincon* interior-point and *fminbnd* algorithms were employed, with a search interval constrained to $[0.0001, 10]$ and a cost function value tolerance set at $1e-5$. For

detailed descriptions of the optimisation algorithms, readers are referred to [274–276] or [277, 278] for *fmincon* and *fminbnd* respectively. Finally, the value of the bandwidth corresponding to the minimum function value of all runs was taken as the optimal value.

Upon determining the optimal bandwidth, the final step of the gait consistency framework involves a counting task. Implementation of the MMD-hypothesis test, comparing all residual bins as previously described, populates the *accept-reject map* with test results, in the form of a yes/no survey, indicating acceptance or rejection of the null hypothesis. Finally, the percentage of null hypothesis rejections is computed across all quadrants (T-T, T-R, R-T, and R-R), serving as measure of gait consistency. A higher rejection percentage implies a higher number of gait anomalies in the residual patterns, indicative of less consistent gait. This final counting step marks the completion of the proposed methodology for the objective quantification of gait consistency.

5.4 A case-study for quantifying gait consistency

5.4.1 Participants and initial data processing

The dataset used in this work consists of IMU acceleration recordings from HC and MS individuals, divided into two distinct groups. The first group (group A) consisted of 14 HCs and 26 individuals with MS. The first group of subjects completed the baseline test and the retest on the same day, one hour apart. Importantly, the sensors were not repositioned between the two tests. An additional group (group B) consisting of 23 HCs and 24 MS-affected individuals performed the retest one week apart from the baseline test. This was done in order to introduce natural variability in changing assessment conditions and its effect on gait consistency within a period in which the disease status would not change.

Table 5.1: Demographics table.

		Age	**Gender**	**MS Subtypes**			**Walking assistive devices**		
		Mean (SD)	*N male*	*PP*	*RR*	*SP*	*None*	*Unilateral*	*Bilateral*
Group A	**HC** (n=14)	27.4 (3.7)	8	-	-	-	14	0	0
	MS (n=26) $\overline{EDSS} = 3.9$	44 (13.1)	5	2	23	1	19	5	2
Group B	**HC** (n=23)	49.4 (8.0)	7	-	-	-	23	0	0
	MS (n=21) $\overline{EDSS} = 5.1$	56.5 (10.4)	6	0	0	21	10	8	3

PP = primary progressive, RR = relapse remitting, SP = secondary progressive

Gait data was acquired using three tri-axial IMUs, which were securely fastened to the body through elastic straps, om the anterior aspect of both lower shanks and on the lower back (L4 - L5). The testing procedure is schematically depicted in Figure 5.1. Group A performed the walking test traversing along a 14-m corridor, while the group B walked along a 10-m corridor, going back and forth for 6 minutes. Following the methodology outlined in Chapter 3, Section 3.2.2, all turns and resting breaks were automatically removed, and only straight-line walking bouts of continuous steady-state walking were retained in the subsequent analysis. For the sake of brevity, the details of the data segmentation procedures are not duplicated here. Instead, the reader is referred back to Section 3.2.2. Finally, in order to avoid the effects of possible undesired minor movements of the sensors between sessions, this study uses the raw resultant acceleration. Here, is essential to recognize that while the study employed raw acceleration data, for future applications seeking device-agnostic methodologies, data filtering techniques could be utilised prior to the ARX modeling task. However, in this specific instance, such a necessity was deemed unnecessary, since the same devices were consistently utilized throughout the entire data acquisition process.

5.5 Results

This section presents the results of the novel approach for gait consistency quantification, comprising of utilising ARX modelling and MMD-hypothesis testing. The results are evaluated on the dataset discussed in Section 5.4.1. Specifically, this study targeted two main objectives:

1. To verify whether the consistency of gait is altered by the presence of a locomotor disease (i.e., MS in the case of this work);

2. Quantify the effect of variations in testing conditions.

To begin with, ARX model coefficients were computed on the training set, consisting of the first straight line walking bout of the baseline test. The model order selection, however, was been conducted on a separate validation set utilising data from the second straight line walking bout of the baseline test. Once both the model order and coefficients were established, model residuals were computed across the entirety of the baseline and retest data. Figure 5.4 exemplifies the residual patterns observed for two HCs and two individuals diagnosed MS. Within this figure, the baseline test is denoted by the blue shaded region, while the red shaded region represents the retest. The blue dotted lines situated at the beginning of the baseline test highlight the training and validation regions employed for constructing the ARX models. Upon visual inspection, qualitatively, it can be seen that both HCs are able to maintain a stable gait across both the baseline test and the retest, as indicated by the constant variance of the residuals. In contrast, individuals affected by MS exhibit erratic gait patterns during both the baseline test and the retest. Specifically, the first MS-affected individual demonstrates an increase in residual variance towards the end of the baseline test, while second individual displays a notably different gait pattern during the retest compared to the baseline test.

Next, the corresponding *accept-reject maps* can also be seen on the right in Figure 5.4. These were created by cross comparing the residual bins using the MMD-hypothesis test, for which the number of permutations was set to 500, and the confidence level was set to 99%. Firstly, the *accept-reject maps* depicted on the left-hand side of Figure 5.4 distinctly illustrate the consistent gait pattern observed among HC individuals. The scarcity of grey squares signifies the limited instances where residual bins differed significantly from the norm, leading to rejection of the null hypothesis. in contrast, the *accept-reject maps* for the PwMS confirm the qualitative observations from the corresponding residual plots. The presence of grey areas in the *accept-reject maps* provides additional evidence of gait inconsistencies between affecting this population.

Once the test-retest residual sequences have been computed and the *accept-reject maps* have been generated across all participants, further statistical analysis was required to accomplish the two rather exploratory objectives of this study. Specifically, statistical comparison of the percentages of null hypothesis rejections, indicative of gait anomalies across all quadrants of the *accept-reject map*, was performed. The Mann-Whitney U-test was utilised for this comparison, with a minimum significance

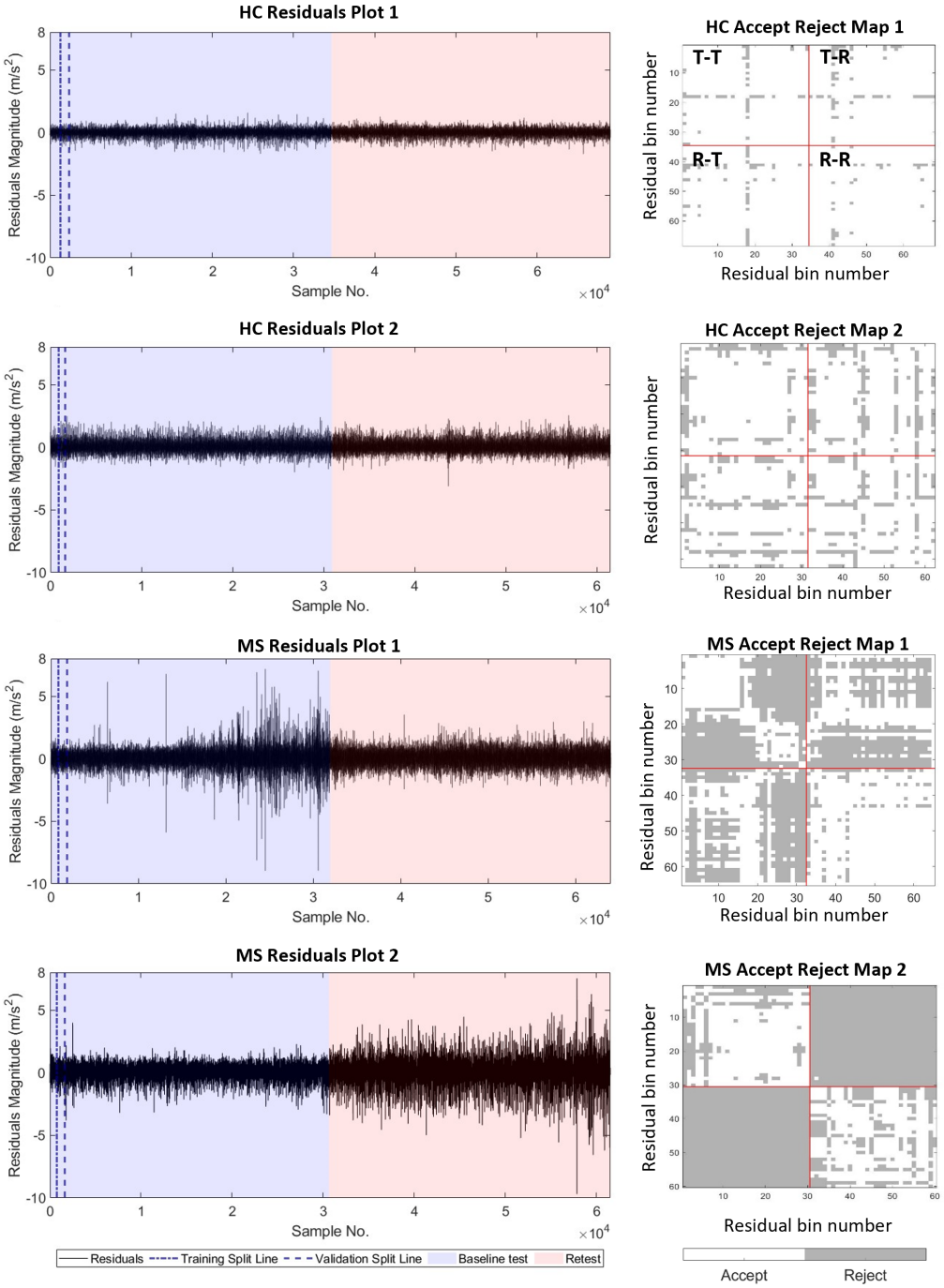


Figure 5.4: Left: Typical residuals patterns for HC (top 2 figures) and MS (bottom 2 figures) individuals. Right: Examples of the corresponding *accept-reject maps*. Here, the red lines mark the T-T, T-R, R-T and R-R quadrants. Gait inconsistencies in the form of null hypothesis rejections are flagged by the grey squares across all bin comparisons.

level of 1.25%, following Bonferroni correction, accounting for multiple comparisons ($\alpha^* = 0.05$ / 4 comparisons). Type II error was evaluated using Cohen's d estimate, with threshold values of 0.1, 0.3, and 0.5 representing small, medium, and large effect sizes, as per [279]. All statistical tests discussed in this paragraph were performed in MATLAB 2021b (MathWorks, Inc., Natick, MA, USA).

The comparison between the percentages of null hypothesis rejections in all the quadrants of the *accept-reject maps* are shown in Figure 5.5. The statistical comparison between the HC and MS groups who completed the retest one-hour apart is shown in Figure 5.5A, while Figure 5.5B shows the comparison for those who completed the retest one week apart.

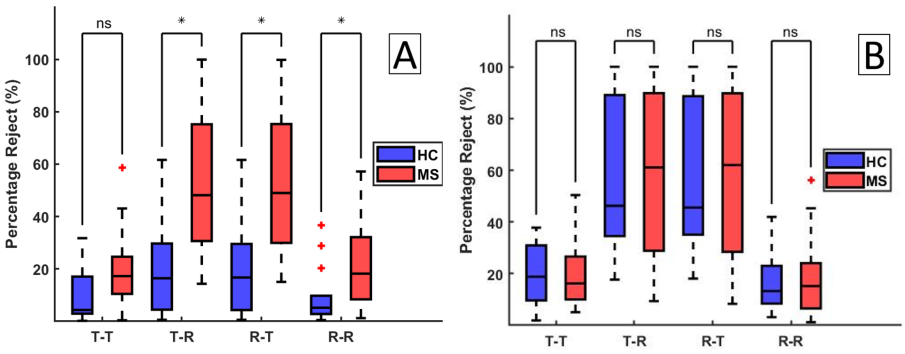


Figure 5.5: Summary of the results. (A) - one-hour apart group, (B) - one-week apart group. The horizontal line inside the boxes represents the median value, while the box is showing the interquartile range. The whiskers indicate the 2.7 standard deviations range, considering a Gaussian distribution. Outlier data is displayed using red crosses. * represents statistical significant difference, *ns* represents a non-significant result.

Examining Figure 5.5A, as indicated by the T-R and R-T comparisons, significant differences in gait pattern consistency among the MS-affected individuals were found even when the retest was performed one hour apart ($p = 0.0002$ and $p = 0.0003$ for T-R and R-T respectively). While the T-T comparison did not reach statistical significance ($p = 0.0208$), it suggests that the MS group displayed a higher number of gait anomalies during the baseline test. This is also indicated by the large effect size ($d = 0.77$) recorded for this comparison and might be interpreted as an indication of fatigue or balance and coordination difficulties during prolonged periods of gait. Conversely, a statistically significant increase in gait inconsistencies during the retest was observed for the MS group ($p = 0.0048$), as indicated by the R-R comparison, while the HC group exhibited a reduction in variance. Overall, large effect sizes were

recorded across all comparisons for group A. Notably, although the T-R and R-T comparisons may appear identical, they differ due to the bootstrapping procedure inherent in the MMD hypothesis test.

Table 5.2: Descriptive statistics for the investigated comparisons, together with p-values for the independent Mann-Whitney U test with Bonferroni correction and associated effect sizes.

	HC (A)	**MS (A)**	**HC (B)**	**MS (B)**	**HC vs. MS (A)**	**HC vs. MS (B)**
	Median (min, 25^th percentile, 75^th percentile, max)				p-value (d)	
T-T (%)	4.33 (0.17, 2.98, 16.62, 31.74)	17.25 (0.39, 10.58, 24.11, 58.69)	18.83 (1.90, 10.21, 30.01, 37.78)	16.20 (5.10, 10.21, 24.25, 50.38)	0.0208 (0.77)	0.8694 (0.10)
T-R (%)	16.44 (0.57, 5.09, 27.39, 61.62)	48.14 (14.33, 30.91, 75.03, 99.90)	46.23 (17.69, 35.96, 88.13, 100)	61.06 (9.38, 30.20, 89.44, 100)	**0.0002** (1.33)	1 (0.07)
R-T (%)	16.73 (0.57, 4.88, 27.48, 61.62)	48.98 (15.06, 30.21, 75.18, 99.79)	45.54 (18.11, 36.17, 87.88, 100)	61.98 (8.33, 29.90, 89.35, 100)	**0.0003** (1.32)	0.9812 (0.07)
R-R (%)	5.20 (0.48, 2.82, 9.09, 36.73)	18.20 (1.21, 8.66, 31.76, 57.18)	13.28 (3.17, 8.49, 22.95, 41.94)	15.19 (1.22, 6.69, 23.44, 56.12)	**0.0048** (0.84)	0.7070 (0.12)

$kp - value < 0.05$ (k = number of multiple comparisons, equal to 4) are in bold.
A: Group 1, who performed the retest one hour apart; B: Group 2, who performed the retest one week apart.

Examining Figure 5.5B, no significant differences were observed between the HC and MS groups across all comparisons. Interestingly, the within-test comparisons (T-T and R-R) did not reveal any discernible differences between HCs and MS, contrasting the differences observed within group A. Moreover, a small effect size was recorded across all comparisons. The associated descriptive statistics characterising the boxplots are presented in Table 5.2.

5.6 Discussions

This study introduces ARX residual modelling as a novel approach for identifying gait inconsistencies in both healthy individuals and those with pathological conditions. By monitoring the residuals, deviations from normal stable gait patterns can be promptly detected when the previously identified ARX model can no longer make good predictions, as a result of significant changes in the system's dynamics. Notably, although model orders and coefficients are subject-specific, meaning that they are uniquely selected for each participant, once computed during the baseline assessment, they do not have to be recomputed again. This feature proved invaluable in elucidating the influence of confounding factors during follow-up assessments. However, it should also be noted, that given the uniqueness of the gait patterns, the models are not

transferrable between individuals, as a result of significant differences in the number of lags and coefficients. Furthermore, it was also noticed that the residual patterns obtained during training and validation walking-bouts were unaffected by factors such as impaired movement, slow walking, asymmetry, compensatory movements, or the use of walking aids. Inconsistencies were observed at later points in time during the baseline assessment or during the retest one week apart. An inherent advantage of this approach lies in its avoidance of gait event detection algorithms, which may be prone to inaccuracies due to the aforementioned factors.

Following the successful implementation of the residual modelling task, the quantification of gait consistency involved evaluating the similarity of residual sequences at various time points using non-parametric statistical hypothesis testing. For this task, a non-parametric hypothesis test is a necessity. Therefore, in the second part of the proposed methodology, this paper introduced the MMD-based hypothesis test, which offers a kernel-embedding of the residuals, and effectively accounting for all the information present in the distributions. This feature offers an advantage as it eliminates the need for end users to specify in advance the specific features of residual distributions that the statistical test should detect. Instead, the kernel trick allows the user to effectively assess infinite statistical moments through the use of inner products in a feature space [212]. Finally, this paper also introduced the concept of *accept-reject maps* as a menas of quantifying gait consistency in an objective manner. The idea of monitoring the ARX model residuals is fundamentally connected to the requirement of a hypothesis test, as the two parts of this novel methodology can only exist in conjunction. Next, it is perhaps important to revisit and further discuss the implications of the results presented in Section 5.5.

The development of this methodology facilitated the achievement of the first objective, which aimed to evaluate whether the presence of a disease influences gait consistency. Specifically, the disease under investigation was MS, which is characterized by gait balance and coordination deficits. As depicted in Figure 5.5A, even when the retest was performed one-hour apart and all the external factors were controlled (i.e., the sensors were not repositioned between the two consecutive tests, consistent footwear was worn, and sufficient rest was provided between tests), individuals with MS encountered challenges in maintaining a consistent gait. Consequently, the proposed methodology may have the potential of being a suitable tool for assessing the impact of short-term clinical interventions, such as the Remote Ischaemic Preconditioning (RIPC) [196]. Moreover, the increased sensitivity of the proposed methodology renders

it suitable for quantifying within-test gait consistency in pathological populations, thereby providing an overall consistency metric, given that less data is used for the kernel bandwidth optimisation task.

The second objective aimed to investigate whether variations in testing conditions can influence gait consistency. The results depicted in Figure 5.5B clearly demonstrate that further work is required in order to improve the generality the models and remove the influence of the confounding factors. In this instance, these factors appeared to exert a greater influence than the disease itself, potentially masking changes in gait patterns during follow-up assessments. Here, the one-week interval between tests allowed for the evaluation of the proposed methodology under more realistic follow-up assessment scenarios, where variations in sensor placement, timing of assessments, preceding physical activity, footwear differences, and other variables may occur. It is important to clarify that the included MS participants did not undergo any disease-related therapeutic interventions, and the one-week interval was deliberately chosen to ensure a stable disease status throughout the study period. By controlling these factors, the effects of disease progression or treatment effects were isolated. The findings of this study also revealed statistically significant results in the healthy population when comparing the baseline assessment with the one-week apart retest, indicating inherent variability in gait patterns, even for the HC population. As such, had the modeling approach demonstrated sufficient generality, no significant differences would have been expected for the T-R and R-T comparisons in group B, particularly within the HC population. Such outcomes would have indicated successful isolation of the natural variability. Remarkably, similar challenges are well recognized in the SHM field [280–283], from which this modeling approach draws inspiration. This highlights the fact that the confounding factors arose from environmental changes in testing conditions are detrimental to the assessment of the condition of the system being analysed, irrespective of the nature of the system. To address these challenges, various tools developed for SHM applications, such as cointegration [282, 283], warrant future exploration.

Due to the unique nature of the methodology employed in this study, a direct comparison with the existing literature is challenging. While the introduced concepts differ fundamentally, perhaps the closest resembling study is the work of Angelini et al. in [22], which investigated the between-session reliability of several temporal, variability and balance gait metrics using data collected one-week apart from both HCs and individuals affected by MS. Their study focused on the consistency of gait

metrics over time, reporting strong agreement between repeated tests. Conversely, the present study directly assessed the consistency of gait patterns themselves, revealing significant discrepancies between test-retest gait patterns. Although the same group of subjects was used for the one-week apart comparison, these differences were expected, given the enhanced sensitivity of the ARX-based method and the MMD-based hypothesis test. In addition to this, several other attempts have been made in the literature towards quantifying the reliability of various spatio-temporal metrics in MS populations. For instance, Morris et al. [10] evaluated gait consistency over a five-hour period in MS patients compared to HCs. Their study revealed that despite discernible differences in gait metrics between MS patients and HCs, the metrics maintained consistency throughout the monitoring period, in contrast to the short-term comparison outcomes elucidated in our study. However, the analysis of [10] was based on a 10-m walking test, recognised for its inherent lack of precision [284] and was only limited to a small set of gait metrics, as only the gait speed, cadence, stride length and double limb support percentage were examined. Another study utilising test-retest data was the attempt presented in [64] for quantifying the potential effects of rehabilitation in MS subjects. In this study, significant differences between test were found across all spatio-temporal metrics included. However, the absence of a control group precludes attributing these changes solely to rehabilitation or varying testing conditions. In contrast, the case-study presented in this chapter includes a control group and demonstrates that gait inconsistencies may serve as an indicator of MS, providing that environmental testing conditions are maintained constant.

It is also worth to consider the potential impact of walking aids and the existence of asymmetry on the outcomes of the present study, given that the latter is frequently recognized as a hallmark of MS [17]. The novel data-driven methodology employed in this research expands its scope beyond the examination of typical gait patterns, enabling the comprehensive analysis of intricate pathological gait, irrespective of its severity or reliance on walking aids. To support this claim and assess the robustness of the proposed methodology, 18 MS-affected individuals relying on walking aids were included in the analysis. The results of the study demonstrated that walking aid utilisation did not compromise the effectiveness of the residual modeling task. This suggests that employing a sufficient number of lags to capture the dynamic relationship between the lower limbs and the upper body ensured that the residuals exhibited characteristics akin to white noise, without any discernable patterns. Furthermore, it is also noteworthy that the ARX modelling procedure employed in

this study inherently addresses scenatios involving asymmetry between the lower limbs. Such instances may manifest as temporal differences, fluctuations in signal amplitude or increased noise levels in acceleration signals recorded on one leg, relative to the other. This aspect is managed by treating the left and right limbs as distinct entities and allocating varying number of lags and unique coefficients as necessary. As such, the model's parameters are automatically adjusted to effectively handle asymmetry, using the BIC, as detailed in Section 5.3.

Having stressed the advantages and potential uses of the newly proposed methodology for quantifying gait consistency, some thought must also be given to the possible limitations. One primary limitation concerns the subject-specific ARX-type models, as no significant differences were observed for the within-tests comparisons between HC and MS, except for the R-R comparison in group A. Several factors may be responsible for these outcomes. Firstly, if the model learns the pre-existing impaired gait pattern of an individual with MS and if that individual consistently maintains the same impaired gait pattern throughout the entire walking test, the residuals will continue to resemble white noise and exhibit minimal variance fluctuations. Secondly, the kernel bandwidth optimisation task was conducted using the first half of the baseline test data. While this method is preferred for validating gait consistency across repeated tests, it might decrease the sensitivity of the hypothesis test throughout the baseline assessment. To mitigate this issue, a practical alternative would involve utilising a smaller subset of the baseline test data for the kernel bandwidth optimisation task. Furthermore, while this approach utilized ARX-type models, which are a linear representation of a dynamic system in discrete time, it is crucial to recognize that, akin to many other engineering applications, the modelled process is inherently non-linear. As such, a more flexible class of regressors will be introduced in the next chapter.

In the second part of the methodology, a potential disadvantage of the MMD-based hypothesis test is the computational overhead, which was mitigated by avoiding the quadratic-time computational cost in the permutation loop, as explained in Section 5.3.3. Moreover, although the T-R and R-T comparisons may not always yield identical results due to the bootstrapping procedure generating two artificially symmetric distributions, the differences are negligible. It is worth noting that the computational time can be further halved by solely considering the upper or lower diagonal matrices of the *accept-reject map*. Moreover, other improvements would involve using the updated versions of the MMD, such as its linear time estimate

[273] or the B-tests [285].

5.7 Conclusion

This chapter has introduced a novel methodology for objectively quantifying gait consistency in both healthy and pathological individuals. While clinical conclusions are probably not advisable, instead, it is perhaps more important to discuss the main ideas introduced by this chapter and their potential future usages. Firstly, the idea of a data-based modelling approach, in the form of ARX residual modelling, has been applied in the context of gait analysis. The aim has been to investigate whether monitoring the residuals can lead to insight into, or enhancement of, the understanding of gait consistency in both healthy and pathological populations. Thus, upon establishing a suitable ARX model, accurate predictions can only be obtained providing a stable and controlled gait, similar to the observed patterns during the learning phase. As a result, the residuals should have a constant variance and resemble white noise. Conversely, in the presence of gait inconsistencies, an obvious departure from the constant residual variance should be recorded. This modelling approach immediately lends itself as an useful tool for monitoring the gait consistency during clinical walking assessments. However, obtaining an objective measure of gait consistency is only possible if the residual modelling task is utilised in conjunction with statistical hypothesis testing. To this end, the MMD-based hypothesis test has been introduced, offering enhanced sensitivity to gait inconsistencies, by effectively extending the comparison of the residual signals to an infinite array of statistical moments via inner products in through a RKHS. Finally, by considering smaller data segments, a single objective measure of consistency has been provided, by cross comparing all the smaller data segments and visually displaying these comparisons using *accept-reject maps*.

While the data-driven approach used in this study is only designed to augment traditional gait analysis, the result of the most immediate importance is that this newly proposed methodology revealed the detrimental effects of varying assessment conditions on gait pattern consistency. Therefore, the obvious direction of the future work targets the exploration of more flexible modelling procedures, which will then allow the long-term monitoring of gait progression in longitudinal studies. This challenging modelling task will be further explored in the upcoming chapters, where

a probabilistic modelling approach will be undertaken, in order to account for the any unforeseen variations or complexities in the gait patterns.

Chapter 6

Towards probabilistic modelling of kinematic gait patterns

In view of the results presented in Chapter 5, where confounding factors were found to have a detrimental effect on longitudinal gait pattern monitoring in multiple sclerosis (MS), this chapter takes a different approach to gait analysis, introducing an innovative and adaptable Bayesian modelling alternative. To this end, due to its role in characterising lower limb distal motion, which is often impacted by alterations in distal muscle involvement, the shank angular velocity emerges as a signal of interest. However, within this pathological population, the shank angular velocity signal exhibits considerable heterogeneity, adding complexity to the modelling task by introducing both within-subject and between-subject variability. Additionally, the data's inherent organisational structure—arising from wearable sensor recordings across contralateral limbs, individual subjects, and broader population groups—necessitates an effective modelling strategy.

Unlike traditional approaches that rely on discrete gait parameters, the methodology proposed in this chapter examines the continuous biomechanical motion throughout the gait cycle, capturing the full dynamics of locomotion. The proposed framework is designed to model gait variability while remaining sensitive enough to detect clinically meaningful changes. This is particularly important in MS, where gait patterns vary significantly not only between individuals but also within the same person over time. To address these challenges, this chapter proposes the Hierarchical Variational Sparse Heteroscedastic Gaussian Process as a flexible and scalable

modelling framework. In essence, this method extends Gaussian Processes by incorporating a hierarchical structure to capture variability at multiple levels (e.g., within and between individuals), employing variational inference to efficiently handle large datasets, and accounting for changing variability levels in the data (heteroscesticity). By modelling the functional form of the shank angular velocity across the entire gait cycle—rather than reducing gait to a set of summary features—this approach allows for a more comprehensive analysis of gait patterns. As such, this framework provides a robust and interpretable methodology for longitudinal gait monitoring, offering automatic uncertainty quantification and facilitating a range of comparisons, from patient-specific assessments to population-level analyses.

6.1 Introduction

For an enhanced understanding and quantification of MS, it is necessary to accurately characterise the lower limb distal motion, as it is often affected as a result of alterations in distal muscle involvement [1, 84, 286]. As such, the shank angular velocity emerges as a potential signal of interest within this context. While the existing literature predominantly employs the shank angular velocity for gait cycle detection, emphasising its effectiveness in identifying key gait event landmarks [74], the full potential of this signal remains largely unexplored. This chapter proposes a new approach which seeks to model the full kinematic signal of the shank angular velocity using a data-driven approach. As a first step before exploring longitudinal gait changes, this model is used to reveal regions of the gait cycle that are mostly affected by the disease or exhibit the greatest variation between contralateral limbs or individuals.

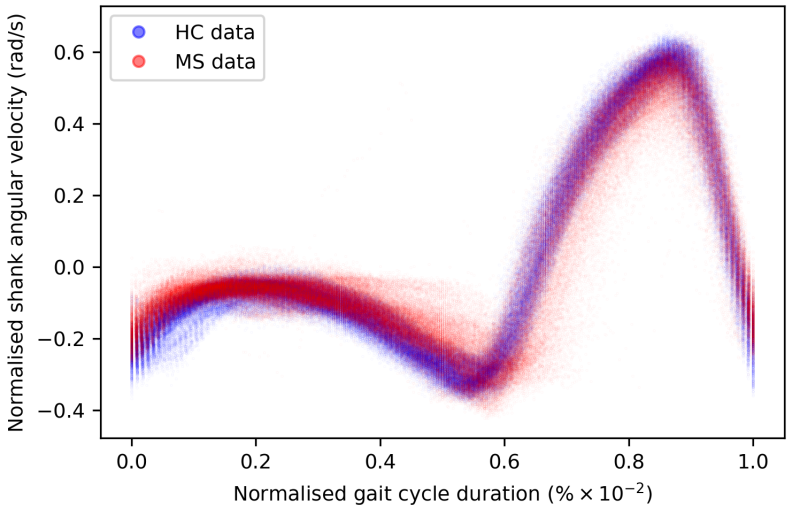


Figure 6.1: Comparison of shank angular velocity data between healthy controls (HC) and individuals affected by multiple sclerosis (MS). The figure presents an aggregate of data points collected using inertial measurement units placed on both shanks, containing a total of 7,899 gait cycles from 28 healthy controls and 7,105 gait cycles from 28 individuals affected by MS.

Typical examples of the shank angular velocity signals collected using wearable sensors are shown in Figure 6.1, for two distinct groups of individuals: HCs, and PwMS. Relative to the HC group, it can be seen that the MS group exhibits increased gait pattern variability, which is particularly discernible from mid-stance to mid-

swing. For clarity, the term 'variability' is often employed in gait analysis to signify stride-to-stride fluctuations during walking, often serving as an indicator of gait impairment [287]. However, here, it is used to denote the dispersion of shank angular velocity around its mean characteristic pattern throughout the gait cycle. Although a direct comparison with the relevant literature presents challenges, as the majority of the studies focus on joint kinematics, rather then segment kinematics (as in the case of the present study), similar trends have been previously documented. [288] reported a significantly increased knee and ankle joint angle variability for MS-affected individuals. [240] proposed that individuals with MS experience insufficient propulsion from the ankle plantar flexor muscles and lack fine motor control during the swing phase, perhaps as a result of favouring the more proximal muscle groups. A reduced range of ankle flexion was also confirmed by [159], even for MS-affected individuals in the prodromal phase of the disease. Similar trends were also reported by [289]. More recently, [290] investigated the deviation phase as a measure of coordination variability and reported significant differences between HC and MS gait during the stance and swing phase. The increased variability observed in the MS group may be attributed to factors such as muscle weakness, spasticity, fatigue or balance impairments, prompting inefficient gait compensations [159, 194, 291].

From a modelling perspective, it can be seen that the gait signals are exhibiting a number of interesting features:

1. The relationship between the input domain and the shank angular velocity is not linear.
2. The variance in the data is not constant across the input-space.
3. There is a common trend shared across all individuals.

While the first point made above is self-evident, the other two observations need additional justification. The process resulting in a non-constant variance across the input-space is known as a *heteroscedastic* noise process, where the variance in the data is dependent on the input, or the noise variance changes across the input domain and can be modelled as a function of the input. In contrast, the process when the variance is independent of the input is referred to as a *homoscedastic* noise process. In regards to the third feature highlighted above, it should be noted that Figure 6.1 displays datapoints from repeated gait cycles collected during straight-line walking for multiple individuals belonging to both groups. Following this clarification,

it becomes clear that there is common pattern shared across all individuals. This observation not only underscores the commonalities in gait dynamics between HC and MS-affected individuals, but also prompts an exploration of the hierarchical structure inherent in the process of data acquisition during gait assessments. Starting with the collection of data from contralateral limbs, this initial stage of organisation extends to an individual level, encapsulating the unique characteristics of each participant's gait, which holds particular relevance in the context of neurological conditions [292, 293]. Subsequently, the aggregation of individual-level data contributes to the formation of distinct groups (HC and MS in the case of this work), which, in turn, become nested within the broader population of individuals, representing a diverse spectrum of human gait patterns. This approach may offer a comprehensive understanding of the lower limb distal movement that considers individual variations, group dynamics, and shared trends across the entire population. However, to the best of author's knowledge, the existing methods present in the gait-analysis community do not necessary capture the hierarchical structure of the acquired data, as presented here. Conversely, the current practice of individual- or group-level walking pattern characterisation involves deterministic approaches, typically using crude method of averaging over multiple gait cycles and potentially overlooking valuable information, or fine-tuning generic models, which marginally improves the accuracy of personalized models. From Figure 6.1, it can be seen that the shank angular velocity is inherently stochastic. Here, it is argued that the full richness of information in this dataset cannot be fully captured via a deterministic approach. Instead, for maximum utility, it should be modelled probabilistically. To counteract these limitations, it becomes a necessity to establish robust methodologies for disease characterisation based on the shank angular velocity, considering both the inherent variability within the MS group, as well as the hierarchical organisation of the acquired data.

In view of the gait characteristics of PwMS and considering the distinctive features observed in the dataset depicted in Figure 6.1, this study pioneers a novel methodology for constructing a robust probabilistic model specifically tailored to the angular velocity. For this task, Gaussian Processes (GPs) appear a to be viable approach. Although a comprehensive introduction to GPs is postponed until Section 6.2, briefly, they can be thought of as distributions over *non-parametric functions* that best fit the data. The probabilistic framework proposed in this chapter can offer valuable insights in the field of gait analysis, especially in the challenging context of assessing and quantifying the degree of gait deficit associated to a patient's health status. In the context of neurological disorders, such as MS, which is marked by intrinsically

unpredictable disease progression [46], the proposed probabilistic framework becomes particularly relevant. A probabilistic framework will provide distributions over the expected gait patterns along with a measure of confidence, allowing informed decision-making and empowering clinicians to make data-driven assessments of pathological gait. This is the opposed of a deterministic approach, where uncertainty is not accounted for, and therefore implying perfect models. A deterministic approach can lead to shortcomings in planning for unforeseen variations or complexities in the individual's gait. Furthermore, the hierarchical extension can advance the analysis capabilities, allowing for a granular analysis of the gait patterns. Here, idiosyncrasies of individual's gait patterns can be immediately captured, together with the corresponding confidence bounds. Moreover, group-level differences can be revealed, as well as isolated. It is therefore clear that a hierarchical probabilistic modelling approach offers tangible benefits for the gait analysis community.

This chapter aims to provide a methodology for accurately modelling the shank angular velocity through an extension of the hierarchical GPs model proposed by Hensman et al. [294], which effectively manages heteroscedasticity and facilitates sparse inference. It is hypothesised that such a model would be able to showcase similar trends consistent with the existing literature on lower limb joint kinematics. Importantly, the model is anticipated to achieve this alignment in a manner that reflects the inherent organisation of the dataset. The contribution of this chapter can be summarised into three key modelling ideas:

1. The hierarchical structure inherent in the data is leveraged to capture the *temporally structured covariance* between contralateral limbs, individual subjects and groups. This hierarchy accommodates the shared underlying population patterns across both groups, while also accommodating the characteristic group patterns present in all individuals in each group. Additionally, it considers the extension of distinctive individual patterns to contralateral limbs, in order to deal with the potential presence of lower limb asymmetry characterising the MS-affected group [286].

2. Given the substantial amount of data collected during clinical assessments, this work addresses scalability challenges through *variational sparse approximations* [295]. This ensures the efficiency of the GP in handling large datasets.

3. Recognising the non-constant variability of the shank angular velocity across the gait cycle, *heteroscedasticity* is introduced into the GP framework by modelling

the variance as an input-dependent function [176].

The modelling approach proposed here could lead to a sensitive and informative method for characterising the shank angular velocity patterns.

6.2 An introduction to Gaussian Processes

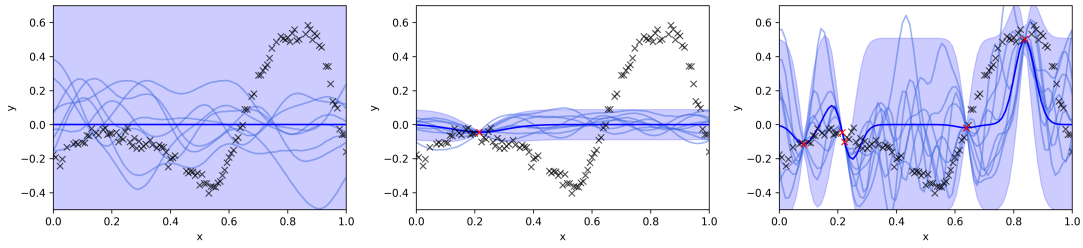
Within engineering, a large number of problems can be considered to be a regression task. Various approaches are available to do this, a practical example being the linear autoregressive approach presented in Chapter 5 for verifying the consistency of the gait patterns between gait assessments at different time points. This chapter provides an additional practical example, focusing on identifying the most probable set of functions that characterise the shank angular velocity kinematic gait pattern over the entire gait cycle (i.e., as a function of time).

In the most general case, regression problems can be solved by firstly defining the D-dimensional input vector, denoted as $\boldsymbol{x}$, while the output (often referred to as the target) is denoted as $\boldsymbol{y}$, forming a training dataset $\mathcal{S}_{train}$ of N observations, $\mathcal{S}_{train} = \{(\boldsymbol{x_i}, \boldsymbol{y_i}) |\ i = 1, \ldots, N\}$. Given the provided training set $\mathcal{S}_{train}$, the objective is to make predictions for new inputs $\boldsymbol{x}_\star$ that have not been seen in the training set. Hence, it is evident that this constitutes an inductive task, as it requires transitioning from the finite training data $\mathcal{S}_{train}$ to a function f capable of making predictions for all possible input values. In order to achieve this objective, it is necessary to make assumptions about the characteristics of the underlying function. Although a wide array of methods have been proposed in order to deal with this problem, two general methods are discussed here. The first method is to impose a bias restriction on the class of functions being considered, for example by limiting the regression problem only to linear functions of the input. Immediately, this approach reveals an obvious problem, in that a decision needs to be made regarding the richness of the class of the function considered. This means that if the underlying function is not well modelled by the chosen function, then predictions will be poor. In fact, within engineering, many processes are inherently non-linear. In response to this, one might also be tempted to increase the *flexibility* of the *function class*. However, this approach is further complicated when it becomes infeasible to write down the exact equations describing a particular system (which can be due to insufficient domain knowledge

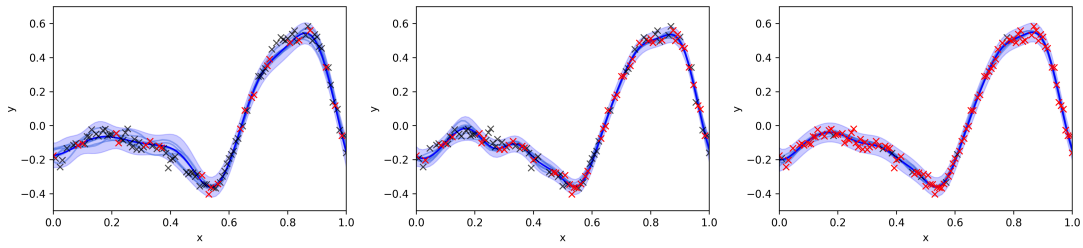
or the system being modelled is so complex, rendering it impractical for the user to completely describe its physical behaviour). Additionally, increasing the richness of the function class might also result in overfitting. Fortunately, the second approach (under a Bayesian paradigm) is to give a prior probability to every possible function, where the higher probability is given to the functions considered more plausible.

The second approach presents a notable challenge: how might one handle an uncountably infinite set of potential functions with finite computational resources? This is where the Gaussian *process* (GP) comes to the rescue.

Although not as widespread as neural network-based technology, GPs have been growing in popularity within the machine learning literature over the recent years. They represent a versatile non-parametric Bayesian machine learning approach for resolving regression problems, enabling the characterisation of *distributions over functions* [166]. In other words, a GP is a generalisation of the Gaussian probability distribution. Whereas probability distributions describe random variables, which are represented by scalars or vectors (in the case of multivariate distributions), a stochastic *process* extends the concept of probability distributions to encompass *functions*. As such, the GP is constructed as a probability distribution, from which any sample is a continuous function over the whole D-dimensional input space [296]. The 'learning' process in the GP is then to determine which functions, from within the initial infinite set, are the most plausible representations of the observed training data. The fundamental nature of the GP is that for new inputs, resembling those previously encountered in the training set, the corresponding targets will also be similar. Before delving into the mathematical framework that facilitates this learning process, it can be beneficial to visually explore this concept. As such, Figure 6.2 shows the progression of the GP 'learning' process, using shank angular velocity data as an example.



(a) GP with 0 training points (b) GP with 1 training point (c) GP with 5 training points



(d) GP with 25 training points (e) GP with 50 training points (f) GP with 100 training points

Figure 6.2: Visual representation of the GP posterior predictions as more training points are added. Discrete data observations are shown by crosses where the ones highlighted in red are used in the training process. The dark blue line corresponds to the mean predictions, whereas the lighter shade lines represent posterior samples. The 95% confidence interval is depicted by the light-blue area.

Figure 6.2a shows the prior model of the GP. Under this prior, very little information is known about the shape of the function describing the system. As such, the function has a zero mean across the entire input domain (indicated by the dark blue line) and exhibits equal Gaussian uncertainty on both sides of this zero mean function (depicted by the light-blue area). Here, samples extracted from the GP are illustrated as lighter blue lines in Figure 6.2. These samples represent non-linear functions that span the space corresponding to the uncertain regions. It is also important to note that uncertainty quantified by the GP does not manifest as uncorrelated white noise. Rather, the uncertainty is correlated across the family of the functions that are being modelled. This correlation ensures that the functions drawn from the GP exhibit smoothness [296]. The elegance of the GP 'learning' process is demonstrated in Figures 6.2b-f. Here, the form of the most likely function representing the data is discovered *nonparametrically*, rather than specifying the number of parameters describing the function *a-priori*. It becomes clear that GP predictions at proximate locations corresponding to the input (i.e. in the vicinity of the training data highlighted by the red crosses) will confidently resemble previously seen values. Visually, this phenomenon is represented by a 'pinching' of the uncertain

region near previously observed training points. As more data are added to the training set, the model's confidence grows in the regions with a higher density of observations.

Having visually introduced GPs, the mathematical considerations will now be presented. Reiterating, a GP is an infinite set of random variables that exhibit a joint Gaussian distribution for any finite subset. GPs have gained significant popularity across a diverse range of applications owing to their ability to automatically quantify uncertainty in predictions, minimal requirement for *a priori* input, and modelling capabilities, even in the presence of high noise levels in the measured data. The GP is developed to model data as the output of some function $f(\boldsymbol{x})$, operating on a D-dimensional input, $\boldsymbol{x}$, as described by Equation 6.1.

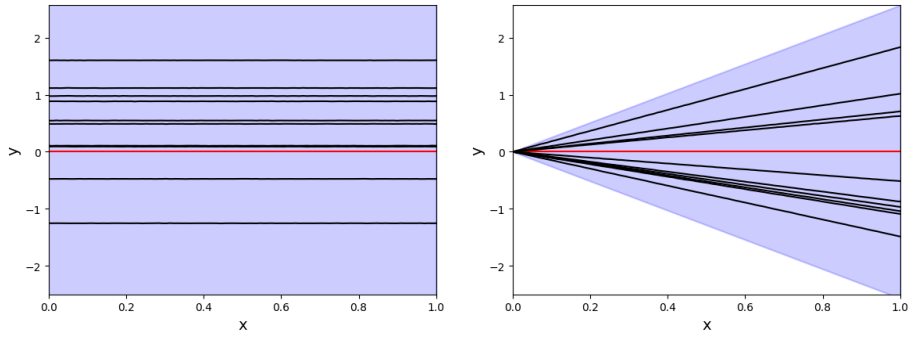
$$y = f(\boldsymbol{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{6.1}$$

Here, it is assumed that the measured values $\boldsymbol{y}$ differ from the latent function values $f(\boldsymbol{x})$ by some additive noise ϵ with zero mean and a predetermined variance σ_n^2. Equation 6.2 formally defines a GP, where $\boldsymbol{x}$ and $\boldsymbol{x'}$ are a pair of inputs to the function of interest:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})) \tag{6.2}$$

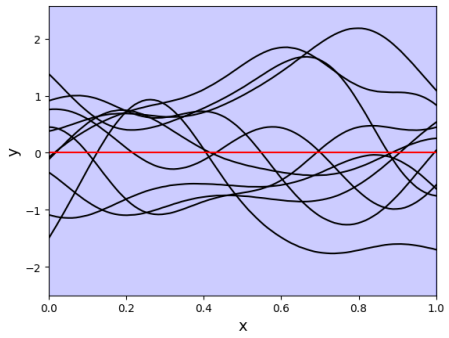
It follows that a GP is completely specified by its mean function $m(\boldsymbol{x})$, and the covariance function $k(\boldsymbol{x}, \boldsymbol{x'})$. The mean function $m(\cdot)$ can be chosen to be any parametric function of the inputs. However, this is commonly set to zero in the relevant lierature [166] (as in the case of this work). The covariance function, (also known as the *covariance kernel*, or simply the *kernel*), is defined as the inner product in a feature space [212], encoding the similarity between any pair of inputs. Although there are numerous covariance functions, a popular choice is the 3/2 Matérn kernel, which is defined in Equation 6.3. It should be noted that the Kernel function is defined by a set of two hyperparameters: the variance σ_f, controlling the vertical scaling (amplitude) of the kernel, and the length-scale l, which controls the smoothness of the functions.

$$K_{\boldsymbol{x}\boldsymbol{x'}} = \sigma_f^2 \left(1 + \frac{\sqrt{3}(\boldsymbol{x} - \boldsymbol{x'})^2}{l}\right) \exp\left\{\frac{-\sqrt{3}(\boldsymbol{x} - \boldsymbol{x'})^2}{l}\right\} \tag{6.3}$$

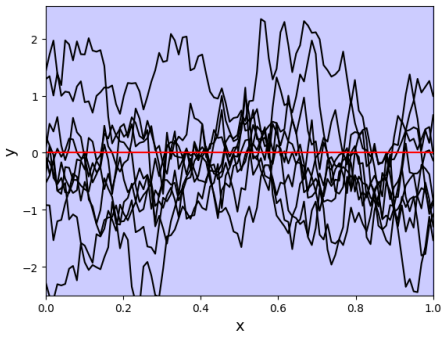


(a) Constant kernel

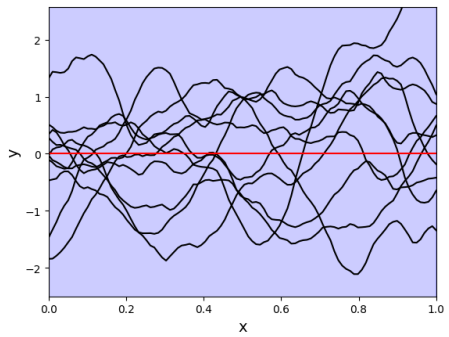
(b) Linear kernel



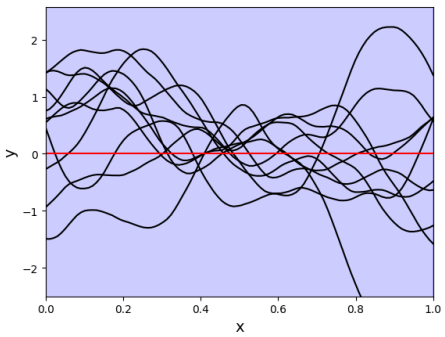
(c) RBF kernel



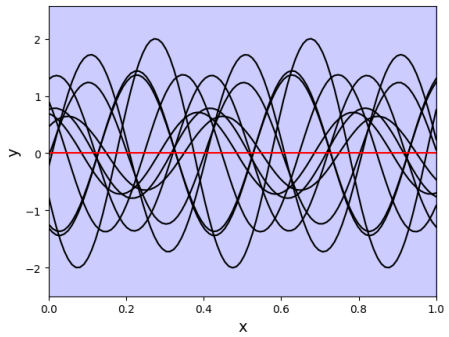
(d) Matérn 1/2 kernel



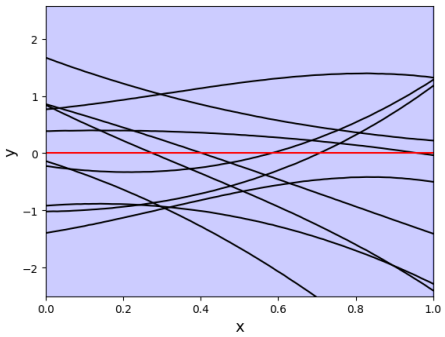
(e) Matérn 3/2 kernel



(f) Matérn 5/2 kernel



(g) Cosine kernel



(h) Polynomial kernel

Figure 6.3: Realisations from the prior of the Gaussian Process using 8 commonly used kernel functions.

The covariance function governs the *function class* which comprise the prior of the GP. That is to say that the covariance function directly impacts the spectrum of functions within the infinite set, from which samples are extracted. Hence, the choice of this function significantly impacts the GP's modelling performance [296]. Perhaps a more useful way to visualise what the covariance function is to leverage the generative nature of the GP, since it is possible to draw samples from the prior over the function space and plot potential outputs using a particular kernel before observing any data. As such, the plots in Figure 6.3 show realisations from the prior for eight different kernels functions. Here, the mean is highlighted by the red line, whereas the distribution over the possible functions (up to 3σ) is shown by the blue area. The realisations are plotted using the black lines. The influence of the covariance function is immediately evident here. However, although automatic kernel selection strategies have been proposed in the literature (see [297]), determining an 'optimal' kernel choice is can be a a very difficult task and should be a key concern for the practitioners, before attempting any kind of GP modelling tasks. In the case of this work, the optimal choice for modelling the abrupt changes in the slope of the gait traces was found to be the 3/2 Matérn kernel.

Having established the mean and covariance functions, the prediction task is achieved by assessing the joint Gaussian distribution of the observed target values $\boldsymbol{y}$ and the function values at the new test locations $\boldsymbol{y}_\star$ under the prior, as:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(X) \\ m(\boldsymbol{x}_\star) \end{bmatrix}, \begin{bmatrix} K_{XX} + \sigma_n^2 \mathbb{I} & K_{X\boldsymbol{x}_\star} \\ K_{\boldsymbol{x}_\star \boldsymbol{X}} & K_{\boldsymbol{x}_\star \boldsymbol{x}_\star} + \sigma_n^2 \mathbb{I} \end{bmatrix} \right) \tag{6.4}$$

In Equation 6.4, X denotes a set of N, D-dimensional training inputs, where $X \in \mathbb{R}^{N \times D}$, whereas $\boldsymbol{y} \in \mathbb{R}^{N \times 1}$ is the corresponding set of N measured training outputs. By Assessing the joint Gaussian distributions, it is now possible to make predictions $\boldsymbol{y}_\star$ at new test input locations $\boldsymbol{x}_\star$, given the training inputs X and their corresponding outputs $\boldsymbol{y}$. Thus, the predictive distribution over $\boldsymbol{y}_\star$ is now given in Equation 6.5:

$$p(\boldsymbol{y}_\star|\boldsymbol{x}_\star, \boldsymbol{y}, \boldsymbol{y}) \sim \mathcal{N}(\mathbb{E}[\boldsymbol{y}_\star], \mathbb{V}[\boldsymbol{y}_\star]) \tag{6.5a}$$

$$\mathbb{E}(\boldsymbol{y}_\star) = m(\boldsymbol{x}_\star) + K_{\boldsymbol{x}_\star \boldsymbol{X}}(K_{XX} + \sigma_n^2\mathbb{I})^{-1}(\boldsymbol{y} - m(\boldsymbol{x})) \tag{6.5b}$$

$$\mathbb{V}(\boldsymbol{y}_\star) = K_{\boldsymbol{x}_\star \boldsymbol{x}_\star} - K_{\boldsymbol{x}_\star X}(K_{XX} + \sigma_n^2\mathbb{I})^{-1}K_{X\boldsymbol{x}_\star} + \sigma_n^2\mathbb{I} \tag{6.5c}$$

In order to learn the hyperparameters, a Type-II maximum likelihood (ML-II) approach is used [166], by maximizing the *marginal likelihood* of the model. However, for convenience and numerical stability, the optimisation is performed as a minimization task over the negative log marginal likelihood. Therefore, the set of hyperparameters, denoted by θ are chosen through the following optimisation:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} -\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{6.6}$$

with,

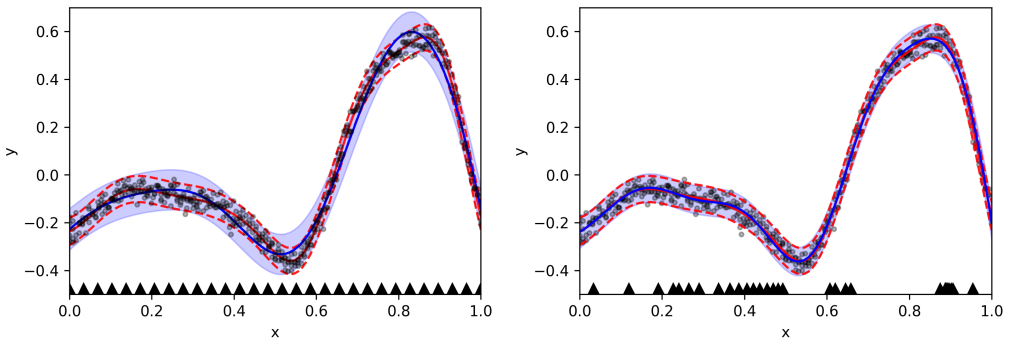
$$\begin{aligned} -\log p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = -\log \mathcal{N}(\boldsymbol{y}|, m(\boldsymbol{x}), K_{XX} + \sigma_n^2\mathbb{I}) = \\ = \underbrace{\frac{N}{2}\log(2\pi)}_{constant\ \text{term}} + \underbrace{\frac{1}{2}\log|K_{XX} + \sigma_n^2\mathbb{I}|}_{complexity\ \text{term}} + \underbrace{\frac{1}{2}\left((\boldsymbol{y} - m(\boldsymbol{x}))^T(K_{XX} + \sigma_n^2\mathbb{I})^{-1}(\boldsymbol{y} - m(\boldsymbol{x}))\right)}_{model\ \text{fit term}} \end{aligned} \tag{6.7}$$

The annotated terms in Equation 6.7 have readily available interpretations, and it can be clearly seen that there is a trade-off between model fit and model complexity. This property is known as the Bayesian Occam's Razor [166, 298]. Thus, the hyperparameters of the kernel can be learnt and the GP is completely defined by Equations 6.4 and 6.5.

6.2.1 Sparse GPs for large datasets scaling

Either learning the hyperparameters of the GP or making predictions involves taking the inverse of the covariance matrix with noise, $(K_{xx} + \sigma_n^2\mathbb{I})^{-1}$, which is an operation scaling as $\mathcal{O}(N^3)$ in computational complexity. Hence, in practice, it is not feasible to perform full GP regression tasks on datasets involving more than roughly ten

thousand datapoints. This is also one of the limitations preventing the use of full GP regression on gait data, as the number of datapoints collected during a visit often exceeds ten thousands points per subject. To address this limitation, a number of approximation methods have already been proposed in the literature [295, 299–301]. Broadly speaking, these approaches are divided into two main classes, namely *model approximations* and *posterior approximations*. For the sake of brevity, the reader is referred to [166, 299, 300] for more details about these approaches. The posterior approximation approach is widely recognised to generally provide more robust approximations and possess inherent mechanisms to counteract overfitting. Thus, the present study employs a posterior approximation method, specifically the Variational Free Energy (VFE) method proposed by [295]. The main advantage of this approximation method is the reduction in time complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$, where M are the number of inducing points introduced. Clearly, this becomes advantageous when $M \ll N$.



(a) Before optimisation of inducing points.

(b) After optimisation of inducing points.

Figure 6.4: Sparse GP predictions. The mean variational approximation predictions are shown by the solid blue line, with the corresponding 95% confidence interval shown by the light blue area. The mean predictions of the full GP are shown by the red lines, while the full GP 95% confidence interval is shown by the red dashed lines. The locations of the inducing points are shown on the bottom of the plots using the black triangles (a) - before optimisation and (b) - after optimisation.

The *variational approximation* of the full posterior is handled through the use of small set of auxiliary points, called *inducing points*, $\{Z, \boldsymbol{u}\}$ (where Z contains the locations of the inducing points and $\boldsymbol{u}$ are the values of the latent functions at these points). They can either be selected as a subset of the training inputs, or simply as auxiliary pseudo-inputs. A practical example of GP posterior approximation can be seen in Figure 6.4, where the inducing points allow to rigorously approximate the exact GP

by minimising the distance between the sparse model and the exact one. As such, the variational model can then be learnt by minimising the Kullback-Leibler (KL) divergence between the approximate joint posterior and the full joint GP posterior. This minimisation is equivalent to maximising the following lower bound, $F(Z)$, of the true log marginal likelihood [295]:

$$F(Z) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log|Q_{XX} + \sigma_n^2\mathbb{I}| - \frac{1}{2}(\boldsymbol{y} - m(X))^T[Q_{XX} + \sigma_n^2\mathbb{I}]^{-1}(\boldsymbol{y} - m(X)) \\ -\frac{1}{2}\sigma_n^{-2}tr(K_{XX} - Q_{XX}) \tag{6.8}$$

where $tr(\cdot)$ is the trace operator and Q_{XX} is the approximate covariance matrix, defined as[1]:

$$Q_{XX} = K_{X\boldsymbol{u}}K_{\boldsymbol{uu}}^{-1}K_{\boldsymbol{u}X} \tag{6.9}$$

Here, the kernel functions, evaluated at the data points X, inducing input points Z, and between the data and inducing points, are represented by the kernel matrices K_{XX}, $K_{\boldsymbol{uu}}$ and $K_{X\boldsymbol{u}}$ respectively. It can be seen that the bound resembles the one used in Equation 6.7, with the novelty of this approach being the inclusion of the regularisation trace term. This bound derived in Equation 6.8 can then be used for hyperparameter optimisation. For the complete derivation of this bound, the reader is referred to [302].

Following optimisation, making predictions can be done in a comparable manner to the standard GP. Hence, noting that the explicit conditioning of the posteriors on the training data, test input locations, inducing points and hyperparameters was dropped here for simplicity of notation, the predictive distribution is given by:

$$p(\boldsymbol{y}_\star|\boldsymbol{x}_\star, X, \boldsymbol{y}, \boldsymbol{u}) = \mathcal{N}(\mathbb{E}[\boldsymbol{y}_\star], \mathbb{V}[\boldsymbol{y}_\star]) \tag{6.10a}$$

$$\mathbb{E}[\boldsymbol{y}_\star] = Q_{\boldsymbol{x}_\star X}(Q_{XX} + \sigma_n^2\mathbb{I})^{-1}\boldsymbol{y} \tag{6.10b}$$

$$\mathbb{V}[\boldsymbol{y}_\star] = K_{\boldsymbol{x}_\star\boldsymbol{x}_\star} - Q_{\boldsymbol{x}_\star X}(Q_{XX} + \sigma_n^2\mathbb{I})^{-1}Q_{\boldsymbol{X}\boldsymbol{x}_\star} \tag{6.10c}$$

[1]Following [295, 300], this notation can be generalised, such that $Q_{\boldsymbol{ab}} = K_{\boldsymbol{au}}K_{\boldsymbol{uu}}^{-1}K_{\boldsymbol{ub}}$

The main benefit of this sparse approximation is that the computational requirements are reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$ for M inducing points[2]. Therefore, with this approximation method, the standard GP can be scaled up to large datasets, such as those containing gait data collected during clinical assessments for multiple patients.

6.2.2 Heteroscedastic noise GP models

With reference to the dataset presented in *Figure 6.1*, it is evident that the *homoscestic* noise assumption of the GP model is not satisfied, since an increase in signal variance is immediately observed towards the termnations of the stance phase, as well as during the swing phase. To address this limitation, *heteroscedastic* GP models (i.e. those using input-dependent additive noise) have been developed [176]. Focusing on the MS group, it can be seen that the variability of the gait patterns is one of the key aspects of the disease. Hence, enhancing the model's ability to effectively capture the inherent variability will enable the establishment of more accurate confidence intervals for predictions. In this case, the regression model introduced by Equation 6.2 would then become:

$$y = f(\boldsymbol{x}) + \epsilon(\boldsymbol{x}), \quad \epsilon(\boldsymbol{x}) \sim \mathcal{N}(0, r(\boldsymbol{X})) \tag{6.11}$$

This means that the variance of the noise process is now a function of the model inputs. Notably, the *heteroscedastic* noise model presented above reduces to a *homoscedastic* one when $r(\boldsymbol{x})$ is a constant. The first derivation of the homoscedastic GP model was firstly introduced by [176]. To define the heteroscedastic model, a GP prior is firstly placed on the unknown function $f(\boldsymbol{x})$, such that:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(0, k_f(\boldsymbol{x}, \boldsymbol{x'})) \tag{6.12}$$

To ensure positivity of the noise variance, $r(\boldsymbol{x})$, an exponential transform is applied, as:

[2]Naively, it is not clear where the computational speed-up is found. However, note that the Woodbury inversion lemma can be applied to $(Q_{XX} + \sigma_n^2\mathbb{I})^{-1} = (K_{Xu}K_{uu}^{-1}K_{uX} + \sigma_n^2\mathbb{I})^{-1} = \sigma_n^{-2}\mathbb{I} - \sigma_n^{-2}\mathbb{I}\underbrace{(K_{uu} + K_{uX}(\sigma_n^{-2}\mathbb{I})K_{Xu})^{-1}}_{\mathbb{R}^{M\times M}}K_{uX}\sigma_n^{-2}\mathbb{I}$

$$r(\boldsymbol{x}) = \exp(h(\boldsymbol{x})), \quad h(\boldsymbol{x}) \sim \mathcal{GP}(\mu_0, k_h(\boldsymbol{x}, \boldsymbol{x'})) \tag{6.13}$$

Here, $h(\boldsymbol{x})$ is modelled by a GP, whose covariance function is denoted by $k_h(\boldsymbol{x}, \boldsymbol{x'})$. The GP has a constant mean, μ_0, which controls the *scale* of the noise process. Here, μ_0 is introduced as a learnable parameter that models the average noise level across the function. This approach deviates from standard function modelling, where $\mu = 0$ is typically assumed. By learning μ_0, the model can account for the inherent noise present in the data, corresponding to the homoscedastic case. The addition of the secondary GP placed on the log noise variance increases the expressiveness of the GP, albeit simultaneously increasing the complexity of the learning and inference processes.

As a result of the inclusion of the heteroscedastic noise model, the log-likelihood becomes untractable and can no longer be computed analytically. Similarly to the sparse GP extension, a variational method is used to approximate the distribution over the posterior and form a new lower bound, as described in Equation 6.14.

$$F(\boldsymbol{\mu}, \Sigma) = \log \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, K_f + R) - \frac{1}{4} tr(\Sigma) - KL(\mathcal{N}(\boldsymbol{h}|\boldsymbol{\mu}, \Sigma)||\mathcal{N}(\boldsymbol{h}|\mu_0 \boldsymbol{1}, K_h)) \tag{6.14}$$

To ensure clear notation, K_f and K_h are used here to denote the covariance matrices of the two GPs used to model $f(\boldsymbol{x})$ and $h(\boldsymbol{x})$ respectively. R is a diagonal matrix whose elements are are $R_{ii} = \exp(\boldsymbol{\mu}_i - 0.5\Sigma_{ii})$ Moreover, $\boldsymbol{\mu}$ and Σ are the variational parameters to be determined, defined according to Equation 6.15, for some positive semidefinite diagonal matrix Λ:

$$\boldsymbol{\mu} = K_h \left(\Lambda - \frac{1}{2} \mathbb{I} \right) \boldsymbol{1} + \mu_0 \boldsymbol{1}, \tag{6.15a}$$

$$\Sigma^{-1} = K_h^{-1} + \Lambda \tag{6.15b}$$

It can be seen that this implementation requires the optimisation of $N + N(N+1)/2$ free variational parameters. This is achieved through the reparametrisation of $\boldsymbol{\mu}$ and Σ in terms of Λ [176]. As a result, in practice, the computational complexity is roughly twice that of the homoscedastic GP.

Another challenge that has emerged from the use of a heteroscedastic Gaussian Process (GP) is the lack of the complete predictive distribution in a closed form. Fortunately, it is still possible to approximate the first two moments (i.e. the mean and variance) of the predictive distribution. These are expressed in the following equations:

$$\mathbb{E}_q[\boldsymbol{y}_\star] = \boldsymbol{a}_\star \tag{6.16a}$$

$$\mathbb{V}_q[\boldsymbol{y}_\star] = \boldsymbol{c}_\star^2 + \exp\left(\boldsymbol{\mu}_\star + \frac{1}{2}\boldsymbol{\sigma}_\star^2\right) \tag{6.16b}$$

Here, the following notations have been used:

$$\boldsymbol{a}_\star = k_f(x_\star, X)(K_f + R)^{-1}\boldsymbol{y} \tag{6.17a}$$

$$\boldsymbol{c}_\star^2 = k_f(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - k_f(\boldsymbol{x}_\star, X)((K_f + R)^{-1})k_f(X, \boldsymbol{x}_\star) \tag{6.17b}$$

$$\boldsymbol{\mu}_\star = k_h(\boldsymbol{x}_\star, X)(\Lambda - \frac{1}{2}\mathbb{I})\mathbf{1} + \mu_0 \tag{6.17c}$$

$$\boldsymbol{\sigma}_\star^2 = k_h(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - k_h(\boldsymbol{x}_\star, X)(K_h + \Lambda^{-1})^{-1}k_h(X, \boldsymbol{x}_\star) \tag{6.17d}$$

With these definitions, it is now possible to make predictions using a heteroscedastic GP. Here, it is assumed that the first two moments presented above are representative of the true underlying distribution. The implementation presented above enables probabilistic estimation of the latent underlying functions, along with accurate prediction of variance, at any given inputs and at a cost comparable to that of a standard analytically tractable homoscedastic GP. The additional information regarding the uncertainty of the process will prove to be an important capability for modelling kinematic gait patterns, especially in pathological populations. It is now straightforward to extend this variational method to enable sparse heteroscedastic GP inference.

6.2.3 Sparse heteroscedastic GP regression

Having established the foundation of both a sparse and a heteroscedastic GP model and given the gait data shown is Figure 6.1, it is reasonable to investigate the

integration of these two approaches, giving rise to a sparse heteroscedastic GP. To this end, the Variational Sparse Heteroscedastic Gaussian Process (VSHGP) has already been derived in the literature by [303]. Inspired by the model presented in [176] and in Section 6.2.2, [303] have shown that is possible to derive an analytical $ELBO$ using m inducing points for the mean function GP, $f(\boldsymbol{x})$, and u inducing points for approximating the log noise variance GP modelling $g(\boldsymbol{x})$. As a result, this approach is scalable to large datasets, given its $\mathcal{O}(nm^2 + nu^2)$ complexity.

Similarly to the approaches presented in the previous sections, the approximation of the VSHGP model also reduces to computing a lower bound on the marginal likelihood. Although there is a non-trivial amount of algebra to arrive at this point, to learn the optimal set of hyperparameters for the GP model, the problem reduces to maximizing the following lower bound, which is defined as:

$$\begin{aligned} F(\boldsymbol{\mu}, \Sigma) = log\mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, Q^f_{XX} + R_h) - 0.25Tr[\Sigma_h] \\ -0.5Tr[R_h^{-1}(K^f_{XX} - Q^f_{XX})] - KL(\mathcal{N}(\boldsymbol{h}|\boldsymbol{\mu_u}, \Sigma_u)||\mathcal{N}(\boldsymbol{h}|\mu_0, K^h_{uu})) \end{aligned} \tag{6.18}$$

where the diagonal matrix $R_h \in \mathbb{R}^{N\times N}$ is defined as: $R_{ii} = \exp(\boldsymbol{\mu_{h}}_i - 0.5\Sigma_{hii})$, with the mean and variance

$$\boldsymbol{\mu_h} = \Omega^h_{Xu}(\boldsymbol{\mu_u} - \mu_0\boldsymbol{1}) + \mu_0\boldsymbol{1} \tag{6.19a}$$

$$\Sigma_h = K^h_{XX} - Q^h_{XX} + \Omega^h_{Xu}\Sigma_u(\Omega^h_{Xu})^T \tag{6.19b}$$

$$\boldsymbol{\mu_u} = K^h_{uX}(\Lambda - 0.5\mathbb{I})\boldsymbol{1} + \mu_0\boldsymbol{1} \tag{6.19c}$$

$$\Sigma_u^{-1} = (K^h_{uu})^{-1} + (\Omega^h_{Xu})^T\Lambda\Omega^h_{Xu} \tag{6.19d}$$

Here, Λ is a positive semi-definite diagonal matrix, and Ω_{Xu} is defined according to Equation 6.20. For details regarding the full derivation, the reader is referred to [303].

$$\Omega^h_{Xu} = K^h_{Xu}(K^h_{uu})^{-1} \tag{6.20}$$

The sparse approximation for the two GPs employed to model the heteroscedastic noise introduces several additional hyperparameters associated to the inducing points

used in $f(\boldsymbol{x})$ and $h(\boldsymbol{x})$. It should be noted that the number of inducing points does not necessarily have to be equal for both functions. Here, the covariance matrices are indexed by a superscript f or h, which denotes the function and corresponding hyperparameters under consideration. The subscripts are denoting which sets of points the covariance is computed between, with X being the full set of points and u being the set of inducing points for the corresponding function. Therefore, as an example, K^h_{Xu} is the covariance matrix between the training points and the inducing points for $h(\boldsymbol{x})$ given the hyperparameters of the kernel for the log-noise GP. As a result of the introduction of the two sets of inducing points, the number of hyperparameters has now increased to include the kernel hyperparameters for $k_f(\boldsymbol{x}, \boldsymbol{x'})$ and $k_h(\boldsymbol{x}, \boldsymbol{x'})$, the constant mean for the log noise variance, μ_0, the location of the m inducing points for $f(\boldsymbol{x})$, and the location of the u inducing points for $h(\boldsymbol{x})$, as well as the n variational parameters composing the Λ diagonal matrix.

At this point, the attention can shift towards making predictions with the VSHGP model. However, as with the non-sparse heteroscedastic GP, computing the predictive distribution $p(\boldsymbol{y}_\star|\boldsymbol{y}, \boldsymbol{x}_\star)$ at the test points $\boldsymbol{x}_\star$ requires the computation of an intractable integral. Analogous to the non-sparse case, [303] have derived an approximation for the first two moments of the predictive distribution using the Gauss-Hermite quadrature. Therefore, the resulting mean and variance, $\boldsymbol{\mu}_\star$ and $\boldsymbol{\sigma}^{\boldsymbol{2}}_\star$ respectively were found to be:

$$\boldsymbol{\mu}_\star = \boldsymbol{\mu}^f_\star, \quad \boldsymbol{\sigma}^2_\star = \boldsymbol{\sigma}^{f\,2}_\star + e^{\boldsymbol{\mu}^h_\star + \boldsymbol{\sigma}^{h\,2}_\star/2} \tag{6.21}$$

where,

$$\boldsymbol{\mu}^f_\star = K^f_{\star u} K_R^{-1} K^f_{uX} R_h^{-1} y, \tag{6.22a}$$

$$\boldsymbol{\sigma}^{f\,2}_\star = K^f_{\star\star} - K^f_{\star u}(K^f_{uu})^{-1} K^f_{u\star} + K^f_{\star u} K_R^{-1} k^f_{u\star}, \tag{6.22b}$$

$$\boldsymbol{\sigma}^{h\,2}_\star = K^h_{\star\star} - K^h_{\star u}(K^h_{uu})^{-1} K^h_{u\star} + K^h_{\star u} K_\Lambda^{-1} K^h_{u\star}, \tag{6.22c}$$

$$K_R = K^f_{uX} R_h^{-1} K^f_{Xu} + K^f_{uu} \tag{6.22d}$$

$$K_\Lambda = K^h_{uX} \Lambda^{-1} K^h_{Xu} + K^h_{uu} \tag{6.22e}$$

Here it is important to note that the correction term, $K^f_{\star u} K_R^{-1} K^f_{u\star}$ in Equation 6.22b contains the heteroscedasticity information form the noise term R_h. Finally, with

these definitions it is now possible to approximate the first two moments of the predictive distribution: $\boldsymbol{\mu}_\star$ and $\boldsymbol{\sigma}_\star^2$ respectively. In addition, [303] also improved the scalability of the VSHGP model presented above through the addition of the stochastic and distributed extensions. For the sake of brevity, the extensions to the VSHGP model are not duplicated here, instead the reader it referred to the original paper.

6.2.4 Hierarchical expansion

Revisiting the dataset presented in Figure 6.1, which consists of two known subgroups (namely HCs and patients affected by MS), this section proposes a hierarchical expansion of the VSHGP model presented in Section 6.2.3, expanding on the model previously introduced by Hensman et. al. in [294]. The key intuition of hierarchical modelling is that there exists a common trend across a pool of data from multiple individuals performing the same walking test, regardless of their group label. However, for the sake of simplicity, this section will only consider a two-layer hierarchy, i.e. a single group consisting of multiple individuals. As such, the initial emphasis is placed on modelling the overarching group-level dynamics, which contains the collective information from all individuals within that particular group. As a result, the shared-group pattern is being learnt initially. Subsequently, once the shared group pattern has been established, attention can shift towards the individual patterns, which are known to exhibit unique variations from the shared group trend, as a result of biomechanical differences, as well as corruptive noise. Therefore, as this stage, individual-specific variations are incorporated, building upon the shared group pattern while allowing for slight deviations to accommodate individual characteristics. This sequential refinement process enables the preservation of overarching trends while accommodating individual nuances, thereby fostering a comprehensive understanding of the dataset's hierarchical structure.

For clarity, this section will only present the hierarchical modelling approach using a standard GP. Integrating the variational sparse heteroscedastic extension in a hierarchical framework is straightforward, as it will be explained in the upcoming paragraphs, by considering block-wise relationships within the hierarchical covariance.

Let $\boldsymbol{y}_{gi}$ denote the vector of measurements for the i-th individual in group g. The corresponding time points are stored in vector $\boldsymbol{x}_{gi}$. To combine the data acquired

during all the walking tests from all participants in a particular group, a Bayesian hierarchical approach is being used. The underlying trend for the g-th group, $f_g(\boldsymbol{x})$ is presumed to be drawn from a GP with a zero-mean and whose covariance function is denoted by $k_g(\boldsymbol{x}, \boldsymbol{x'})$. Further down in the hierarchy, the underlying trend describing the gait pattern belonging to an unique participant in that particular group, $f_{gi}(\boldsymbol{x})$, are drawn from the group GP. However, the mean of the individual GP is $f_g(\boldsymbol{x})$, as described in Equation 6.23.

$$\begin{aligned} f_g(\boldsymbol{x}) &\sim \mathcal{GP}(0, k_g(\boldsymbol{x}, \boldsymbol{x'})) \\ f_{g,i}(\boldsymbol{x}) &\sim \mathcal{GP}(f_g(\boldsymbol{x}), k_i(\boldsymbol{x}, \boldsymbol{x'})) \\ where \;\; i &= 1, 2, \ldots, n \end{aligned} \tag{6.23}$$

It should be noted that the two covariance functions k_g and k_i used for the group and individual levels may be different. This model is graphically shown in Figure 6.5, where the function dependency is highlighted. The prior over the underlying group pattern $f_g(\boldsymbol{x})$ is shown at the top, as a dotted line. The shaded area represents the variance of the function, which is controlled by σ_g^2. The smoothness of the function is controlled by the length-scale of the group kernel, l_g. A single sample from this prior is then shown as red solid line and the length-scale of the covariance function is also highlighted. The individual level is shown in the second row, where samples conditioned on the sample shown in $f_g(\boldsymbol{x})$ are displayed, representing three unique individuals. The three samples follow the trend of $f_g(\boldsymbol{x})$, but are allowed to independently vary by a small amount (σ_i^2) with a short length-scale l_i. Therefore, although the main features of the common trend are preserved, each of the individuals exhibit their own characteristics. Finally, the hierarchical covariance matrix is shown at the bottom of Figure 6.5, demonstrating the block-wise relationship between individuals.

Let $\boldsymbol{Y_g} = \{\boldsymbol{y}_{gi}\}_{i=1}^n$ be the collection of noisy observations for n patients in group g and $\boldsymbol{X_g} = \{\boldsymbol{x}_{gi}\}_{i=1}^n$ the corresponding time points. Due to the conjugacy property of Gaussian distributions, the model described above can be mathematically represented as a joint Gaussian distribution, and it is possible to write down the likelihood as:

$$p(\boldsymbol{Y_g}|\boldsymbol{X_g}, \theta) \sim \mathcal{N}(\hat{\boldsymbol{y}}_{\boldsymbol{g}}|\boldsymbol{0}, \Sigma_g) \tag{6.24}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{g}} = \left[\boldsymbol{y}_{g,1}^T, \boldsymbol{y}_{g,2}^T, \ldots, \boldsymbol{y}_{g,n}^T\right]^T$ has been used to denote the row-wise concatenation

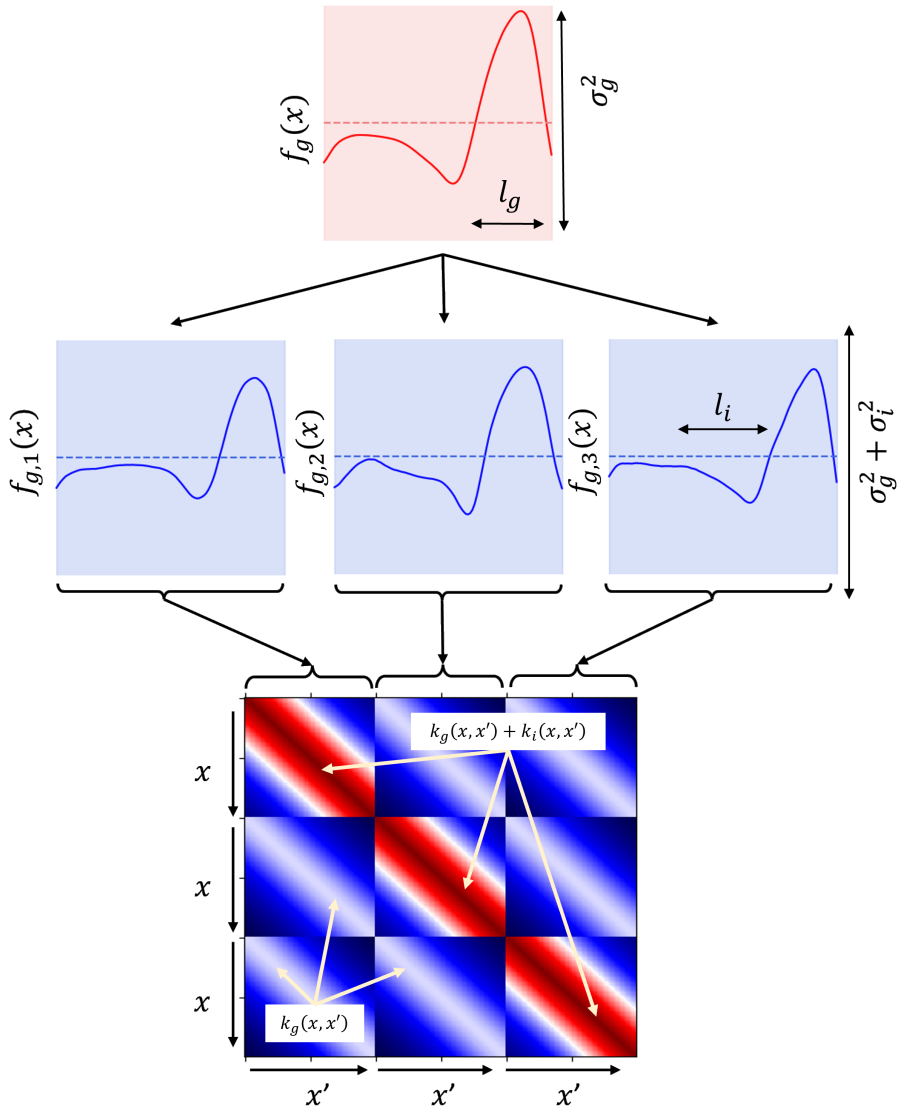


Figure 6.5: **An illustration of a simple hierarchical GP**. *Top*: solid line - a single sample from the prior over the underlying group function $f_g(x)$. Dotted line - zero-mean function. Shaded area: variance of the functions, controlled by σ_g^2 *Middle*: three samples conditioned on $f_g(x)$ and corresponding to three distinct individuals. The individual samples follow the trend of $f_g(x)$, but vary by a small amount, σ_i^2. The length-scale of the group and individual functions are denoted by l_g and l_i respectively. *Bottom*: block-wise covariance matrix used to generate samples.

of $\boldsymbol{Y_g}$. θ are the hyperparameters of the covariance functions $k_g(\cdot)$ and $k_i(\cdot)$. Finally, the block of Σ_g is given by:

$$\Sigma_g[i, i'] = \begin{cases} K_g(\boldsymbol{x}_{gi}, \boldsymbol{x}'_{gi}) + K_i(\boldsymbol{x}_{gi}, \boldsymbol{x}'_{gi}) + \sigma_n^2 \mathbb{I} & \text{if i = i'} \\ K_g(\boldsymbol{x}_{gi}, \boldsymbol{x}'_{gi}) & \text{otherwise.} \end{cases} \tag{6.25}$$

In order to make inferences about the functions $f_g(\boldsymbol{x})$ and $f_{g,i}(\boldsymbol{x})$, it is necessary to compute the covariances between the data $\boldsymbol{Y}$ and the functions. Employing the superscripted notation $\boldsymbol{y}_{gi}^{(x)}$ to represent the element of $\boldsymbol{y}_{gi}$ observed at input location x, the predictive covariance functions are given in Equation 6.26. This means that group predictions can be made simply by using the group kernel, $k_g(\cdot)$, whereas an additive kernel, $k_g(\cdot) + k_i(\cdot)$, is required for individual predictions.

$$cov\left(\boldsymbol{y}_{gi}^{(x)}, f_g(x')\right) = k_g(\boldsymbol{x}, \boldsymbol{x'}), \tag{6.26a}$$

$$cov\left(\boldsymbol{y}_{gi}^{(x)}, f_{gi'}(x')\right) = \begin{cases} k_g(\boldsymbol{x}, \boldsymbol{x'}) + k_i(\boldsymbol{x}, \boldsymbol{x'}) & \text{if i = i'} \\ k_g(\boldsymbol{x}, \boldsymbol{x'}) & \text{otherwise.} \end{cases} \tag{6.26b}$$

Finally, inference can then be made using the standard methods outlined in the preceding sections, while hyperparameters of the covariance functions can also be optimized in a similar fashion. However, the scalability of the hierarchical model is similar to that of a standard GP, limiting it's applicability to large datasets. To reduce the computational cost, variational approximation methods can be used, as in Sections 6.2.1 or 6.2.3, acknowledging that all the inducing points may have to be shared across all GPs in the hierarchy.

6.2.5 Assessing modelling performance

Having introduced the mathematical details for GP regression, the quality of the predictions can be established in several ways. Here, two methods are employed, namely the *Normalised Mean Squared Error* (NMSE) and the *Mean Standardised Log Loss* (MSLL), which is derived from the *Standardised Log Loss* (SLL), providing a probabilistic measure [166].

The NMSE is computed as:

$$NMSE = \frac{100}{n\sigma_y^2}(\boldsymbol{y}_\star - \boldsymbol{y})^T(\boldsymbol{y}_\star - \boldsymbol{y}) \tag{6.27}$$

where n is the sample size, σ_y^2 is the signal variance, $\boldsymbol{y}_\star$ is the model prediction and $\boldsymbol{y}$ is the true measured data. Here, a NMSE score of 0 is returned when model predictions precisely align with the target values, whilst a value of 100 corresponds to predicting with the mean of all observations. Therefore, the NMSE measures the averaged squared difference between the predicted and actual values normalised by the variance of the target variable. This returns accuracy of the model's predictions and provides an indication about how much the variance in the target variable is captured by the model.

Additionally, because GP predictions are effectively returned in the form of distributions, it is also feasible to compute the negative probability of a prediction under the model, often referred to as the model's loss. By standardising this value in relation to the mean and variance of the training set, the MSLL can then be computed as:

$$MSLL = \frac{1}{n}\sum_k \Big\{ -\log p(\boldsymbol{y}_{\star,\boldsymbol{k}}|X, \boldsymbol{y}, \boldsymbol{x}_{\star,\boldsymbol{k}}) + \log p(\boldsymbol{y}_{\star,\boldsymbol{k}}; \mathbb{E}(\boldsymbol{y}_{\boldsymbol{k}}), \mathbb{V}(\boldsymbol{y}_{\boldsymbol{k}})) \Big\} \tag{6.28}$$

where k indexes a particular test point, $\log p(\boldsymbol{y}_{\star,\boldsymbol{k}}|X, \boldsymbol{y}, \boldsymbol{x}_{\star,\boldsymbol{k}})$ is the log predictive likelihood of the model, $x_{\star,k}$ represents the test location, X is the set of training inputs and $\boldsymbol{y}$ are the training targets. Therefore, the MSLL is obtained by taking the average of the negative log likelihood over the test set and subtracting the trivial model which always predicts the mean and variance of the training set, therefore providing a quantitative metric for how well the model quantifies the uncertainty in predictions. A MSLL value of zero is associated with predicting with the training set mean and variance, and increasingly negative values indicate improved predictions [166]. By using both NMSE and MSLL, users can gain a better comprehensive understanding of the model's performance.

6.3 Modelling gait patterns using hierarchical GPs

6.3.1 Participants and initial data processing

The dataset used in this work consists of the shank angular velocity recordings from the baseline assessment presented in Chapter 5. Here, only a subset of 28 HC individuals and 28 MS-affected individuals were included for analysis, combining data from both A and B groups, as presented in Section 5.4.1. Specifically, from the HC individuals from Group A (those who performed the retest one hour apart from the baseline assessment), only 13 HC subjects were included here. Additionally, another 15 HC individuals were included from group B (those who performed the retest one week apart from the baseline assessment). This selection was performed as a result of inconsistencies in correctly identifying the heel strike events. Specific to MS-affected individuals, those utilising walking aids were automatically removed from the analysis in this chapter. As such, only 18 PwMS from group A were included, in addition to the 10 individuals from group B. The complete demographics details can be seen in Table 6.1.

Table 6.1: Chapter 6 demographics table.

		Age	**Gender**	**MS Subtypes**	
Group	**Participants**	*Mean (SD)*	*N male*	*RR*	*SP*
HC	n=28	39.2 (12.7)	13	-	-
MS	n=28 $\overline{EDSS} = 3.35$	47.7 (12.2)	8	18	10

RR = relapse remitting, SP = secondary progressive

Data segmentation into individual strides was performed following the peak-detection procedure detailed in [185]. However, as noted in Chapter 3, where inclusive outliers were detected specifically because of misclassification of gait events, this procedure is suboptimal. Therefore, in order to avoid problems caused by misclassification of the gait events [304], it was decided to remove the transient part at the beginning of the signals (first 8% of the samples). Additionally, a rotation to a vertical-horizontal coordinate system was applied, as described in [184], mitigating any sensor misalignment effects. Moreover, to ensure a device-agnostic methodology, a zero-phase, low-pass, Butterworth filter with a 10 Hz cut-off frequency was applied to the raw angular velocity signals, followed by scaling using a zero-mean, unit range normalisation method. Finally, for each individual limb, the resultant gait cycles

were normalised along the time axis, effectively eliminating the pace component from the signals and facilitating direct comparative analysis.

6.3.2 A case study to demonstrate Hierarchical Variational Sparse Heteroscedastic Gaussian Processes (HVSHGPs)

The aim of this case study is elucidate the intrinsic advantages of the novel Hierarhcial Variational Sparse Heteroscedastic Gaussian Process (HVSHGP) methodology for modelling the lower limb distal movement through the analysis the shank angular velocity. As such, this work proposes a novel modelling viewpoint which respects the natural hierarchical structure of the data collected from contralateral limbs, from multiple individuals, and from multiple groups of individuals forming a population. Moreover, the angular velocity across the complete gait cycle is modelled functionally, including its variation, as opposed to a set of summary features. The objectives of this case study are three-fold:

1. Considering the variability of the gait patterns in MS-affected individuals, the first objective is to showcase the added benefits of heteroscedastic noise modelling.

2. By verifying previously reported trends in joint kinematics, the second objective of this case study is to quantify group-level differences between HC and MS, followed by individual-level comparisons, relative to the HC group.

3. Finally, considering contralateral limb data collection and the time normalisation procedure used in this study, the third objective of this study is to showcase the utility of the probabilistic models by introducing a novel methodology for lower limb asymmetry quantification.

With regards to the third objective of this study, it is important to note that the concept of asymmetry presented here differs from the traditional understanding of gait asymmetry, as presented in the relevant literature, which is commonly assessed as the absolute difference in temporal metrics between contralateral limbs or as the natural logarithm of the absolute ratio between their mean values [66, 187]. Although the full motivation for alternative asymmetry metrics is postponed until Section 6.3.5, the new asymmetry metrics proposed in this study aim to quantify the degree

by which the functional form of the shank angular velocity differs between the left and right limbs, but also how much of the individual limb dynamics is explained by the combined model (i.e., aggregating data from both limbs).

Following pre-processing, the data are separated into two distinct sets: the *training* set and a *held-out* test set. The training set consists of 70% randomly selected samples from the aggregate dataset, i.e. combined data from all individuals, regardless of group affiliations. Only this set was used for *training* and hence the optimisation of the hyperparameters. The test set contains the remaining 30% of the data, meaning that these samples remained *unseen* until predictions were made. Following data partition, the training set comprised 1,942,881 data points, while the test set contained 832,664 points. It can be therefore seen that given the size of the datasets, it becomes unfeasible to fit standard GPs with any reasonable amount of computational resources.

The HVSHGP model, combining the hierarchical approach, together with sparsity and heteroscedasticity was implemented in Python, using GPflow [305]. As an implementation note, instead of constructing the block-wise covariance matrix, the underlying dependencies and interdependencies between different subjects and groups were modelled by incorporating a block-wise relationship into an additive $ELBO$, which was then optimized using the NADAM algorithm [219]. As a result, a four-layer HVSHGP model was achieved (see Figure 6.6 for the graphical representation). Here, each layer introduces new hyperparameters that need to be optimised. At the top layer of the hierarchy, the entire population is being modelled, in order to capture the overall trend shared across the entire dataset. The mean function at the group level is then conditioned on the population samples, and consequently, all the individual mean functions are then subsequently conditioned on their corresponding group samples. Finally, a fourth layer was introduced to address asymmetric gait patterns exhibited by contralateral limbs. Unless otherwise stated, all statistical comparisons performed in the subsequent sections of this chapter and the next chapter were performed using GraphPad Prism v.9.5.1 software (GraphPad Inc., La Jolla, CA, USA).

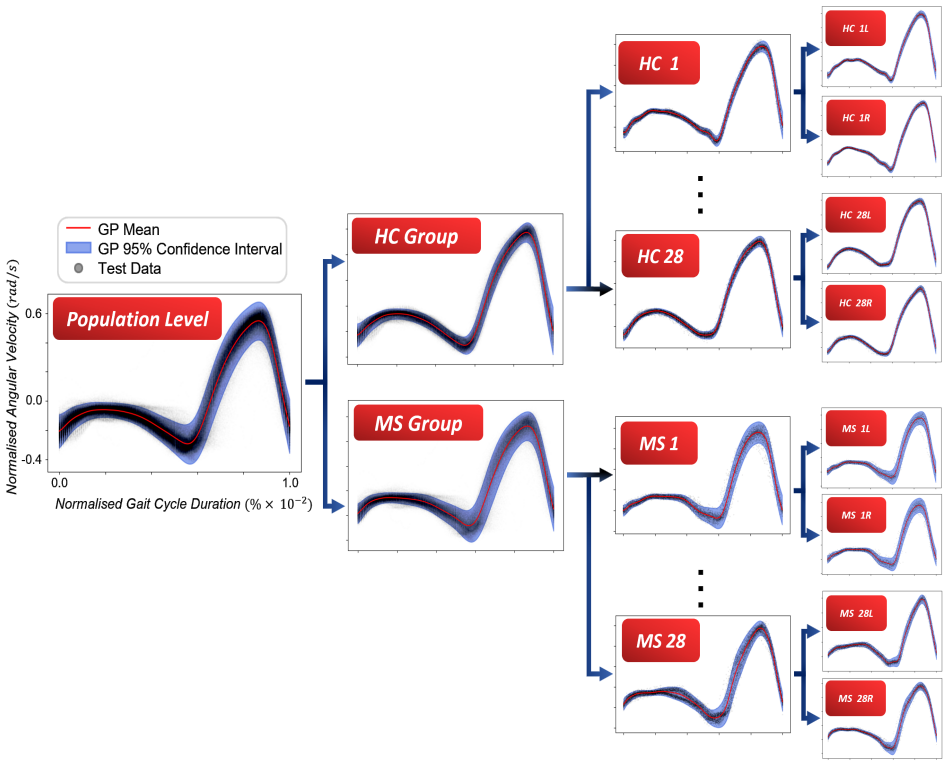


Figure 6.6: Hierarchical model structure: Progression from left to right includes the population layer, group layer (HC/MS), individual layer (combining contralateral limbs for specific individuals), and individual limb layer (independently modelling left and right limbs).

6.3.3 Comparative analysis of homoscedastic and heteroscedastic models for gait data

In this section, the first objective of the case study will be addressed, which involves showcasing benefits of modelling the non-constant variability of the shank angular velocity across the entire gait cycle, by integrating heteroscedasticity capabilitites into GP regression. To support the proposed modelling approach, group-level comparisons will be presented between a homoscedastic four-layer hierarchical variational sparse GP model (HVSGP) and the four-layer hierarchical variational sparse GP heteroscedastic model proposed in this work (i.e. the HVSHGP model).

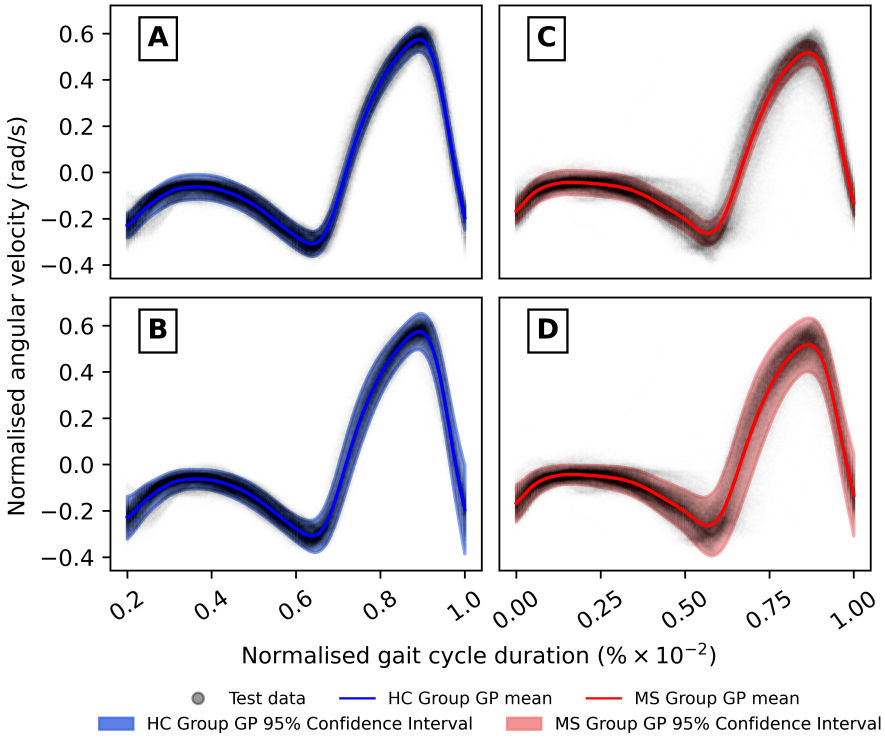


Figure 6.7: Group-level GP predictions against test data for homoscedastic models (first row) and heteroscedastic models (second row). **A** - HC Homoscedastic model, **B** - HC Heteroscedastic model, **C** - Homoscedastic MS model, **D** - Heteroscedastic MS Model.

The comparison of the model predictions against the test set can be seen in Figure 6.7, for both HC and MS groups, where the first row of predictions belong to the homoscedastic model, and those on the second row belong to the heteroscedastic model. Qualitatively, it can be seen that the homoscedastic model has failed to capture the confidence interval representative of the data, especially in the case of the MS group, where many test datapoints lie outside the confidence interval. Particularly, the increased variability in the gait signals is not captured adequately during the swing phase of the gait cycle, which accounts for approximately 33 to 38% of the gait cycle [17, 31, 36]. According to [202], this behaviour may be likely a result of unstable equilibrium during the coordination of the swing and forward movement. Here, the heteroscedastic model further expands the confidence interval of the GP and allows for better predictions across the entire duration of the gait cycle. Even though both models appear to be comparable during the stance phase, the heteroscedastic model also improves the predictions towards the termination of the stance, i.e., during the double support phase.

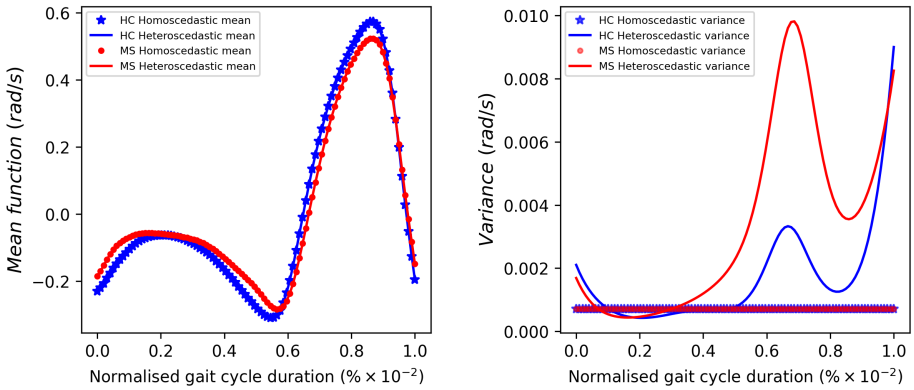


Figure 6.8: Mean and variance differences between the homoscedastic and heteroscedastic models.

The improvement of the heteroscedastic model, relative to the homoscedastic model is most evident when examining the predictive mean and variance of the models for both HC and MS. This comparison can be seen in Figure 6.8. Considering only the means, a notable resemblance is noted between the homoscedastic and heteroscedastic models, regardless of group affiliations. This behaviour was perhaps expected, given that the predictive mean of the heteroscedastic model has a very similar formulation to the predictive mean equation for the homoscedastic case. However, the key distinction between the models, from a predictive standpoint, lies in the computation of the variance of the predictive density, particularly preceding the onset of the swing phase. Additionally, an increased variance magnitude is evident for the MS group, exceeding that of the HC group. Consequently, the variability arising from uncontrolled coordination during the swing phase of the MS group is effectively captured within the confidence bounds of the heteroscedastic model, contrasting the homoscedastic alternative. Therefore, it is clear that the integration of heteroscedastic modelling capabilities into GP regression leads to a more accurate representation of group-level gait patterns in patahological gait. The performance metrics of both the homoscedastic and heteroscedastic models are presented in Table 6.2.

Table 6.2: Performance comparison of homoscedastic and heteroscedastic models in group-level predictions.

	NMSE (%) **Train**		NMSE (%) **Test**		MSLL **Train**		MSLL **Test**	
	HC	MS	HC	MS	HC	MS	HC	MS
Homoscedastic	2.869	7.403	2.873	7.397	-1.305	0.532	-1.305	0.527
Heteroscedastic	2.869	7.403	2.873	7.397	-1.494	-1.520	-1.493	-1.523

From Table 6.2, it is evident that there are minimal distinctions in model performance between the training and test sets. This highlights the robustness and generalisation capability of the GP modelling procedure applied to model the functional form of the shank angular velocity. Subsequently, upon analysis of the NMSE values, it is unsurprising that both modelling strategies yielded similar results. This is because the NMSE is independent of the predictive variance of the model. Thus, solely relying on this performance metric might lead to a misleading interpretation. However, upon examining the MSLL values, the improved uncertainty quantification of the heteroscedastic model is evident, as indicated by the negative MSLL value. While the addition of heteroscedastic noise modelling led to a modest enhancement in predictive performance for the HC group, the MS group exhibited a notably more significant improvement. This underscores the efficacy of the heteroscedastic approach, particularly for pathological populations where accurate uncertainty estimation is paramount. As such, accounting for the non-constant variance along the gait cycle is crucial for maintaining high sensitivity when deploying this modelling strategy for longitudinal gait monitoring purposes in PwMS.

6.3.4 Comparative analysis between HCs and PwMS: group and individual-level comparisons

Drawing from the validated utility of heteroscedastic noise modelling, this section targets the second objective of the case study, aiming to present a robust methodological foundation for systematically examining and quantifying the differences in shank angular velocity patterns across multiple scales. Initially, the investigation between group-level discrepancies between HC and MS is presented. This analysis aims to establish a comprehensive understanding of the model's capacity to capture the inherent variability in MS gait dynamics. Moreover, this comparative analysis acts as

a validation point, confirming previously reported trends in the literature. Expanding upon the hierarchical structure of the model, in contrast to conventional approaches that predominantely focus on group-level distinctions, the analysis exploration extends to facilitate nuanced comparisons at lower hierarchical levels, allowing for individual-level comparisons.

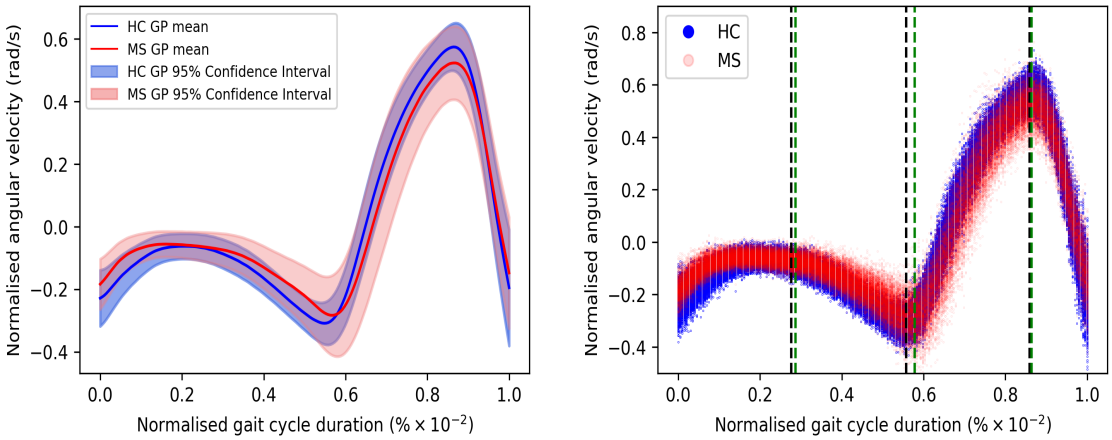


Figure 6.9: Left: GP group predictions, Right: Samples drawn from the GPs. The vertical dotted lines, from left to right, correspond to the mid-stance, toe-off and mid-swing events. The black line corresponds to the HC group, while the green line corresponds to the MS group.

Regarding the HC and MS group predictions, the achieved NMSE scores on the test data are 2.873 and 7.397 respectively. It can be seen that the HC group model had a marginal decrease in point-wise error. This outcome may stem from the presence of outliers outside the confidence bounds of the GP prediction in the MS group, particularly preceding the onset of the swing phase. Intriguingly, the MSLL scores for both group predictions were relatively similar, achieving values of -1.493 for the HC group prediction and -1.523 for the MS group prediction. Moving on, it proves more beneficial to concentrate on the specific regions in the gait cycle that exhibit the most significant differences.

Considering Figure 6.9, one way of quantifying the differences is by computing the unbiased formulation of the *Maximum Mean Discrepancy* (MMD) [247] at each test location between samples generated from the predictive distributions of both groups, according to Equation 5.7. The mathematical details regarding the MMD computation have been provided in Section 5.3.2. The advantage of employing the MMD lies in its formulation as a non-parametric distance metric, which employs kernel embeddings of the distributions under comparison. This is advantageous, as

the kernel trick allows the end user to effectively assess infinite moments through the use of inner products in a feature space [212]. In this context, the radial basis function (RBF) [269] is employed as the chosen kernel, defined according to Equation 5.8, where the σ parameter governing the bandwidth of the kernel is established as the median distance between points within the aggregate sample [247].

The predictive distributions are visible on the left of Figure 6.9. On the right, 500 samples are generated at each of the 200 equidistant test points along the input space. Figure 6.10 displays the comparison between the predictive distributions of the two groups, where the MMD test statistic values are normalized between 0 and 1. While interpreting absolute value of the MMD is challenging, lower values indicate greater similarity between distributions, while higher values signify increased disparities between the models. Furthermore, solid white lines represent absolute differences in mean predictions, while dashed white lines depict differences in confidence intervals, facilitating comparison. Despite noticeable similarities observed in certain regions of the gait cycle (highlighted by the black and dark blue regions), specific locations stand out as being dissimilar. These regions include the first 15%, the range of 35 to 55% and the range of 60 to 90% of the gait cycle. The first two highlighted regions coincide with the double support phases, where both feet are in contact with the ground [306], while the most significant disparities are prominent during the onset of the swing phase.

Comparison with other findings in the relevant literature poses a challenge, as most characterisations of pathological gait predominantly focus on joint kinematics rather than segment kinematics. Nevertheless, in the case of the MS group, there is an observable progressive variability in the gait pattern from mid-stance to mid-swing. Consistent with the outcomes of this study, other researchers have also reported greater variability in joint angles among individuals with MS, even among those with mild disability [159, 240, 288, 289]. Similar trends have been found by [290], where significant differences between HC and MS were reported during the stance and swing phases when considering deviation phase as a measure of coordination variability. Additionally, the flatter trajectory observed during the stance phase may suggest a reduced range of plantar flexion, as reported by [159]. This reduction can result in diminished power generation at the ankle, particularly during the onset of the swing phase, potentially affecting the stability control throughout this latter phase. Such outcomes might be attributed to factors such as muscle weakness, spasticity, fatigue, or balance impairments [159, 194]. Therefore, while confirming author's

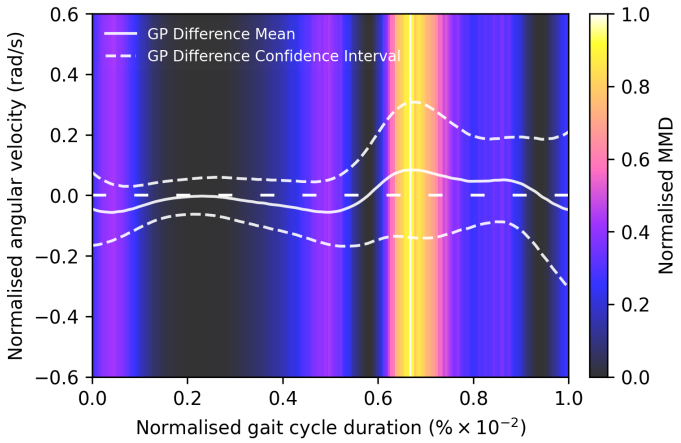


Figure 6.10: Visualisation of the group differences: The solid white line represents the difference between the means of the two GPs, whereas the dotted lines illustrate the differences in confidence intervals. Notably, although the confidence interval may not seem symmetric above and below the mean upon initial observation, it is, in fact, symmetric. Disparities between the HC and MS models are highlighted across the input space using the normalised MMD.

expectations, the comparative results between HC and MS groups underscore the HVSGP model's capability to effectively capture the inherent variability in MS gait patterns, facilitating an informed analysis.

Next, to showcase the capabilities of the proposed hierarchical modelling approach to facilitate nuanced comparisons at lower levels in the hierarchy, individual gait pattern comparisons have been investigated, relative to the control group. This analysis aligns with the overarching aim of providing personalised assessments. Detailed predictions of shank angular velocity were generated for all participants in the study, as depicted in Figure 6.11, where the held-out test data has been overlaid.

The individual NMSE and MSLL scores are provided in Appendix B.1. Notably, individuals with MS exhibited significantly higher NMSE scores, indicating lower predictive capabilities of the MS model. Not surprisingly, the MS group also displayed significantly higher MSLL scores, further supporting the diminished predictive performance of the model for this population, given the increased variability in the gait patterns as well as the presence of asymmetry. To highlight individual level discrepancies, samples drawn from the HC group predictive distribution were compared to samples drawn from the individual GPs, in a similar way to the MMD-based approach used above. Specifically, 500 samples were generated again at each of the 200 equidistant test locations along the input space, and the MMD was computed

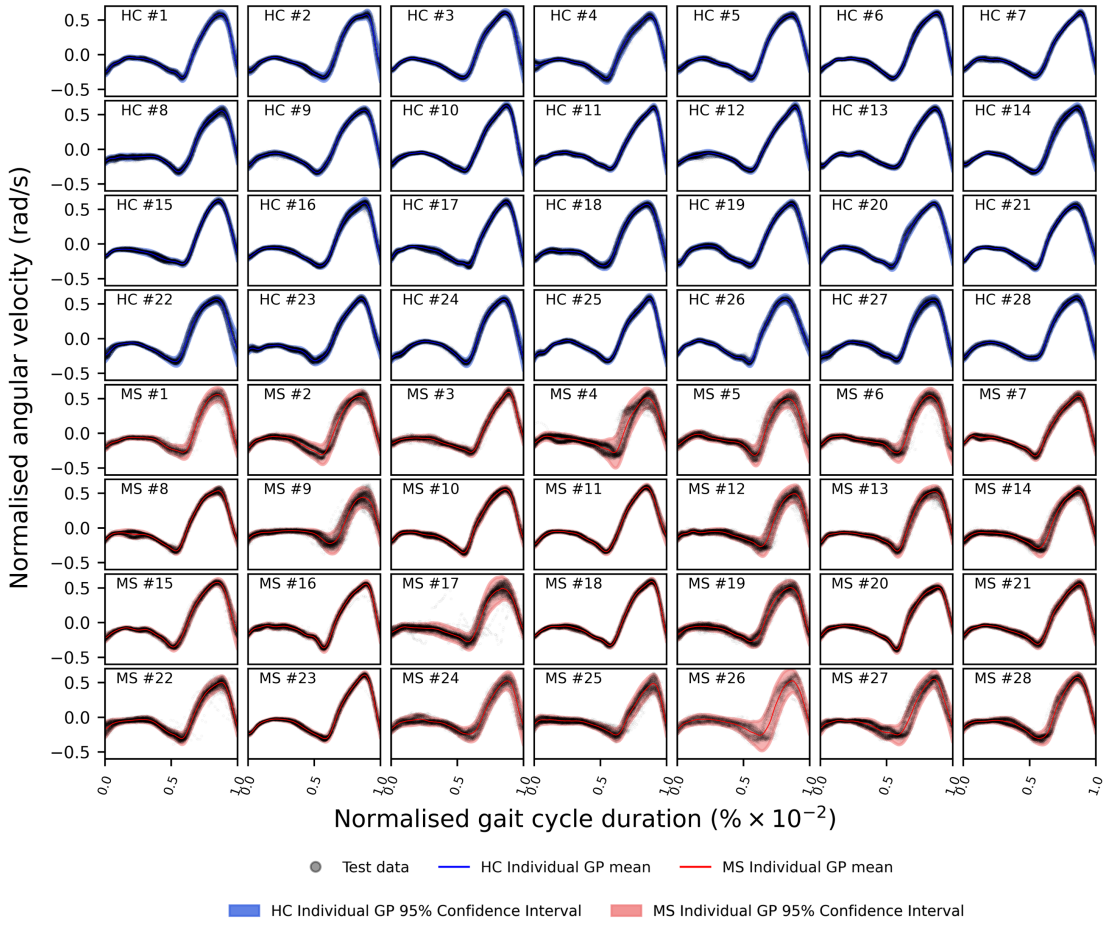


Figure 6.11: Comparison of individual predictions and *held-out* test data: the first four rows represent predictions for HC individuals, while the last four rows depict predictions for individuals affected by MS.

at every test location, for each subject. Then, the MMD values from all subjects were aggregated and normalised between 0 and 1, highlighting regions in the gait cycle where differences relative to the control group are most prominent, as shown in Figure 6.12. It is important to acknowledge that solely comparing individual-level predictions of MS indiviudals with the predictive distribution of the HC group is insufficient. This limitation arises due to the presence of overlapping confidence intervals and comparable means observed between the HC group and MS-affected individuals. Such similarity arises from the increased variability evident in individual predictions within the MS group.

Figure 6.13 showcases the individual heteroscedastic variance, revealing deviations from the typical healthy control (HC) gait patterns among several individuals affected by multiple sclerosis (MS), especially towards the end of the stance phase and during

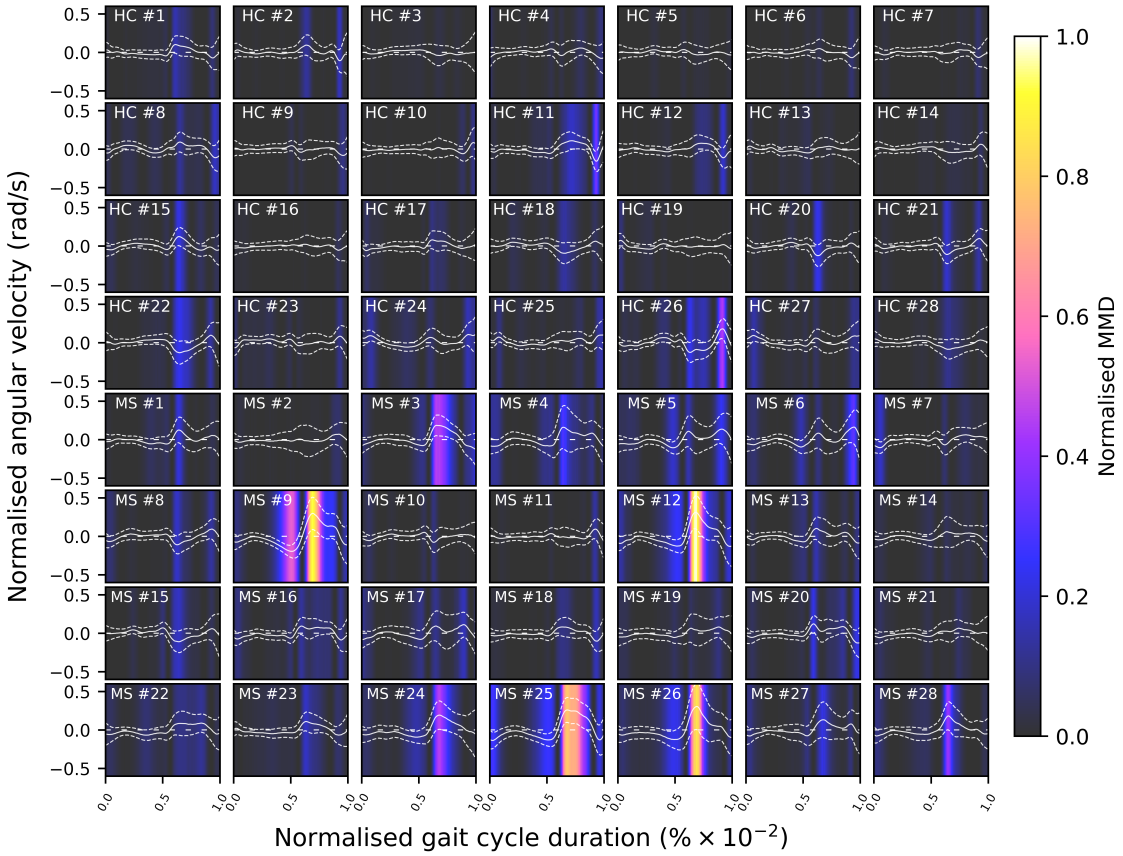


Figure 6.12: Individual differences highlighting the MMD. The first 4 rows correspond to HC individuals, while the last four rows correspond to MS-affected individuals.

the swing phase. Consequently, the extension of the hierarchical methodology to model individual gait patterns introduced a novel feature — the heteroscedastic variance — that may serve as an indicator of neurological or musculoskeletal deficits. Two primary factors likely contribute to these findings. The first one might associated with the possible lack of control during the swing phase, as a result of muscle spasticity, fatigue or balance impairments [159, 194], while the second one relates to the presence of asymmetry. The concept of asymmetry will be elaborated upon in subsequent paragraphs, where the hierarchical model has been further expanded to separately model the left and right limbs. This extension has been added since the distribution over the outputs must be symmetric about the mean, which is not representative of the bi-modal distribution when asymmetry is present. However, the asymmetric gait patterns are still captured within the confidence bounds of the GP predictions. Thus, the variance predicted throughout the gait cycle could hold significant potential for integration into clinical gait analysis, noting that this methodology is not intended

as a replacement for standard gait analysis, but rather as a valuable augmentation to the clinical gait assessment of MS-affected individuals.

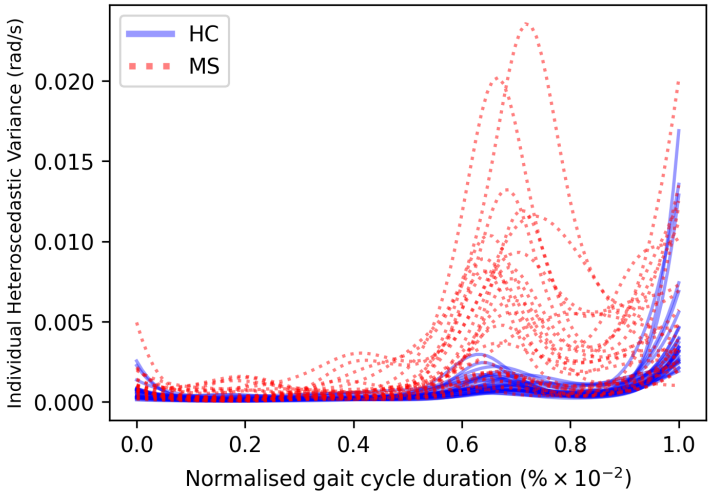


Figure 6.13: Comparison of the predicted heteroscedastic variance across HC and MS individuals. Each line corresponds to a single subject.

6.3.5 A novel proposal for gait asymmetry quantification

In light of the results presented in Section 6.3.4, where the predictive heteroscedastic variance emerged as a discerning factor between HCs and MS-affected individuals, and accounting for the data normalisation procedure used in this study, this section addresses the third objective of the case study, specifically focusing on proposing a novel methodology for quantifying asymmetry. Recognizing asymmetry as one of the key factors for the increased variability in the gait patterns when assessing the individual-level models (i.e. those combining data from left and right limbs), the hierarchical model has been extended to separately address contralateral limbs. This extension not only facilitates a detailed exploration of contralateral limb data but also lays the groundwork for introducing a novel methodology for quantifying lower limb asymmetry from a probabilistic standpoint.

Traditionally, gait asymmetry has been defined as the absolute difference between the temporal metrics extracted from the left and right lower limbs [66] or as the natural logarithm of the absolute ratio of the shorter and the longer mean value of the temporal metric [187]. Gait asymmetry is often regarded as an indicator of MS [17]. However, within the context of this study, the term ‘asymmetry’ is used to

denote non-overlapping confidence intervals or a clear trend of one signal pattern significantly exceeding or falling below the other. To provide enhanced visual insights into the presence of asymmetry, individual predictions for the left and right limbs are presented in Figure 6.14, where the unseen held-out test data has also been overlaid.

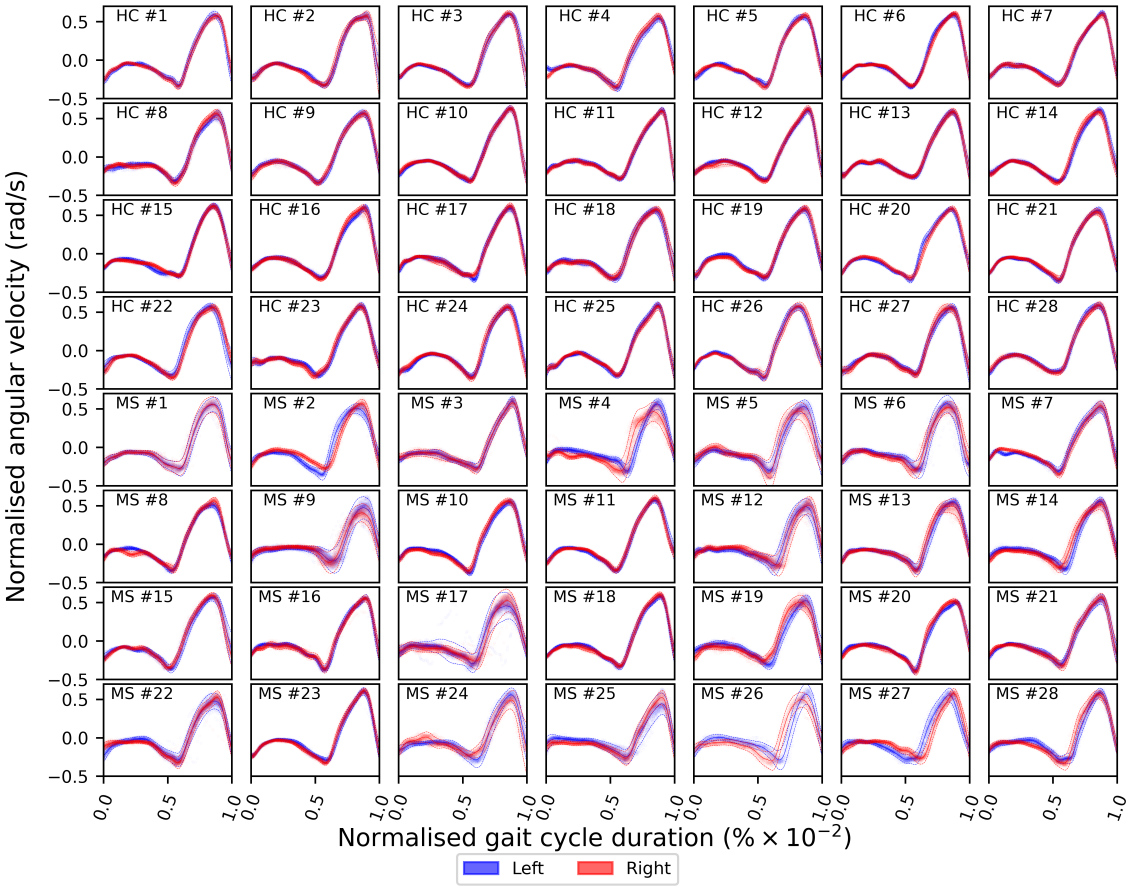


Figure 6.14: Individual limb GP predictions. The first four rows correspond to HCs, while the last four rows correspond to MS-affected individuals. The solid lines represent the mean GP predictions, while the dotted lines correspond to the 95% confidence interval. The dots represent the held-out test data

Examining Figure 6.14, discernible disparities emerge in the gait patterns observed across contralateral limbs for PwMS, with several subjects manifesting notable deviations. Conversely, the gait patterns observed among the HC individuals demonstrate a higher degree of symmetry, i.e. the same functional form throughput the gait cycle, despite the sparse presence of slight deviations between the left and right shanks, which are deemed negligible here. Thus, in order to quantify gait asymmetry, a comprehensive methodology is necessarily. Initially, the Wasserstein distance (WD) [307], has been used to quantify the dissimilarity between probability

distributions associated with gait patterns of the left and right limbs, as modelled by the fourth layer in the HVSHGP model. This metric is intuitive, as it effectively measures the minimum cost of transforming one probability distribution to another, for example the probability distribution of the GP predictions from the left leg to the right one through the entire gait cycle. In the context of this work, a smaller WD vale would suggest that the gait patterns are more similar, while a larger WD value would denote greater dissimilarity. Similarly to the computation of the MMD, 500 samples were drawn from each GP's predictive distributions at 100 equidistant points spread across the input space. Subsequently, the WD was computed at each test location (see Appendix B.2, Figure B.1), enabling end-users to visually identify prominent locations within the gait cycle where asymmetry is pronounced. Moreover, the overall effect of gait asymmetry was quantified by computing the area under the curve, thus providing an unified gait asymmetry metric. The comparison of the individual differences for all subjects included in the study is presented in Figure 6.15 (a) and (b).

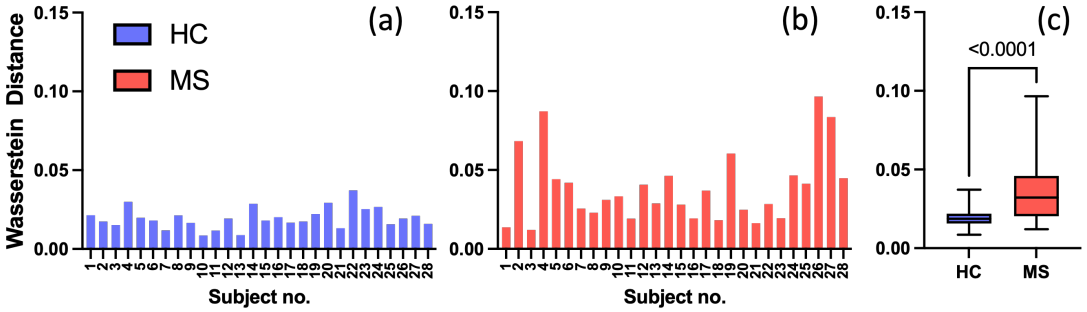


Figure 6.15: (a) & (b) - Unified Wasserstein distance computed between samples drawn from the left GP and the right one for each of the individuals in both HC and MS groups. (c) - Statistical comparison between the HC and MS groups.

Here, the PwMS individuals numbered 2, 4 , 19, 26 and 27 particularly stand out. Subject 2 displayed a reduced range of mobility on the right shank during the stance phase, as well as an increased stance duration, when compared to the left leg. An uncontrolled movement was recorded on the right leg of subject 4. For this subject, the one-sided balance and coordination deficits were evident across the entire gait cycle, albeit more pronounced during the swing phase. Subject 19 also displayed temporal differences between the left and right limbs. A reduced range of mobility was also recorded for subject 26, for the right shank. This was recorded in conjunction with temporal differences between the two limbs, as well as increased

variability during the swing phase and the end of the double support phase. Lastly, subject 27 manifested an uncontrolled movement in the right shank, along with temporal differences and higher movement variability in the left leg. To summarise the asymmetry disparities between the HC and MS groups, the non-parametric Mann-Whitney U test was conducted, with a significance level set at 5%, revealing statistically significant differences between the two groups ($p < 0.0001$, see to Figure 6.15 (c)). The relationship between the WD and clinical scores, as represented by the EDSS score, was investigated. A weak correlation was observed (Pearson's $r = 0.33$), indicating that asymmetry in MS may manifest independently of disease severity. Despite some arguments proposing a link between asymmetry and disease severity, such as Pau et al. [286], which demonstrated moderate correlations between joint kinematics asymmetry and the EDSS score using trend symmetry [288] the findings of this study emphasise asymmetry relevance across various levels of MS disability. However, because of the considerably lower number of subjects included in the present study, drawing definitive clinical conclusions is probably not advisable.

Secondly, an additional dimension of asymmetry analysis was pursued using the KL divergence. This procedure entailed comparing the third-layer of the HVSHGP model (considering both limbs collectively) with separate fourth-layer GP models focusing on the left and right limbs individually. This comparison is depicted in Figure 6.16 (a) and (b). The KL divergence provides insight into the extent to which the combined third-layer model accurately represents the dynamics of the fourth-layer models consisting of individual limbs. A high KL divergence indicates that the combined model might not fully capture the nuances of each of the individual limbs, potentially reflecting asymmetry or distinctive characteristics. It should be noted that computing the KL divergence involves treating the GPs as multivariate Gaussian distributions, and requires access to the full covariance matrix. In contrast to the previous asymmetry analysis utilising the WD, examination of the KL divergence metrics in Figure 6.16 (b) reveals two additional MS subjects of interest, namely subjects 10 and 14. Although the outcomes for subject 10 may not be readily discernible, this result can be attributed to subtle disparities between the left and right gait trajectories, which may not be apparent upon visual inspection. Similar findings were observed for subject 14, where the presence of asymmetry was notably pronounced. The statistical comparison is presented in 6.16 (c). The non-parametric Mann-Whitney U test was also employed in this case, with a minimum significance alpha level of 5%, highlighting statistically significant differences between the two groups ($p = 0.0029$). However, no correlations were found between the KL-based

asymmetry metrics and the EDSS score ($Pearson's\ r = 0.1015$ and 0.0051 for the left and right shanks respectively), necessitating additional validation procedures before drawing any definitive clinical conclusions. However, the extension of the hierarchical model to individually consider contralateral limbs provided new insights into asymmetry levels in individuals affected by MS through a novel, multifaceted approach employing both the Wasserstein distance and the KL divergence. It is important to note that while the results presented here may suggest that asymmetry could be a challenging aspect of MS even in the early disease stages, caution should be exercised due to the small sample size and the need for further validation. Nonetheless, these insights may hold promise in assisting end-users toward providing improved personalized treatment or rehabilitation plans.

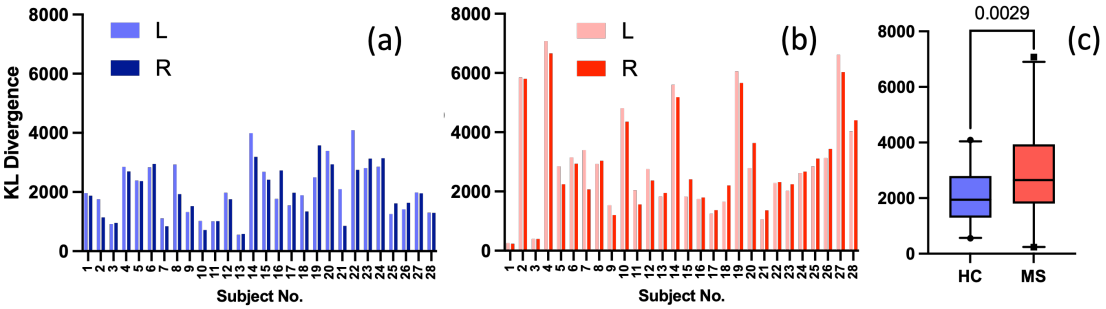


Figure 6.16: KL divergence computed between the individual level GP, combining the left and right limbs (third layer), and the individual limbs GP models, treating each limb individually (fourth layer).(a)- HC, (b) - MS. The lighter shades represent the left shank, while the darker one represents the right shank. (c) - statistical comparison.

6.4 Conclusions

The work presented in this chapter targeted the final level of the proposed hierarchical framework for assessing the condition of PwMS, aiming to investigate the longitudinal progression of the disease. Given the relevancy of the shank angular velocity for this pathological population, this chapter proposed a novel probabilistic, Bayesian hierarchical modelling perspective. The proposed approach makes a departure from understanding gait with respect to a set of summary features. Instead, it models the functional form of the shank angular velocity, including its varying uncertainty throughout the entire gait cycle. In addition, recongising the common trend across a pool of data from multiple individuals performing the same walking test, regardless

of their group label, the model proposed here obeys the underlying hierarchical structure of the data, exploiting similarities but respecting differences.

The novel Hierarchical Variational Sparse Heteroscedastic Gaussian Porcess (HVSHGP) model proposed in this chapter combined three main ideas. Firstly, the hierarchical structure employed here directly models the natural organisation of data, progressing from contralateral limbs to individuals and then to groups of individuals forming a population. This framework accounts for the temporally structured covariance between different levels of groupings, maximising the richness of available information. Then, the problem of scaling GPs to handle large datasets, such as the ones obtained during clinical gait assessments, has been addressed by using variational approximation methods. Finally, the variability in the gait cycle has been captured by input-dependent noise modelling, i.e. heteroscedasticity. This was achieved by modelling the log-noise variance of the process as an additional GP, which is also learnt in a variational manner. While the combination of heteroscedasticity and sparse inference has been previsouly addressed in the machine learning literature, to the best of the author's knowledge, the additional integration of hierarchical modelling presents a novelty element.

Although clinical recommendations are probably not advisable, this study highlighted some potential significant differences between the HC and MS groups, particularly evident during the swing phase of the gait cycle and towards the end of the stance phase, aligning with previously reported trends in joint kinematics. Additionally, the usefulness of the methodology was demonstrated though its granular analysis capabilities. The findings of this work indicate that the characteristic feature that distinguishes anomalous gait patterns in individuals with MS is the presence of heteroscedastic variability throughout the gait cycle, and most notably, the amplitude of the log-noise variance. Moreover, a novel understanding of gait asymmetry between the lower limbs has also been presented, suggesting that its presence may not be strictly conditioned upon the severity of the disease. The potential utility of the novel probabilistic framework introduced in this study for gait analysis across multiple scales represents an important outcome deserving immediate attention by the research community, especially because this approach can be further extrapolated for the analysis of several other kinematic signals. Finally, the most natural continuation of this work is the exploration of the generality of the model at follow-up assessments. This aspect is explored in the next chapter of this thesis, which aims to validate the proposed HVSHGP model within a time frame of (assumed) constant disease status.

If successful, it will be then used for longtudinal monitoring and prognosis.

Chapter 7

An exploration of longitudinal gait data

One of the biggest limitation of all major studies involving MS-affected individuals is that the majority are overwhelmingly based on observational cross-sectional data, thereby preventing a clear understanding of disease progression over time [1, 45]. While these studies offer valuable snapshots of disease states, they are inherently limited by their ability to capture the dynamic nature of MS progression. The lack of longitudinal data hinders the ability of researchers to predict disease trajectories and develop personalised treatment plans. This trend, however, appears to be shifting favourably, evidenced by the increasing prevalence of longitudinal studies emerging in the literature [16, 45, 49, 50]. Nonetheless, even these advancements face challenges. One critical obstacle lies in ensuring good dataset coverage. Machine learning algorithms that are often employed in gait analysis are data-driven. This means that their performance hinges on the quality and comprehensiveness of the training data. As such, models trained using baseline assessment data, for instance might perform well for short-term regression tasks, within an assumed stable disease state. However, their ability to predict longitudinal changes, where the disease state evolves, remains an open question. For this reason, this chapter represents a continuation of the work presented in the previous two chapters, and showcases an exploratory study into the longitudinal changes in the gait patterns associated with the lower limb distal motion.

Building upon the limitations discussed above and the findings presented in Chapter 5,

this chapter delves into evaluating the applicability of hierarchical GP modelling of the shank angular velocity for longitudinal assessments in MS. Chapter 5 highlighted the challenges associated with the quantification of subtle longitudinal changes in gait patterns, primarily due to potential confounding factors. These factors, such as variations in sensor placement, timing of assessments in the presence of medications, among others, can induce variability that may mask the true underlying changes in gait patterns related to disease progression. The complexity of longitudinal gait analysis in individuals with MS is thus underscored by this variability. Given these limitations, this chapter presents a two-part approach to assess the suitability of GP modeling for the task. In doing so, it serves as a case study for further exploration of longitudinal assessments and underscores the potential of integrating prior clinical knowledge into data-driven models through physics-informed machine learning approaches.

In the first part of this chapter, before presenting the exploratory longitudinal study, it is essential to ascertain whether the hierarchical GP modelling technique presented in Chapter 6 possesses sufficient generalizability within a short time frame. Generalizability, in this context, refers to the model's ability to produce accurate predictions on *unseen* data and avoid overfitting. Such a model is essential to aid end-users in distinguishing between the inherent variability of gait patterns and the impacts of disease progression or treatment interventions. Similar to the approach presented in Chapter 5, this property is evaluated on follow-up data collected at either one-hour or one-week after the baseline assessment. This time frame ensures minimal changes in disease state for the MS subjects. Importantly, the follow-up data remains completely unseen by the model during training, functioning effectively as a test set.

The outcomes of the assessment presented in the first part of this chapter will then serve as the foundation for the subsequent longitudinal assessment in the second part of this chapter. If the hierarchical GP modelling approach demonstrates sufficient generalizability by accurately predicting the shank angular velocity on the completely unseen follow-up data, then it would be natural to explore the utility of the GP models established at the baseline assessment for longitudinal monitoring purposes. Within this context, it is proposed to explore how the longitudinal MS patterns compare to a typical HC gait pattern and also investigate the temporal relationship between predictive modelling outcomes, asymmetry and disease severity. The rationale for prioritising the generalizability assessment lies in the inherent advantages offered

by the GP models. As highlighted in Chapter 5, traditional ARX-type models, which were previously employed, are limited by their reliance on a specific (linear) functional form. Contrarily, GP models present several advantages over ARX models. Specifically, GP models represent entire distributions over functions, enabling them to capture a broad spectrum of complex relationships between inputs and outputs. Additionally, they provide automatic uncertainty quantification, further enhancing their utility. This two-part approach facilitates the systematic evaluation of the suitability of GP models for longitudinal gait analysis, and ultimately contributes to a more accurate understanding of disease progression.

7.1 Part 1: Follow-up consistency check for the HVSHGP modelling approach

The first part of this chapter evaluates the generalizability of the hierarchical variational sparse heteroscedastic Gaussian process (HVSHGP) modelling approach presented in Chapter 6. The completely unseen test data consists of follow-up assessment data from the same participants used for training the model in Chapter 6, i.e. 28 HC individuals and 28 PwMS. Analogous to Chapter 5, these two groups were further divided into two subgroups of subjects: those who underwent retesting after a period of one hour (Group A), and those who performed it after a seven-day interval (Grroup B). Accordingly, 13 HCs underwent retesting one-hour apart, while 15 HCs underwent retesting one week apart. Additionally, 18 PwMS underwent retesting one-hour apart, whereas the remaining 10 MS participants were retested after a seven-day interval. For clarifications, the one-week interval between assessments enabled the novel GP modelling approach to be evaluated through a period of stable disease status, but also under more realistic follow-up assessment scenarios, where changes in sensor placement, differences in the time of assessment, prior physical activity before testing, differences in subject's footwear, and other factors might occur.

To begin with, group-level comparisons will be firstly presented here. As such, NMSE and MSLL metrics obtained using the baseline test (BT) held-out data and the new, completely unseen follow-up (FU) assessment data have been compared. The NMSE and MSLL were defined in Section 6.2.5 of this thesis. Firstly, the metrics describing the group comparison can be found in Table 7.1. For clarity, instead of aggregating

		HC		**MS**	
		NMSE(%)	MSLL	NMSE(%)	MSLL
Group A	BT	3.254	-1.480	9.381	-1.367
	FU	2.741	-1.498	10.527	-1.251
Group B	BT	2.489	-1.506	6.591	-1.591
	FU	2.172	-1.526	6.069	-1.629

Table 7.1: Group-level performance metrics. *Group A* - the one-hour apart retest group; *Group B* - The one-week apart retest group.

the one-hour apart and one-week apart datasets, it was rather decided to present these separately. Observing the results, it appears that the HC group exhibits a slight enhancement in predictive performance at the FU assessment, regardless of the interval between the initial BT and the subsequent FU assessment. Conversely, the MS model demonstrates only a minor decline in predictive performance when the FU assessment has been performed one-hour apart from the BT. Nonetheless, it is important to note that the differences between the BT and the FU assessment are considered to be heuristically insignificant for any practical implications. In fact, it can be seen that the group-level models yield better performance on the FU test set. Based on these findings, it can be concluded that the group-level GP models avoid the risk of overfitting, and may represent a feasible modelling strategy for longitudinal modelling of the gait patterns.

Next, the generality of the model has also been evaluated using the third level of the hierarchy (combining contralateral limb data), for all the HC and MS-affected individuals. To reiterate, within the context of MS gait analysis, achieving generalization refers to a model's capability of maintaining good predictive performance on unseen data after being established using BT data. This necessitates capturing the inherent within-subject variability present in MS patients. In simpler terms, the model should be robust to the unforeseen fluctuations in gait patterns experienced by individuals with the disease. To aid visualisation, the unseen BT data and FU data have been displayed on top of the individual-level GP predictions in Figure 7.1. It should be noted that due to the increased number of test points, the confidence bounds of the GP models are not immediately obvious for the HC models. However, their visibility is improved for the MS models, in the bottom half of the figure. Qualitatively, minimal discrepancies between the held out BT data and the FU data can be observed. The statistical comparison can be seen in Figure 7.2, where NMSE

and MSLL values are presented using the one-hour apart follow-up test data on the first column (A), and the one-week apart follow-up test data on the second column (B). The non-parametric Kruskal-Wallis test, with a minimum level of significance of $\alpha = 5\%$, together with Dunn-Sidak post-hoc correction was used for the analysis. For completeness, the p-values are presented in Table 7.2.

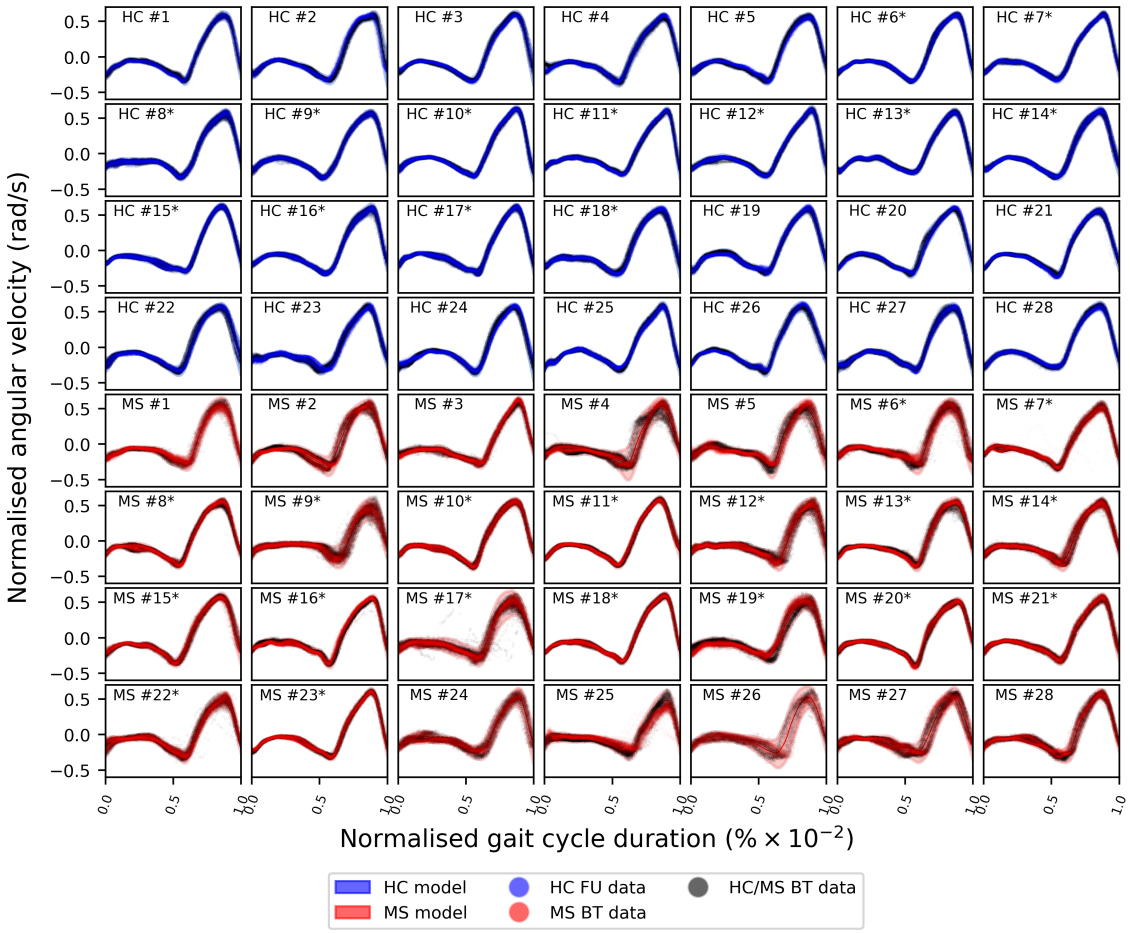


Figure 7.1: Comparison of individual GP predictions, *held-out* BT data, and FU data. The first four rows correspond to HC individuals, while the last 4 rows correspond to MS individuals. Individuals corresponding to group A (one-hour apart FU) are denoted by $*$.

Analysing Figure 7.2, significant differences in the NMSE scores between HC and MS have been recorded, regardless of the timing of the FU assessment (see Table 7.2). The accuracy of the models remains consistent for both groups when comparing the baseline test BT results to the FU assessment, as indicated by the within group BT-FU comparisons. Analysis of the MSLL values presented in Table 7.2 reveals that there are no statistically significant differences when the retest is performed one-hour apart. However, when the retest is conducted one week apart (group B),

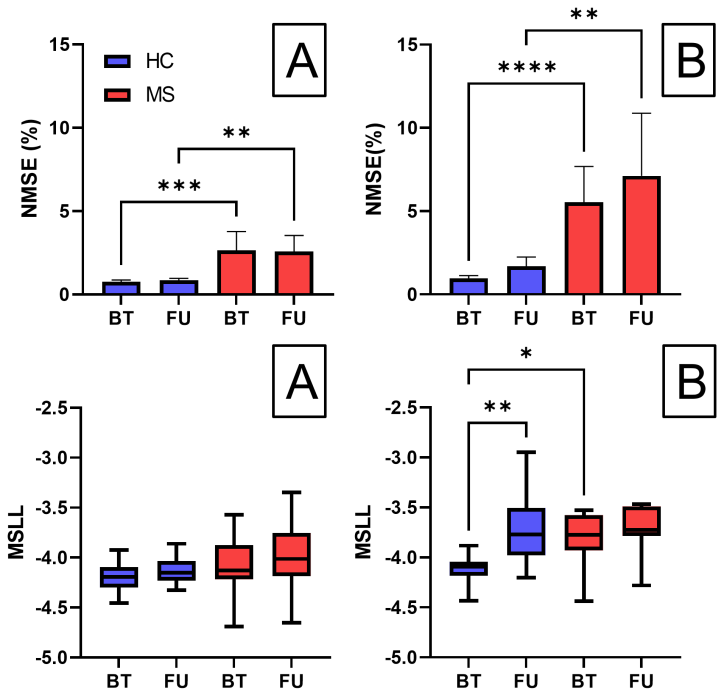


Figure 7.2: Summary of the predictive comparison for the third level of the hierarchy - combined left and right limb data. Here, BT represents the metrics obtained using the baseline-test held-out data, whereas FU refers to the follow-up data. (A) - the one-hour apart group, (B) - the one-week apart group. Significance levels are denoted as: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

significant differences in the MSLL scores are observed between the HC and MS groups, but only on the BT data. This result might have been recorded here, because the patients in this MS group had more pronounced disabilities when compared to the one-hour apart MS group (group A). Particularly, the mean EDSS score for the PwMS in group A was 2.94, whereas for those in group B, the mean EDSs score was 4.1. Interestingly, statistically significant differences in the MSLL scores are also seen within the HC group when the retest was been performed one-week apart. This result is not surprising, given the tight confidence bounds of the HC models learnt at the baseline assessment. Hence, this loss in uncertainty quantification capability is not considered to be important here, since the MSLL values are still within acceptable limits and the average NMSE scores (0.967 and 1.690 using the baseline and follow-up data respectively) heuristically indicate a 'very good' fit. This is because the NMSE can be considered analogous to a percentage error. These results are in line with the findings of [308], who monitored the gait patterns of healthy individuals over a period of two weeks. Their study concluded that even for HCs, gait patterns

are not consistently stable over time, but instead demonstrate distinguishable daily variations, while maintaining previously established good repeatability. The MS group, on the other hand displayed no statistical differences when the retest was performed one-week apart. This results might have been recorded here due to the large confidence bounds learnt at the baseline assessment (see Figure 7.1), which remains consistent over a period of constant disability status, such as the one-week retest period used in this study, in line with previously-reported trends for individuals affected by MS [288].

	NMSE (**A**)	MSLL (**A**)	NMSE (**B**)	MSLL (**B**)
HC BT vs. FU	>0.9999	>0.9999	0.2277	**0.0035**
MS BT vs. FU	>0.9999	>0.9999	>0.9999	>0.9999
HC BT vs. MS BT	**0.0008**	0.5237	**<0.0001**	**0.0447**
HC FU vs. MS FU	**0.0065**	0.6647	**0.01**	>0.9999

Table 7.2: p-values for the Kruskal-Wallis test with a minimum level of significance of $\alpha = 5\%$ and Dunn's correction for the follow-up model consistency comparison using the third level of the HVSHGP model - combining data from left and right limbs. The significant results are highlighted in bold. (A) - the one-hour apart comparison, (B) - the one-week apart comparison.

Finally, the generality of the models was further validated using individual limb models. The visual comparison between GP modelling predictions, held-out BT data, and FU data can be seen in Appendix B.3. As expected, this consistency comparison at the final level of the hierarchy (modelling individual limbs separately) yielded similar results to the previous comparison, where the contralateral limbs were jointly considered. The findings are summarised in Figure 7.3, whereas the corresponding p-values are presented in Table 7.3.

Here it can be seen that no within-group differences have been recorded in group A for both HC and MS-affected individuals. Interestingly, group B displayed statistically significant within-group differences only for the HC individuals, and not for the individuals affected by MS. This result has been achieved given the increased variance observed in the individual-level MS models, which enables them to accommodate the uncertainty present at the follow-up assessments more effectively. The individual HC models, on the other hand, do not possess this feature and any minor deviation at the follow-up assessments will be detrimentally recorded in terms of performance metrics. However, it should be acknowledged that while these differences have been recorded among HC participants, they are relatively insignificant in practical terms,

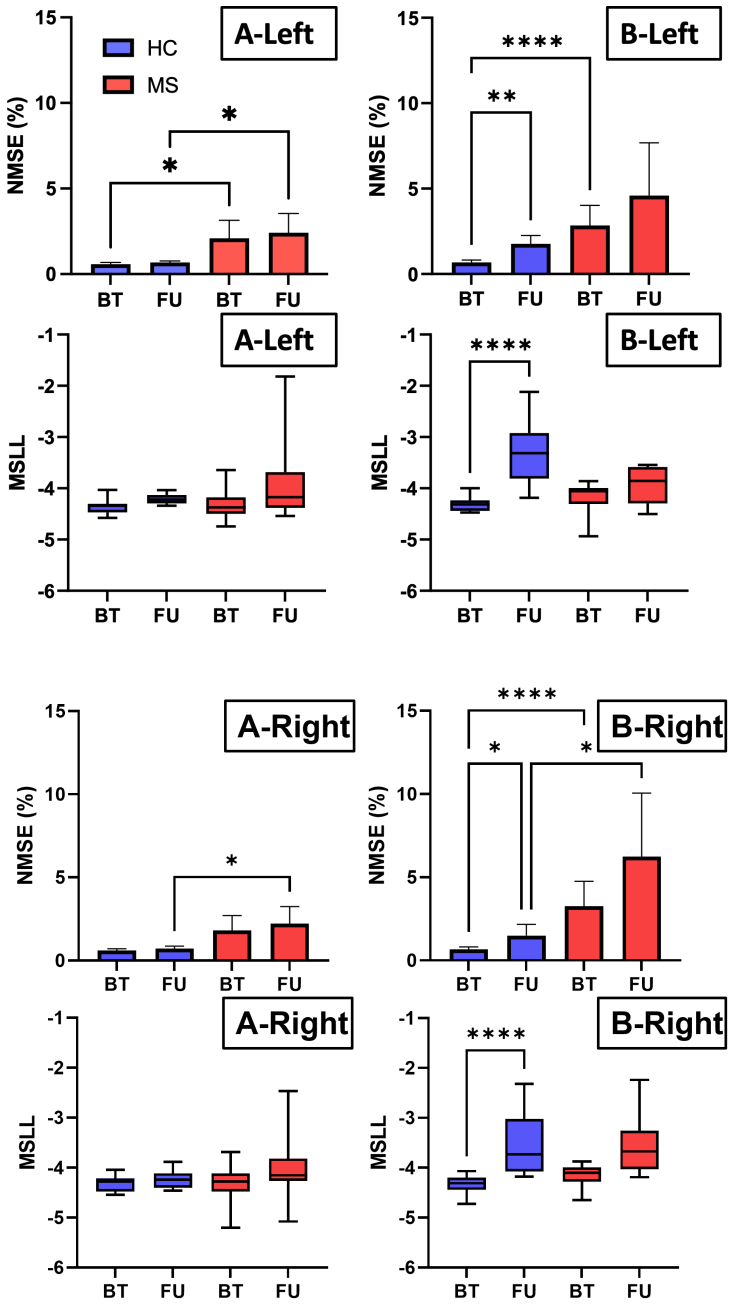


Figure 7.3: Summary of the predictive comparison for the fourth level of the hierarchy - separate left and right limb data. Here, BT represents the metrics obtained using the baseline-test held-out data, whereas FU refers to the follow-up data. (A) - the one-hour apart group, (B) - week apart group. Significance levels are denoted as: * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

considering the magnitude of the NMSE values and the negative values of the MSLL scores obtained. For a visual comparison between the GP predictive distributions, the *held-out* BT data and the FU assessment data, the reader is referred to Appendix B.3.

		NMSE (**A**)	MSLL (**A**)	NMSE (**B**)	MSLL (**B**)
Left	HC BT vs. FU	>0.9999	0.1221	**0.0024**	**<0.0001**
	MS BT vs. FU	>0.9999	0.0758	>0.9999	>0.9999
	HC BT vs. MS BT	**0.0151**	>0.9999	**<0.0001**	0.4939
	HC FU vs. MS FU	**0.0292**	>0.9999	0.2518	0.1428
Right	HC BT vs. FU	>0.9999	>0.9999	**0.0354**	**<0.0001**
	MS BT vs. FU	0.8226	0.1611	>0.9999	0.1045
	HC BT vs. MS BT	0.0528	>0.9999	**<0.0001**	0.2852
	HC FU vs. MS FU	**0.0241**	0.6837	**0.0154**	>0.9999

Table 7.3: p-values for the Kruskal-Wallis test with a minimum level of significance of $\alpha = 5\%$ and Dunn's correction for the follow-up model consistency comparison using the fourth level of the HVSHGP model - separately modelling left and right limbs. The significant results are highlighted in bold. (A) - the one-hour apart comparison, (B) - the one-week apart comparison.

In summary, to assess the consistency of gait patterns and evaluate the generalizability of the proposed models, group and individual-level comparions were categorized into one-hour apart and one-week apart assessments. A robust model that avoids overfitting is essential to aid clinicians in distinguishing between the inherent variability of gait patterns and the impact of disease progression or treatment interventions. In this study, it is hypothesized that if sufficient generality is exhibited by the HVSHGP model, thereby accounting for the stable disease status during this short-term period, then no differences should be detected during the follow-up assessment, irrespective of the influence of variability in testing conditions. To this end, both group-level HC and MS models displayed comparative follow-up results relative to the baselie assessment. Therefore, it can be concluded that the group-level GP models exhibit sufficient generality and can be used as benchmarks in longitudinal investigations. Moving on, although some statistical differences have been recorded for the individual-level HC models, in practical terms, these differences are negligible. On the other hand, no statistical differences have been recorded for any of the performance metrics of the individual MS models, regardless of the timing of the follow-up assessment. This lends the GP modelling approach to be a feasible strategy for the longitudinal investigation of the MS gait patterns. As a result, end-users are now equipped with a general enough model that can effectively accommodate the

natural variability of the gait patterns and can automatically isolate the confounding factors at follow-up assessments. Finally, this generalizability represents an extra step towards studying the longitudinal progression of MS gait patterns over time.

7.2 Part 2: Longitudinal monitoring of gait patterns in MS

Having established the suitability of GP models for longitudinal analysis, the focus can now shift towards monitoring the MS-affected gait in the long term. As such, for this task, a refined group of 16 additional individuals with secondary progressive MS (SPMS) was included in the analysis. These individuals are part of the second dataset presented in Section 1.3. The corresponding demographics details are presented in Table 7.4, where the average age is reported at the baseline assessment. These individuals attended all four visits spread across a period of 96 weeks. Specifically, following the baseline assessment at week 0, the follow-up screenings were attended at week 24, 48 and 96, ensuring consistency across the study's temporal framework. All participants included in this study were diagnosed with SPMS, which is characterised by insidious disability worsening that is independent from clinically apparent relapses [309]. The inclusion criteria was specifically tailored to individuals capable of unassisted ambulation. Finally, data processing was conducted analogous to the procedure described in Section 6.3.1. For brevity, the description of the procedure will not be repeated again here.

Table 7.4: Demographics table for the subjects included in the longitudinal analysis.

MS Subjects	**Age**	**Gender**	**EDSS**
	Mean (SD)	*N male*	*Mean*
$n = 16^{\star}$	56.38 (8.90)	5	5.03

$\star$ = *secondary progressive MS*

In view of the results presented in the first part of this chapter, where it has been established that the group-level HC model exhibits sufficient generality, this section will first employ this model to quantify the degree of gait impairment in MS, by assessing the degree to which the HC model aligns with or deviates from the longitudinal MS data. Secondly, in accordance with Chapter 6, asymmetry has

been considered a significant feature associated with the extent of gait deficit. As such, to provide insights into the progression and variability of the disease, this study targeted two main objectives:

1. To evaluate the predictive performance of the group-level HC GP model in comparison with the longitudinal MS data, and to investigate the temporal relationship between these predictive modelling outcomes and the severity of the disease, as quantified by the EDSS scores.
2. Quantify the longitudinal progression of asymmetric gait patterns in MS and explore the temporal relationship between asymmetry and disease severity.

With respect to the first objective of this study, here it is proposed to use the previously established group-level HC GP model as a benchmark, representing the healthy gait pattern, and compare the alignment of the longitudinal MS data with the predictive distribution of the HC model. This comparison will be quantified by computing the NMSE and MSLL metrics, using the measured MS data as input for the HC model. Regarding the second objective, for each longitudinal assessment, the unified WD and KL divergence-based asymmetry metrics have been recomputed. For the extensive computational details regarding the asymmetry metrics, the reader is referred back to Section 6.3.5. Thus, for each of the longitudinal assessments, four additional three-level HVSHGP models had to be constructed. Each of these models correspond to a specific visit. At the group level, these models incorporate data from all 16 PwMS included in this study. Furthermore, the models expand the analysis to encompass individual-level gait patterns, and also separately model the contralateral limbs, that is, the left and right limbs, thereby providing a more comprehensive evaluation of gait asymmetry. Here, a 70-30 split was employed for training and *held-out* test data respectively, all exclusive of the previous training sets used for the four-layer HVSHGP model used in Chapter 6 and in the first part of the this chapter. The model at the baseline longitudinal assessment (i.e., week 0) was trained using $337,673$ data points and tested with an additional $144,718$ data points. As the study progressed to week 24, the model was trained using an expanded dataset of $423,183$ data points and tested with $181,365$ data points. At week 48, the model was trained using $437,012$ data points, while another $187,292$ data points were used for validation. Finally, at week 96, the model was trained using $412,020$ data points and tested with $176,580$ data points. The performance metrics of the three-layer HVSHGP models across each visit are provided in Appendix B.4.

To begin with, using the group-level HC model established in Chapter 6, Figure 7.4 presents this model's predictions against the longitudinal MS held-out data. For comparative purposes, the individual-level GP predictions of the longitudinal MS models have been also overlaid. Notable dissimilarities are observed between the predictive distribution of the HC model and the longitudinal MS data. Moreover, it is evident that the longitudinal MS models manifest substantial fluctuations across the study's timescale. This observation underpins the rationale for selecting the HC model as a benchmark reference model, indicative of the typical gait pattern observed in healthy individuals. Without a benchmark model, the quantification of the longitudinal gait deficit cannot be accurately conducted. As such, the MS model established at week 0, for example, or in fact at any other assessments, cannot be utilised for comparative purposes since the discrepancy between this model and the HC baseline would have been undetermined. By employing the HC model as a baseline, it facilitates a direct comparison between the gait patterns associated with MS and the healthy condition. This comparison not only simplifies the quantification of degree of gait deviations but also it enables tracking of gait alterations longitudinally. Additionally, the HC model provides a representation of typical gait patterns observed in healthy populations, thus enhancing the interpretability and reliability of the analysis.

The NMSE and MSLL metrics computed across all longitudinal assessments are presented in Figure 7.5, together with the evolution of the EDSS scores, which were reassessed by expert clinicians at each visit. Analysing both Figures 7.4 and 7.5, several individuals stand out. Subject #01 exhibited an atypical gait pattern characterized by a flatter trajectory during the stance phase and a significantly shortened swing phase. Interestingly, despite no change in EDSS score between baseline and the second assessment, the shank angular velocity pattern at week 24 resembled that of a HC. Furthermore, a gradual normalization of the gait pattern towards the HC trajectory was observed, with the greatest similarity occurring at the fourth assessment (week 96). This finding aligns with the observed decrease in EDSS score, suggesting a potential link between improved clinical status and gait recovery. Subject #03, on the other hand, presents a completely different behaviour. Even though the clinical score remained constant throughout the entire longitudinal study, it can be seen that there is a gradual divergence between the normative HC pattern and the longitudinal subject-specific MS pattern. Subject #08 showcased a gradual departure for the typical HC pattern as well. However, the evolution of the EDSS score did not undergo the same trend, given the increase at week 48,

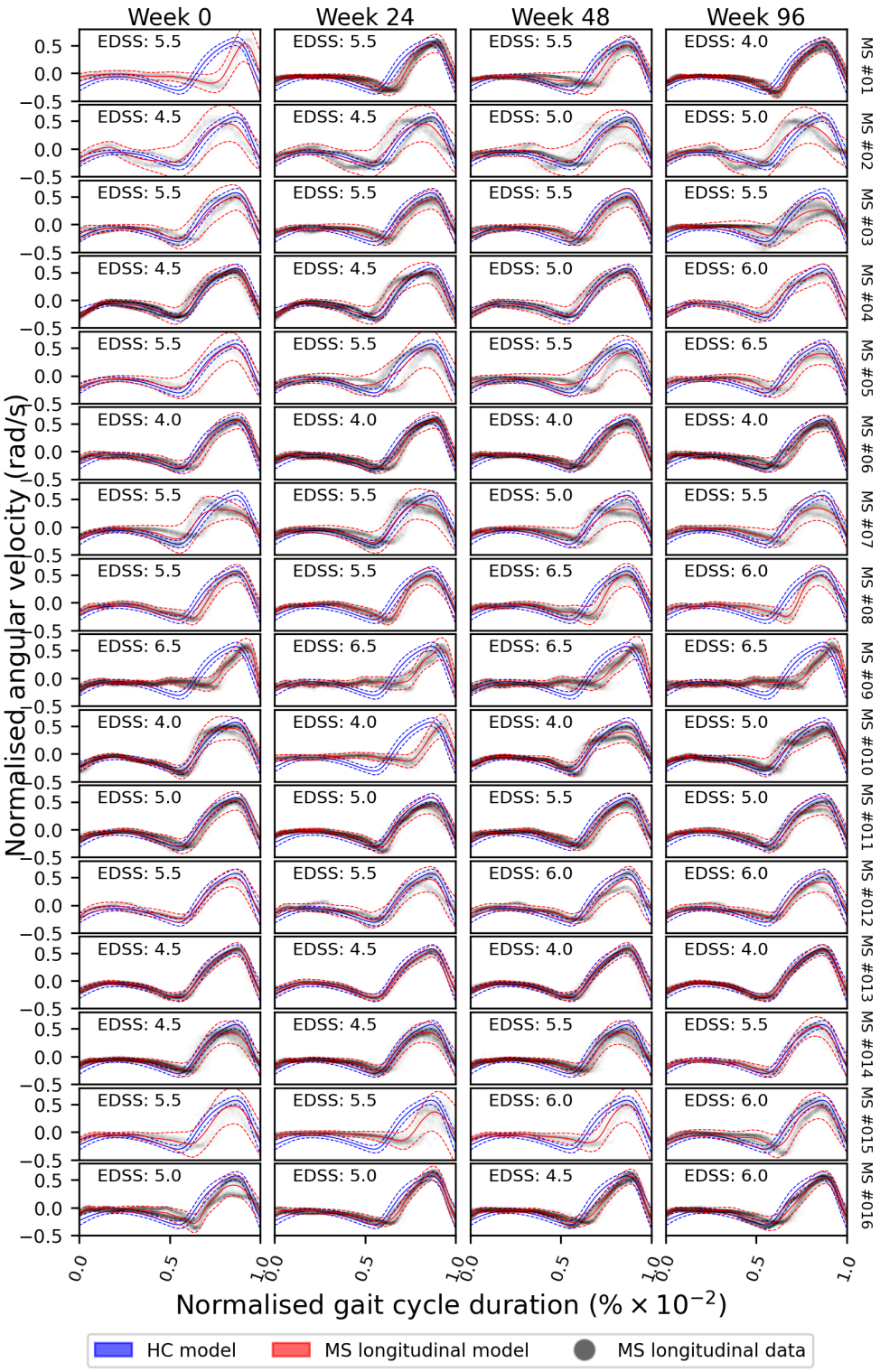


Figure 7.4: Visual comparison between the baseline assessment group-level HC model established in Chapter 6 and the longitudinal MS GP models. The blue lines represent the HC model, while the red ones represent the MS models. Solid lines have been used to mark the mean predictions, while the 95% confidence interval is marked by the dashed lines. Data used for validation of the longitudinal MS models is also overlaid here using the gray dots.

which then was followed by a decrease at week 96. Conversely, looking at individual labelled #09, consistency was recorded throughput the entire study. Additionally, subject #16 also displayed an unexpected progression of gait patterns. While, for this subject, the longitudinal data showed increasing alignment with the normative HC model's GP predictive distribution, suggesting improvement, the corresponding EDSS scores diverged from this trend. The challenges of monitoring disease progression in MS using functional gait assessments are effectively highlighted by these results. This difficulty is attributed to the inherent disease heterogeneity, which leads to unpredictable changes in gait patterns observed across individuals, as reported in [45].

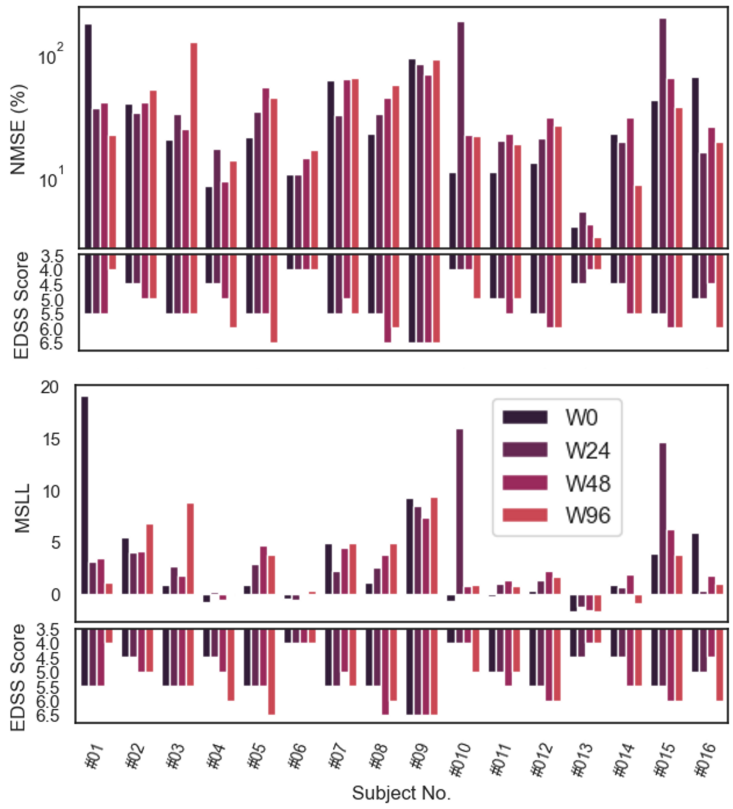


Figure 7.5: Bar chart of the performance metrics obtained using individual-level data across the longitudinal assessments.

To formally investigate the longitudinal relationship between HC model's predictive performance metrics and MS disability status, following the procedure presented in [50], a mixed-effects model with a threshold alpha value of $\alpha = 0.05$ signifying statistical significance was employed [310] to account for repeated measures within the 16 MS-affected individuals across all four longitudinal assessments. Standardised

regression models were unsuitable for this longitudinal data due to the correlation between repeated measures within participants, which violates the independence assumptions of traditional regression methods [311–313]. For brevity, the mathematical details and the summary results of the mixed-effects model are presented in Appendix B.5. Here, a statistically significant ($p = 0.023$) negative correlation was found between the NMSE and EDSS score (as indicated by the estimated coefficient of -0.016). Conversely, the positive correlation (as indicated by the estimated coefficient of 0.185) between the MSLL and EDSS score was also found to be statistically significant ($p = 0.013$). Moreover, enough evidence was found to conclude that the EDSS score increases as the disease is progressing ($p = 0.003$). It should be also noted that for a few PwMS, there was no linear path of disability progression, highlighting the variable and unpredictable disease course in SPMS [16] and further increasing the complexity of the problem. In addition, the discrepancy between the NMSE and MSLL trends seems to be rather counterintuitive. On one hand, the decreasing NMSE suggested improved accuracy in mean predictions over time, signifying closer alignment between the central tendency of the MS longitudinal data and the mean HC model predictions. On the other hand, the increase in MSLL implies worsening performance in uncertainty estimations, suggesting a decreasing capability of capturing the variability of the longitudinal MS data. The divergence between these two trends is rather counterintuitive and may indicate that although the mean pattern of PwMS is converging towards that of healthy individuals, the increase in heteroscedastic variance continues to serve as a significant predictor for the accumulation of MS-related disability. Once again, the heteroscedastic variance may be attributable to two distinct factors: continuous deficits in lack of control during the swing phase, as a result of muscle spasticity, fatigue or balance impairments [159, 194], but also to the presence of asymmetry.

In addition to the comparison to the group-level HC model, the longitudinal progression of the asymmetry levels characterising the MS-affected gait patterns has been also investigated. For this purpose, four additional three-layer HVSHGP models corresponding to each visit have been established. The individual limb-level MS model predictions are displayed in Figure 7.6 against the held-out data, for an improved qualitative overview of the longitudinal asymmetry fluctuations. The quantitative comparisons in terms of WD and KL divergence-based metrics are presented in Figure 7.7, together with the longitudinal EDSS scores.

Here, the WD has been used to quantify the extent of disparity between contralateral

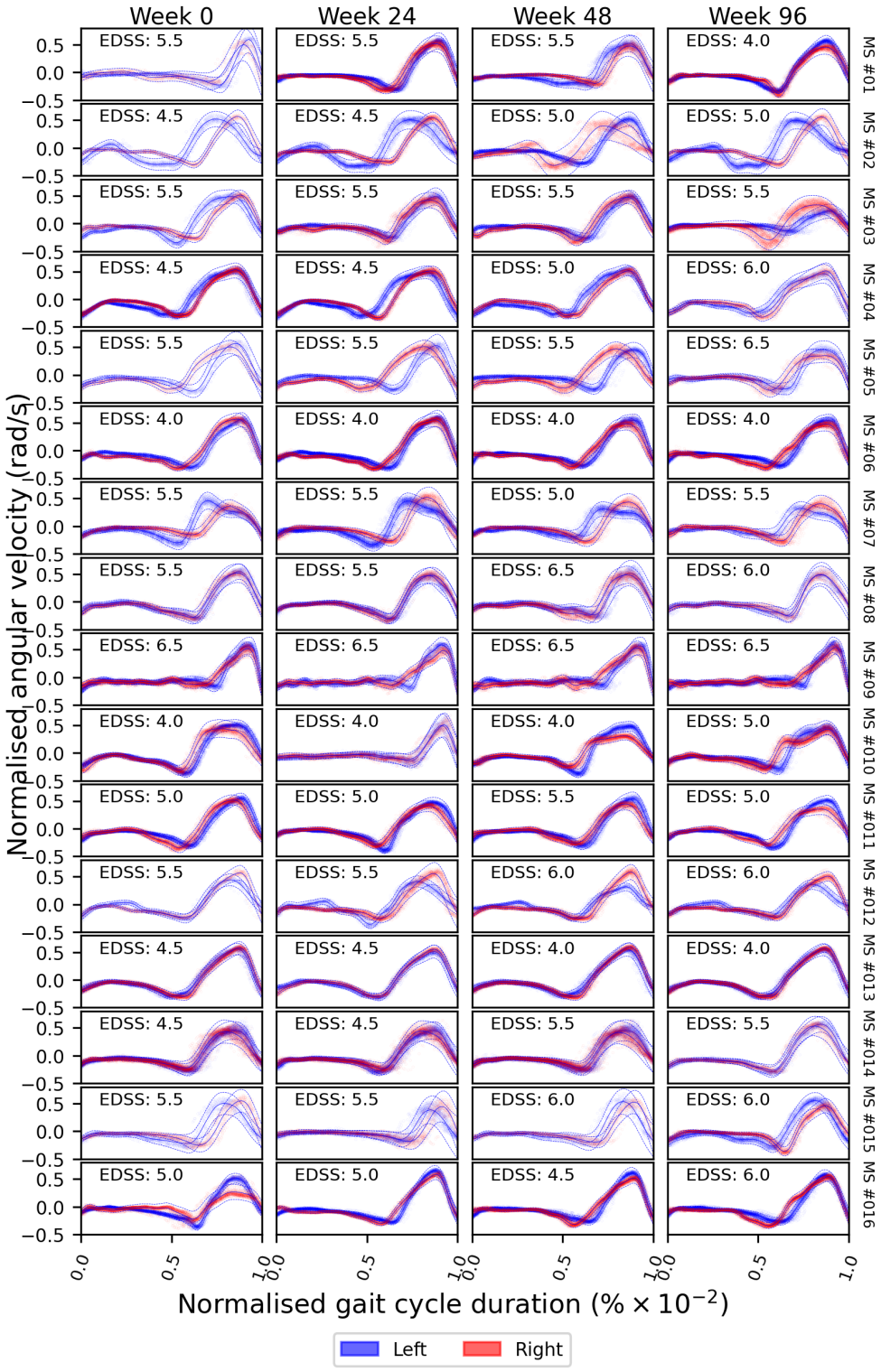


Figure 7.6: Longidtinal contralateral limb GP predictions. The solid lines represent the mean GP predictions, while the dotted lines correspond to the 95% confidence interval. The dots represent the held-out data.

limbs models throughout the entire gait cycle. Similarly to the procedure described in Section 6.3.5, 500 samples were drawn from each GP's predictive distribution at 100 equidistant points spread across the input space. Subsequently, the unified asymmetry metric has then been computed as the area under the WD curve. The KL divergence-based asymmetry metric was also used to quantify the degree to which the individual-layer model (i.e. simultaneously modelling contralateral limbs) accurately represents the dynamics of the individual limb layer models, that is, treating contralateral limbs independently. The KL divergence quantifies the difference between one probability distribution and a second, expected probability distribution. In this context, a high KL divergence would indicate that the individual level GP model does not accurately represent the dynamics of the individual limb models. Conversely, lower KL divergence values indicate that the individual level GP is a good approximation of the individual limb models, indicating that the specific subject does not display significant asymmetry. To ease comparison, $\Delta\ KL$, the absolute difference between the KL divergence metrics computed for each limb has been used in this case. Upon joint examination of Figures 7.6 and 7.7, it becomes evident that there are significant longitudinal variations in the levels of asymmetry. As such, the longitudinal relationship between the EDSS score, the asymmetry metrics and the time of assessment, was investigated using another mixed-effects model with a threshold alpha value of $\alpha = 0.05$, signifying statistical significance, similarly to the previous analysis. For conciseness, the summary results pertaining to the second mixed-effects model are provided in Appendix B.5. Among the investigated metrics, only (ΔKL) exhibited a statistically significant negative correlation with the EDSS score ($p < 0.001$). This finding suggests a counterintuitive relationship, where decreasing asymmetry levels (as indicated by a decrease in ΔKL) are associated with an increase in disability status (as reflected by a higher EDSS score). Therefore, in light of these results, but also due to the small sample size, no clinically meaningful conclusions can be drawn at this time, as the progression pathways exhibit a heterogeneous nature that cannot be effectively predicted. These results further emphasise that longitudinal monitoring of the gait patterns in PwMS is a very challenging endeavour and more comprehensive studies are necessary.

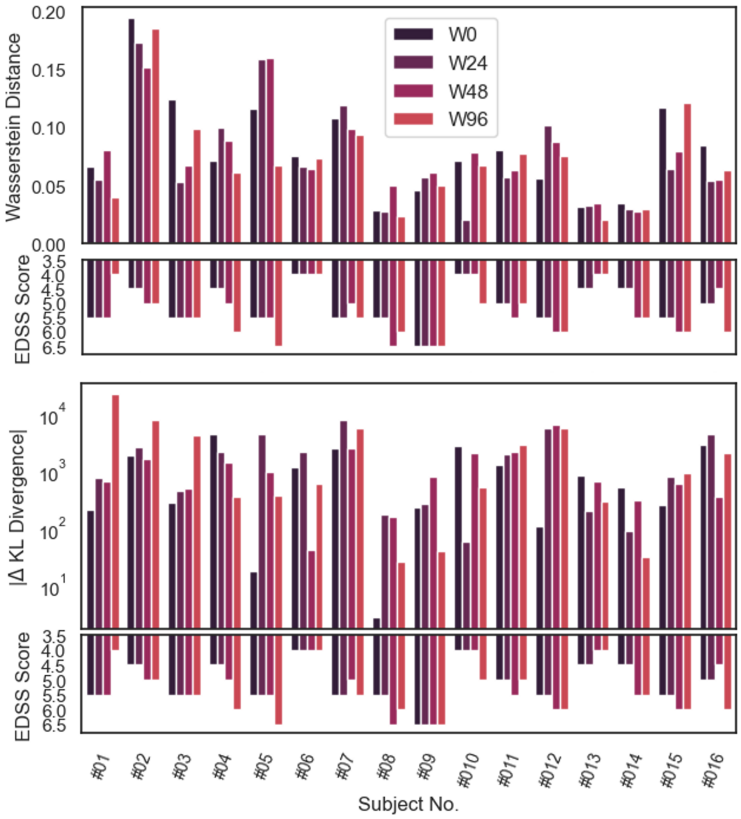


Figure 7.7: Bar chart of the asymmetry metrics obtained using individual limb models across the longitudinal assessments.

7.3 Discussions

This study introduces hierarchical GP modelling of the shank angular velocity as a novel approach for monitoring the longitudinal progression of the lower limb distal motion in MS-affected individuals. To achieve this, an initial validation study has been presented, where it was concluded that group-level HC and MS models established during the baseline assessment exhibit sufficient generality in order to be used for longitudinal assessments. This first step is essential, as follow-up assessments may be further impacted by confounding influences such as variations in sensor placements, differences in timings of assessments, footwear differences, among other variables. The proposed group-level models effectively accounted for both the influence of these confounding factors, as well as the high variability in gait patterns induced by de disease [46]. Here, the one-week apart retest ensured a stable disease status period, effectively isolating the effects of disease progression or other possible

treatment side effects. In addition to this, subject-specific MS models also achieved test-retest generalizability, further reinforcing the suitability of the proposed method for longitudinal monitoring.

Following the successful validation of the proposed HVSHGP modelling approach, the attention shifted towards the rather exploratory study aimed at characterising the longitudinal progression of the shank angular velocity in MS. For this task, an additional dataset of 16 individuals with SPMP was used. Analysing Figures 7.4 and 7.6, it can be seen that the acquired data showed a highly heterogeneous and subject-specific disease course. Here, the degree of ambulatory dysfunction was quantified using four metrics. The initial two metrics, the NMSE and MSLL values, were derived using the group-level HC GP model, established using the baseline data. The advantage provided by this approach is that the HC model effectively encapsulates the range of expected patterns in a HC population, and can therefore serve as a direct comparative benchmark. Specifically, NMSE values approaching zero signify closer alignment of average MS gait kinematics with expected healthy patterns, while more negative MSLL values indicate that the MS gait pattern variance is falling within the confidence bounds of the healthy gait patterns. Moreover, once established, the HC model does not have to be recomputed again and can be used for an indefinite number of follow-up assessments. The last two metrics were the WD and KL divergence-based asymmetry measures computed using the newly constructed models at each of the four longitudinal assessments. This approach allowed for a subject-specific monitoring of the asymmetry metrics, a feature characteristic of MS [17]. Nonetheless, while these metrics provided an overview of the impaired gait pattern, their association with disability levels in MS, as measured by the EDSS score remains ambiguous and no clinical conclusions can be drawn here.

It is believed that one of the main limitations of this study remains the dependence on the EDSS score for measuring disability progression. This is because this measure is not a direct measurement of ambulatory disfunction. Although this may be mostly true for the early stages of the disease, it is possible that the observed changes in EDSS scores among the several PwMS included in this study could be attributed to the deterioration of other functional systems with less impact on mobility, such as the visual or cerebral systems. This possibility might also account for the observed variations in EDSS scores despite the absence of substantial changes in gait patterns, as suggested in [45]. In addition to this, over the years, this scale has also received criticism for several reasons, including lack of objectivity and sensitivity to clinical

evolution [57, 314–317]. As such, while no clinically meaningful correlations were found between the proposed metrics and the EDSS score, the approach suggested in this chapter might still offer valuable insight into the complex progression of gait disability in MS. Nonetheless, the question of establishing a clear and effective method for this monitoring task remains open, warranting further research.

Beyond the current approach, it is also believed that more frequent assessments are likely to further improve the understanding of the progression of the disease. While the path taken by Creagh et al. [16] involving daily, self-administered smartphone-based walking tests offers a distinct methodology, they demonstrate the potential of out-of-clinic assessments for capturing longitudinal changes in MS gait patterns, providing novel perspectives for future research. Furthermore, the data-driven methodology presented here has limitations in extrapolating beyond the training set. To enrich the understanding of gait disability progression, future studies should explore how additional information, such as established biomechanical models or clinical knowledge, could enhance the prognostic capabilities of the model. To this end, physics-informed machine learning, which has shown success in structural health monitoring [318, 319], offers a promising avenue for the healthcare sector by integrating data-driven approaches with established physical knowledge. This approach may improve the generalizability of models to unseen scenarios and provide interpretable insights into the relationship between gait patterns and disease progression in MS patients.

7.4 Conclusions

This chapter attempted the most challenging aspect of the hierarchical assessment framework proposed by the author, consisting of making inference about the progression of the disease. Here, the idea of monitoring the longitudinal progression of the MS gait patterns in relation to disease severity status has been explored using hierarhical GP models. Initially, the generalizability of the HVSHGP model proposed in Chapter 6 has been demonstrated, through a validation study utilising completely unseen, follow-up assessment data. The group-level model established using the HC population then served as an useful benchmark, indicative of the typical gait pattern observed in healthy individuals. The alignment of the longitudinal MS data with the predictive distribution of the HC model has then been investigated by

computing the NMSE and MSLL perfromance metrics. Moreover, the evolution of asymmetry metrics computed using subject-specific longitudinal GP models were also exploited. At this early stage of research, no conclusive remarks can be made about the disease course from a gait analysis perspective, given that no clinically meaningful correlations between any of the metrics and disability status were recorded. While it is suggested to augment the available data with more frequent assessments or attempt a modelling strategy which includes addtional clinical knowledge, this will be the focus of future work, which will be discussed in more detail in the next and final chapter of this thesis.

Chapter 8

Conclusions and further work

This chapter synthesizes the key findings from each results-driven chapter, exploring their contribution to advancing knowledge within the field of gait analysis, with a particular focus on individuals affected by multiple sclerosis (MS). Building upon an overview of the disease and current analysis challenges, as outlined in Chapter 1, the author proposed a novel hierarchical framework for assessing and monitoring the condition of MS patients. As part of the suggested framework, the set of challenges for assessing gait were broken down into four levels, which include detection of gait impairment, classification of gait impairment type, disease severity quantification, and longitudinal monitoring and prognosis. The motivation for a data-driven approach undertaken in this thesis has also been presented in this chapter, along with the datasets and experimental setup. Chapter 2 then provided a background on the current state-of-the-art, as well as a brief introduction to some machine learning technology of importance to the methodological advancements presented later in this thesis. Thus, the subsequent chapters addressed the challenge of improving assessment and monitoring capabilities in MS through various methodologies, which systematically targeted the levels in the proposed hierarchical assessment framework. In the following paragraphs, each of the techniques presented throughout this thesis will be further discussed. In addition, a critical analysis of the limitations inherent to each chapter's methodology and findings is also presented to the reader. Finally, this final chapter concludes with the potential avenues for future research.

8.1 Robust detection of gait anomalies

The work presented in Chapter 3 focused on the first level of the hierarchical assessment framework, specifically on detecting gait impairments attributed to the presence of the disease. This is a key challenge, since being able to detect early signs of the disease may allow end users to design tailored therapeutic interventions, potentially improving patient outcomes. Particularly, in this chapter, the problem of detecting gait impairment has been framed as a novelty detection problem, where the proposed algorithm has been trained using data obtained exclusively from healthy individuals. For this task, inspired by its proven success in the structural health monitoring (SHM) field, the Mahalanobis squared distance (MSD) has been selected as a suitable tool. Following careful data transformations based on well-established biomechanical principles relevant to gait analysis, the extracted gait features were modelled as multivariate distributions. Then, distinguishing abnormal patterns from healthy ones effectively required a comparative analysis against an objective threshold. The idea introduced here is that MSD values exceeding the threshold are considered to be indicative of abnormal gait patterns (possible as a result of altered movement patterns), while values below the objective threshold are deemed as normal. However, it has been shown that this approach, particularly due to a systematic error present in the processing of the healthy control (HC) data, can be compounded by inclusive outliers, that might mask the true, healthy condition. To this end, the minimum covariance determinant (MCD) estimator has been used to mitigate this problem, allowing automatic detection and removal of the inclusive outliers.

Utilising feature selection methods, and particularly a modified version of the sequential forward selection (FSF) algorithm, it was possible to reduce the dimensionality of the problem and ensure the exclusion of irrelevant features that may impede clinical interpretation. As such, an optimal feature set has been presented to the reader, which encompasses not only traditionally extracted gait metrics, but also descriptive statistics and periodicity measures derived from the IMU gait signals. Importantly, the proposed feature set allowed the MSD-based outlier detector to detect gait anomalies even in the early stages of the disease, for individuals who often lack clinically observable walking disabilities. It has also been noticed that, no individual gait feature exhibited a clear differentiation between HC and MS, nor across various levels of MS disability. This highlighted the complex nature of human

gait and emphasizes that distinction between the MS population and HCs requires a comprehensive multivariate exploration of gait data.

One of the main limitations of the work presented in this chapter revolves around the question of whether there might be some other potential useful features that have not been explored, which could be more effective at highlighting the underlying differences between HC and MS. For example, future work should also investigate the usefulness of additional spatial features, which have also been shown to encode potentially useful information [79, 320]. In addition to this, another potential limitation of the outlier detection technique employed in this thesis is the reliance on statistical control charts in order to visualise the objective threshold. While extreme value statistics were utilised in order to obtain a threshold that may be less prone to producing false positives, this concept requires further exploration. To this end, adaptive threshold methods may be worthy of investigation in the future.

8.2 Single sensor severity assessment

Chapter 4 introduced the key findings of the second level of the proposed hierarchical assessment framework, revolving around the development and evaluation of a self-supervised approach for classifying the severity of the disease. This task was achieved using a single wearable sensor approach. The methodology presented in this chapter employed a contrastive methodology in order to achieve a latent embedding space that effectively discriminated between clusters of HCs, mild MS, moderate MS, and severe MS. The results showed that the record-wise cross-validation (CV) approach exhibited slightly superior performance on the validation sets compared to the subject-wise CV approach. However, subject-wise CV was deemed more appropriate for robust evaluation and generalizability of the findings, as it is better aligned with the clinical scenario of diagnosing new patients that were unseen by the model during the training phase. Given that this chapter introduces a multiclass extension to the binary classification task used for detecting gait impairment, it can be argued that the methodology presented here surpasses the previous chapter's approach, as it enables differentiation not only between HC and MS, but also across MS severity levels. Nevertheless, a direct comparison between the two approaches for gait impairment detection would not be possible and fair. The study also utilized the layer-wise relevance propagation (LRP) technique to provide insights into the features most

relevant to the model predictions. This aids model interpretability and builds trust in its predictions. The LRP analysis revealed that the network learned clinically meaningful features. It highlighted that the double support phase and swing phase are relevant, subject-specific characteristics, which are known differentiating factors of MS gait. The relevance scores attributed to specific gait events and phases provided initial evidence of the utility of the proposed network for correctly classifying disability levels in MS. As such, the visual explanations provided as part of this methodology might facilitate further collaborations between machine learning practitioners and clinical experts, leading to more informed assessments for individuals with MS.

The implications of the results presented in this chapter are significant. The deep learning approach presented in this study has the potential to detect impairment in MS and accurately classify disease severity using a single IMU. This approach might prove especially useful in remote assessments, where, for example, monitoring the probabilities associated with the predictions may prove useful for detecting disease progression or transitions between severity classes. It might also have a potential future usage in online embedded systems applications. Moreover, another inherent advantage of this methodology lies in its departure from the traditional reliance on carefully selected 'expert' features for representing gait. Instead, the neural network-based approach presented in this chapter allows the algorithm to automatically extract the most meaningful features through representation learning. However, while the proposed methodology was seen to offer excellent classification results, there is still a need for additional validation procedures before this methodology can be trusted for clinical usage. These include additional validation studies on new datasets, and the exploration of how this approach may perform on more complex tasks, including remote assessment scenarios. Finally, future work should also investigate additional propagation schemes and network architectures.

8.3 A first attempt at quantifying gait consistency

In Chapter 5, a novel methodology for objectively quantifying gait pattern consistency has been proposed. This aspect is of key interest, since a model that can exhibit sufficient generalizability might allow end users to distinguish between natural variability of the gait patterns from disease progression or treatment effects. In view of the observed variability in the gait pattern, owing to both environmental

and pathological influences, this chapter introduced a two-part methodology for the quantification of gait consistency. Firstly, the residual pattern of a dynamic with eXogenous input (ARX) model between both shanks and the lower back was used as a sensitive feature. Secondly, the maximum mean discrepancy (MMD) was introduced to measure the differences in the distribution of the residuals, together with its corresponding hypothesis test. Here, the MMD has been chosen since, no *a-priori* knowledge about the form of the residual distributions should be assumed. As such, the MMD leveraged the kernel trick, effectively enabling assessment of an infinite array of statistical moments via inner products through a reproducing Hilbert space. In turn, this approach led to a sensitive and informative method for evaluating the consistency of the gait patterns.

The development of this methodology provided some interesting insight into the challenges that the gait analysis community might be facing for longitudinal analysis. Particularly, it revealed the detrimental effects of varying assessments conditions on gait pattern consistency and it highlighted a need for more general methods that can effectively overcome the limitations of the ARX modelling technique presented here. This is because this approach is rather pragmatic, particularly because it only provides a linear representation of gait. In fact, gait patterns are inherently non-linear. For this reason, the next chapter provided some additional modelling flexibility, together with uncertainty estimates. Finally, future work should also emphasise more the model-predicted output (MPO) performance, rather than one-step ahead (OSA) performance as presented in this chapter. This task requires feedback of the model prediction itself. Although MPO performance is typically anticipated to be inferior to OSA due to compounding errors, it nonetheless serves as a more accurate indicator of the model's ability to capture the true dynamics of the system.

8.4 The move towards probabilistic modelling of gait patterns

Building on the limitations presented in Chapter 5, Chapters 6 and 7 proposed a completely different outlook for gait monitoring purposes, introducing probabilistic modelling techniques. For this, the Bayesian paradigm emerged as a powerful framework for gait analysis. It enabled the construction of complex models, while

formally incorporating prior knowledge, or the absence thereof. Unlike traditional, deterministic approaches, the Bayesian framework embraces uncertainty as an inherent feature crucial for a comprehensive understanding of gait, which is particularly useful in the challenging context of quantifying the degree of gait impairment and its changes over time. Within a Bayesian framework, Gaussian Processes (GPs) have been introduced for the regression task of modelling the non-linear shank angular velocity as a proxy of lower limb distal movement. Firstly, the standard GP formulation has been extended to allow hierarchical inference, since there is a common trend shared across a pool of data from multiple individuals performing the same walking test, regardless of the labels. Then, the problem of scaling GPs to handle large datasets has been addressed through variational approximation methods. Additionally, the non-constant variability throughout the gait cycle - which has been shown to be a key characteristic of MS gait - has been managed through heteroscedastic noise modelling extensions. This was achieved by modelling the log-noise variance of the process as an additional GP, which was also learnt in a variational manner. While the combination of heteroscedasticity and sparse inference has been previsouly addressed in the machine learning literature, to the best of the author's knowledge, the additional integration of hierarchical modelling presents a novel contribution.

The utility of this methodology was underscored by its granular analysis capabilities. This facilitated a range of quantifiable comparisons, spanning from group-level assessments to patient-specific analyses, addressing the complexity of pathological gait patterns and offering a robust methodology for kinematic pattern characterisation for large datasets. The group-level analysis highlighted notable differences during the swing phase and towards the end of the stance phase of the gait cycle, aligning with previously established literature findings. Additionally, a novel approach for lower limb gait asymmetry quantification has been proposed. The use of probabilistic hierarchical modelling facilitated a better understanding of the impaired gait pattern, while also expressing potential for extrapolation to other pathological conditions affecting gait. However, although the proposed methodology was shown to exhibit sufficient generalizability when attempting to predict the underlying shank angular velocity pattern at follow-up assessments over a short period, during which the health status of MS-affected individuals remained constant, the results presented as part of the rather exploratory longitudinal study were inconclusive. This is because no clear association was found between GP performance metrics, the novel asymmetry metrics proposed and disability status recorded through the Expanded Disability Status

Scale (EDSS) score. In light of these results, the next section explores potential future research directions that could significantly contribute to advancements in longitudinal assessments and may have a direct impact on the understanding of disease progression.

8.5 Future work

The research conducted in this thesis has opened several intriguing avenues for further investigation, which will hopefully inspire future work in this field. As the results in Chapter 7 show, by far, the hardest area of research remains the problem of longitudinal monitoring and prognosis. A possible reason for this is likely to be the infrequency of current clinical assessments, in combination with the lack of available data. As such, the approach taken by Creagh et al. in [16, 44], which involved daily out-of-clinic self-administrated tests where participants performed a 2-minute walking test while IMU data was acquired with a smartphone, might be a potential solution to this problem. Nonetheless, for a reasonable predictive methodology to fully utilise the temporal information acquired using this approach, in order to predict the trajectory of the disease, there is a need to further explore the aggregation of temporal information directly. Although attention mechanisms that identify the most informative temporal patterns and how the predicted disability score is influenced by historical outcomes, such as those developed in [321, 322], seem suitable for this task, this problem will more than likely require further development and integration of various deterministic and probabilistic modelling approaches, as well as the ability to quantify the uncertainty in these predictions.

It is also worth considering how these methodologies might be applied within the context of real-world gait monitoring. So far, the methods presented in this thesis only utilised straight-line walking bouts within a six-minute walking test, which was originally designed for measuring walking capacity, under a controlled environment. While some studies suggest that turnings [50, 323] may also encode complementary information, a major limitation of the data collection setup used in this thesis is that it does not consider any contextual factors or complexities in walking that might be encountered in real-world applications. To this end, before these methodologies can be deployed as part of clinical practice, there is a need for establishing ecological validity, as part of an extensive technical and clinical validation study [205]. To

achieve this, an initial step might entail validating the proposed methodologies using a multi-task, multi-context protocol that simulates real-world environments within a controlled laboratory setting, as proposed by Scott et al. [324]. This should be followed by a larger validation study using real-life continuous monitoring data. To this end, the algorithms proposed in [79] for detection of gait sequences and gait events would certainly become advantageous.

Another area of future research revolves around model adaptability, which becomes important as more data are collected, especially when building predictive models for disease prognosis. This is also motivated by the desire to further refine algorithms in an online setting, where updates and automated decisions occur in real-time during system operation [325]. To this end, some key requirements must be fulfilled. Firstly, gait analysis systems, which encapsulate IMU sensors, signal processing methods, learning algorithms, as well as decision engines, should be adaptive and capable of incorporating novel gait patterns as they are discovered. This is particularly relevant, considering the unpredictable pathway for disease progression characterising MS. Additionally, the algorithms should prioritize computational efficiency to enable real-time operation directly on the device. Finally, the models should be capable of providing accurate diagnostics (ideally be probabilistic, in order to account for the uncertainty of predictions) and recommendations that clinicians can act upon.

These considerations fall under the umbrella of *active learning* [326], which is an area of research aiming to improve the efficiency and effectiveness of machine learning models by selectively querying the most informative data points for labeling. Active learning algorithms can be applied both offline, using a large number of previously collected data (as in the case of this thesis) [327], or online, to novel data streams which evolve through time [325, 328]. Inspired by the successful applications in SHM [325], further research proposes the adoption of probabilistic active learning for analysing wearable sensor data in the context of remote longitudinal monitoring and prognosis. Despite the vast volume of data acquired in remote monitoring settings, critical events are often infrequent, which implies that not every data point demands immediate attention. Thus, by leveraging Bayesian statistics and information theory, an active learning framework may prioritise the most informative datapoints for further analysis by a healthcare professional. This approach might offer significant benefits, such as reducing unnecessary patient visits or tests by focusing more on critical readings that deviate from the patient's typical patterns. Moreover, by prioritizing atypical data, the model may offer the potential to detect early signs of

health decline or disease progression, enabling timely intervention and potentially improved patient outcomes.

Finally, as gait assessments become the norm and data becomes more abundant as a result of the recent proliferation of wearable technology, it is only natural for researchers in the gait analysis community to turn to machine learning methods for diagnostics and prognosis problems. These methods enable learning of complex relationships directly from data, eliminating the need for extensive prior knowledge of the underlying physical laws governing the system under analysis. An example can be drawn directly from this thesis, where neural networks have been used in Chapter 4 to predict the severity of the disease. Another example is the use of GPs in Chapter 6 to model the shank angular velocity. The suitability of the GP to model the kinematic data circumvented the need to incorporate any complex biomechanical models. This type of model is commonly termed a 'black-box' model, indicating that its structure is determined by the data rather than by an understanding of the physical system. On the other hand, at the other end of the spectrum, a 'white-box' model is entirely constructed based on the physical knowledge of the system. While traditional physically-derived models present inherent challenges for modelling pathological gait patterns [243, 259], machine learning offers a powerful alternative. However, a critical consideration when applying machine learning to engineering problems, including gait analysis, is the availability of training data. As these algorithms are data-driven, their predictive capabilities are inherently limited to the scenarios and variations represented within the training dataset and hence cannot be used for extrapolation. This behaviour can be effectively seen in Chapter 7, where models leant using the baseline assessment data are not able to effectively predict future behaviours, even though the underlying severity of the disease remains the same.

In response to these limitations, in combination with a natural wish that any inference over the condition of individuals affected by the disease will be informed by both clinical knowledge and the relevant monitoring data available, the future work may also investigate the suitability of physics-informed machine learning models for MS assessments, monitoring and prognosis. This type of modelling combines black-box and white-box approaches, by incorporating biomechanical knowledge as a prior during model training. Building on their success in SHM applications [318, 319], this approach may potentially prove useful within the healthcare sector as well, by improving the generalizability of the models to unseen scenarios and allow

for interpretable insights into the relationship between gait patterns and disease progression in MS patients. In the context of this thesis, physics-informed ML might be used to develop a model that incorporates established biomechanical principles of gait alongside the rich gait data collected to predict disease severity progression with greater accuracy and interpretability.

Appendix A

Definition of quantitative gait metrics

Table A.1 presents the definition of the quantitative gait metrics used in Chapter 3, along with the corresponding mathematical formulas.

Table A.1: Definition of the 36 quantitative gait metrics comprising the initial feature set used in Chapter 3, adapted from [17].

Gait Metrics	**Description**
Spatial Metrics	
Gait Speed (m/s)	$Gait\ speed = \frac{distance\ walked}{ambulation\ time}$
Temporal Metrics	
Stride time (s)	$Stride\ time = IC(i+1) - IC(i)$ $IC = initial\ contact$ $i = left/right\ foot$
Step time (s)	$Step\ time = IC(j) - IC(i)$ $j = left/right\ foot$ $j = right/left\ foot$
Stance time (s)	$Stance\ time = FC(i) - IC(i)$ $FC = final\ contact$
Swing time (s)	$Swing\ time = IC(i+1) - FC(i)$
Single support time (s)	$Single\ support\ time =$ $(IC(i+1) - FC(i)) + (IC(j+1) - FC(j))$
Double support time (s)	$Double\ support\ time =$ $(FC(i) - IC(j)) + (FC(j) - IC(i))$
Swing phase (%)	$Swing\ phase = Swing\ time/Stride\ time$
Double support phase (%)	$Double\ support\ phase =$ $Double\ support\ time/Stride\ time$

Variability Metrics

Stride time (SD) (ms) Step time (SD) (ms) Swing time (SD) (ms) Stance time (SD) (ms)	$SD_{pooled} = \sqrt{\frac{SD(x_{left})^2 + SD(x_{right})^2}{2}}$ SD_{pooled} = *within-subject combined standard deviation* (ms) SD = *standard deviation* x_{left} = *metric of interest (left limb)* x_{right} = *metric of interest (right limb)*
Stride time (CV) (%) Step time (CV) (%) Swing time (CV) (%) Stance time (CV) (%)	$CV(x) = \frac{SD(x)}{\mu_x} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_x)^2}}{\mu_x}$ CV = *coefficient of variation* x = *metric of interest* μ_x = *mean of the metric* N = *number of samples*

Asymmetry Metrics

Step time asymmetry (ms) Swing time asymmetry (ms) Stance time asymmetry (ms)	$x\ asym = \lvert\overline{x_{left}} - \overline{x_{right}}\rvert$ $\overline{x_{left}}$ = *mean of the metric (left limb)* $\overline{x_{right}}$ = *mean of the metric (right limb)*
Step time asymmetry ln (%) Swing time asymmetry ln (%) Stance time asymmetry ln (%)	$x\ asym\_ln = 100 \times \ln\left\lvert\frac{\min(\overline{x_{left}}, \overline{x_{right}})}{\max(\overline{x_{left}}, \overline{x_{right}})}\right\rvert$

Gait Quality Metrics

Root mean square (RMS) m/s^2 [184, 329]	$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} a_R(n)^2}$ a_R = *resultant acceleration* N = *number of samples*
RMS ratio (−) [67]	$RMS\ ratio_{a_x} = \frac{RMS_{a_x}}{RMS_{a_R}}$ a_x = *acceleration component (x = V, ML, AP)* *Higher values denote an increased degree of walking instability.*
Jerk (JK) (m/s^3) [98]	$JK_{a_R} = RMS\left(\frac{da_R}{dt}\right)$

Jerk Ratio (−) [330]	$JK\ ratio_{a_{ML/V}} = 10 \log\left(\frac{RMS(JK_{a_{ML}})}{RMS(JK_{a_V})}\right)$ $JK\ ratio_{a_{AP/V}} = 10 \log\left(\frac{RMS(JK_{a_{AP}})}{RMS(JK_{a_V})}\right)$
Step Regularity (Ad1) and Stride regularity (Ad2) (−)	*The first and second peak magnitude of the unbiased and normalised autocorrelation function* $Ad(t) = \frac{1}{N-\|t\|} \sum_{i=1}^{N-\|t\|} a_R(i) \cdot a_R(i+t)$ $t = time\ lag$ $Ad1, Ad2 \in (0, 1)$ Lower values correspond to an increased level of variability and asymmetry in gait.
Gait symmetry (−) [331]	$Gait\ symmetry = \frac{\|Ad1-Ad2\|}{(Ad1+Ad2)/2}$ Higher values suggest a greater degree of asymmetry between the left and right steps.
Harmonic Ratio (HR) (−) [17]	Ratio of the first 20 harmonic coefficients for each direction over a given number of strides. $HR_{a_{AP},a_V} = \frac{\sum Amplitude\ of\ even\ harmonics}{Amplitude\ of\ odd\ harmonics}$ $HR_{a_{ML}} = \frac{\sum Amplitude\ of\ odd\ harmonics}{Amplitude\ of\ even\ harmonics}$ Higher HR values denote smoother gait patterns.

Appendix B

Hierarchical GP modelling - supporting results

B.1 Contralateral Limb Model Performance Metrics

The performance metrics of the individual-level GP models used in Chapter 6 (aggregating contralateral limb data) are presented in Table B.1.

Table B.1: Performance metrics - individual limb models.

Healthy Controls					MS-affected Individuals					
	NMSE (%)		MSLL				NMSE (%)		MSLL	
Subject No.	Train	Test	Train	Test	*Subject No.*	EDSS	Train	Test	Train	Test
1	0.841	0.885	-4.177	-4.183	1	4.5	2.838	2.821	-3.870	-3.838
2	1.303	1.277	-4.025	-4.023	2	4.5	3.914	3.925	-3.554	-3.528
3	1.012	0.972	-4.153	-4.134	3	4	1.381	1.347	-4.193	-4.208
4	1.260	1.286	-4.086	-4.092	4	4	8.810	9.162	-3.609	-3.584
5	0.591	0.619	-4.135	-4.109	5	5	6.227	6.344	-3.736	-3.773
6	0.562	0.555	-4.321	-4.320	6	5	4.573	4.589	-3.736	-3.711
7	0.557	0.563	-4.245	-4.231	7	4.5	1.434	1.413	-4.167	-4.176
8	1.014	1.046	-4.108	-4.099	8	2	0.877	0.897	-4.122	-4.124
9	0.742	0.735	-4.087	-4.091	9	3.5	7.875	7.907	-4.677	-4.690
10	0.588	0.598	-4.292	-4.278	10	3.5	1.000	1.032	-4.027	-4.018
11	0.634	0.654	-4.474	-4.456	11	2	0.736	0.738	-4.245	-4.238
12	0.793	0.801	-4.218	-4.207	12	5	6.190	6.275	-4.127	-4.132
13	0.626	0.646	-4.382	-4.366	13	4.5	2.062	2.192	-4.068	-4.030
14	0.992	0.972	-4.109	-4.078	14	2.5	2.899	2.942	-3.887	-3.884
15	0.593	0.586	-4.194	-4.193	15	2	1.434	1.429	-3.878	-3.852
16	0.810	0.813	-4.114	-4.127	16	1.5	1.013	1.004	-4.173	-4.187
17	0.778	0.774	-4.175	-4.173	17	3.5	6.648	6.310	-3.549	-3.570
18	1.108	1.113	-3.913	-3.925	18	2	0.620	0.649	-4.237	-4.211
19	0.800	0.790	-4.029	-4.049	19	2	4.265	4.154	-3.772	-3.801
20	0.927	0.929	-4.087	-4.043	20	2	0.966	0.960	-4.315	-4.293
21	0.575	0.583	-4.366	-4.381	21	3.5	1.320	1.335	-4.191	-4.203
22	1.658	1.662	-3.880	-3.882	22	2	2.933	2.919	-3.998	-3.997
23	0.952	0.973	-4.172	-4.180	23	2	0.729	0.763	-4.307	-4.291
24	1.054	1.093	-4.219	-4.204	24	3.5	5.856	5.974	-3.831	-3.775
25	0.703	0.696	-4.424	-4.433	25	3.5	5.321	5.434	-4.473	-4.438
26	0.973	1.007	-4.055	-4.013	26	4	11.427	11.012	-3.571	-3.553
27	1.094	1.094	-4.098	-4.092	27	4	6.722	6.684	-3.664	-3.663
28	0.637	0.634	-4.073	-4.064	28	4	2.792	2.787	-3.807	-3.823

B.2 Wasserstein Asymmetry

Figure B.1 shows the regions of gait cycle asymmetry between contralateral limbs, as quantified by the individual limb models in Section 6.3.5.

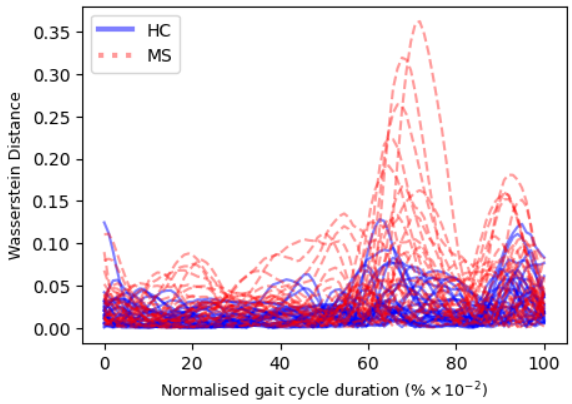


Figure B.1: Wasserstein distance computed between left and right limb models. The blue lines correspond to HCs, while the red lines correspond to PwMS. Each line corresponds to an unique subject.

B.3 Validation of the proposed four-layer HVSHGP model

Figure B.2 shows the comparison of the left individual limb-level GP predictions, against the *held-out* baseline test (BT) data and the follow-up (FU) assessment data, as described in Section 7.1.

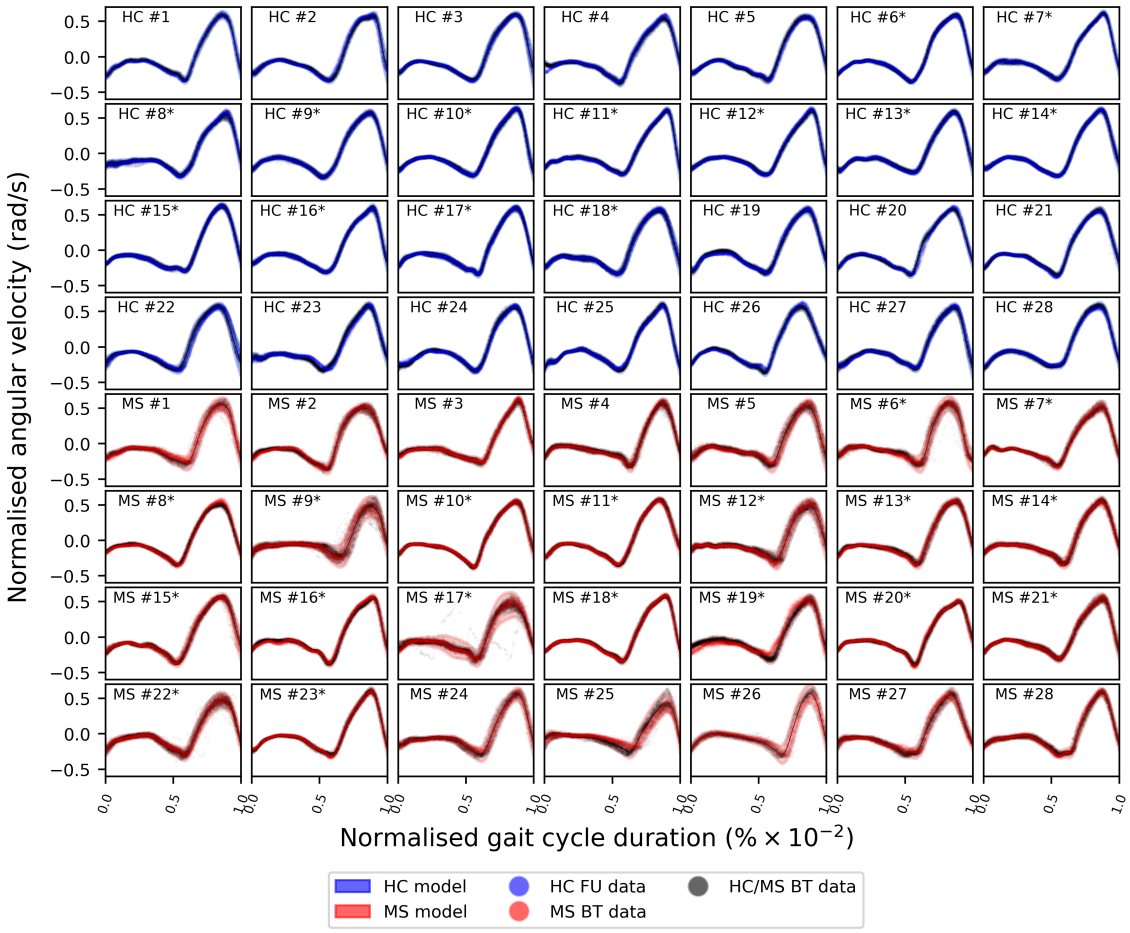


Figure B.2: Comparison of the left individual limb-level GP predictions, *held-out* BT data, and FU data. The first four rows correspond to HC individuals, while the last 4 rows correspond to MS individuals. Individuals corresponding to group A (one-hour apart FU assessment) are denoted by $*$.

Similarly, Figure B.3 shows the comparison of the right individual limb-level GP predictions, against the *held-out* baseline test (BT) data and the follow-up (FU) assessment data.

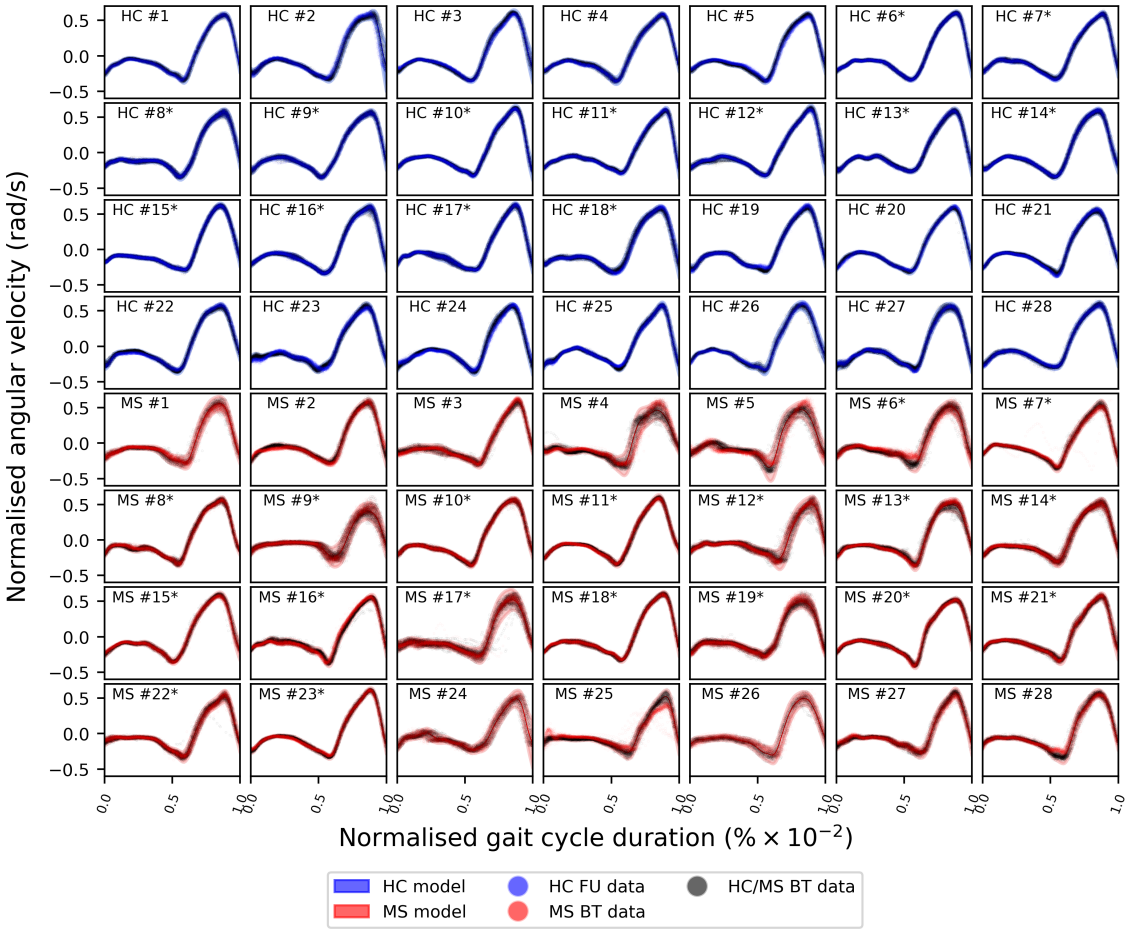


Figure B.3: Comparison of the right individual limb-level GP predictions, *held-out* BT data, and FU data. The first four rows correspond to HC individuals, while the last 4 rows correspond to MS individuals. Individuals corresponding to group A (one-hour apart FU assessment) are denoted by $*$.

B.4 Longitudinal HVSHGP models performance metrics

The performance metrics for the three-layer longitudinal GP models used in Section 7.2 can be found below in Tables B.2 and B.3.

Table B.2: Longitudinal HVSHGP models - performance metrics.

Week 0

Individual Limb (aggregating contralateral limbs)					Individual Limb - Left					Individual Limb - Right				
Subj No.	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**	
	Train	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*
1	17.203	17.370	-1.283	-1.280	1	7.522	8.673	-1.884	-1.798	1	17.259	15.401	-1.254	-1.276
2	25.488	25.406	-0.912	-0.899	2	4.353	4.430	-1.615	-1.590	2	2.255	2.066	-2.303	-2.333
3	14.950	14.505	-1.308	-1.307	3	3.302	3.252	-1.962	-1.971	3	2.403	2.172	-2.242	-2.242
4	4.985	5.034	-1.663	-1.657	4	2.802	2.940	-2.098	-2.091	4	0.924	0.928	-2.556	-2.552
5	15.103	16.510	-1.340	-1.311	5	6.428	8.004	-1.893	-1.776	5	2.893	3.413	-2.003	-1.948
6	4.765	4.738	-1.671	-1.671	6	2.384	2.552	-2.271	-2.257	6	1.684	1.648	-2.347	-2.350
7	23.017	23.108	-1.228	-1.226	7	5.579	5.403	-1.730	-1.740	7	6.643	6.131	-1.878	-1.884
8	5.174	5.460	-1.742	-1.704	8	4.008	4.080	-1.830	-1.813	8	5.314	5.687	-1.823	-1.790
9	6.599	6.470	-1.626	-1.627	9	3.779	3.826	-1.870	-1.865	9	4.065	4.095	-1.783	-1.776
10	7.065	6.807	-1.716	-1.731	10	1.739	1.836	-2.240	-2.216	10	4.643	4.565	-1.817	-1.836
11	6.233	6.158	-1.653	-1.657	11	1.680	1.635	-2.362	-2.379	11	2.330	2.266	-2.026	-2.032
12	5.868	6.137	-1.600	-1.603	12	3.386	3.109	-1.853	-1.885	12	2.516	3.112	-2.278	-2.217
13	2.711	2.822	-2.011	-1.995	13	2.026	2.265	-2.107	-2.081	13	2.226	2.171	-2.153	-2.146
14	8.827	9.000	-1.646	-1.637	14	5.905	5.836	-1.836	-1.837	14	10.078	10.246	-1.649	-1.634
15	21.794	24.377	-1.301	-1.252	15	9.651	11.150	-1.544	-1.489	15	6.876	8.522	-1.947	-1.876
16	12.069	12.436	-1.322	-1.317	16	1.745	1.910	-2.397	-2.371	16	14.042	13.974	-1.365	-1.376

Week 24

Individual Limb (aggregating contralateral limbs)					Individual Limb - Left					Individual Limb - Right				
Subj No.	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**	
	Train	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*
1	4.814	4.846	-1.851	-1.849	1	2.535	2.495	-2.303	-2.313	1	2.809	2.718	-2.216	-2.212
2	22.994	22.213	-0.971	-0.986	2	5.755	5.526	-1.555	-1.564	2	1.851	1.836	-2.270	-2.267
3	8.038	8.116	-1.443	-1.429	3	4.701	4.679	-1.828	-1.803	3	5.827	6.049	-1.608	-1.583
4	8.524	8.253	-1.478	-1.486	4	2.484	2.462	-2.065	-2.055	4	1.186	1.164	-2.490	-2.487
5	22.472	22.446	-1.015	-1.008	5	6.471	6.472	-1.939	-1.932	5	2.709	2.691	-2.020	-2.042
6	4.400	4.242	-1.702	-1.718	6	2.479	2.481	-2.263	-2.246	6	1.884	1.799	-2.341	-2.355
7	18.541	18.378	-1.180	-1.186	7	5.889	5.839	-1.656	-1.661	7	6.110	5.880	-1.809	-1.817
8	4.660	4.439	-1.803	-1.805	8	4.361	4.337	-1.847	-1.831	8	3.615	3.555	-1.914	-1.904
9	7.978	7.900	-1.480	-1.483	9	3.807	3.879	-1.768	-1.764	9	4.932	4.870	-1.709	-1.711
10	11.004	10.850	-1.454	-1.467	10	10.596	9.796	-1.502	-1.541	10	10.227	10.884	-1.490	-1.475
11	5.924	5.718	-1.722	-1.741	11	1.917	1.922	-2.205	-2.215	11	3.329	3.030	-1.915	-1.955
12	11.934	12.561	-1.268	-1.238	12	3.691	3.774	-1.680	-1.673	12	5.939	6.148	-1.924	-1.891
13	2.353	2.406	-1.981	-1.971	13	2.446	2.459	-2.080	-2.082	13	1.054	1.176	-2.388	-2.337
14	7.197	7.171	-1.647	-1.645	14	5.038	5.040	-1.768	-1.765	14	7.839	7.885	-1.680	-1.672
15	31.208	32.348	-1.278	-1.266	15	19.972	20.917	-1.343	-1.345	15	17.447	18.792	-1.492	-1.458
16	3.001	3.138	-1.885	-1.866	16	1.263	1.423	-2.534	-2.485	16	1.232	1.282	-2.497	-2.491

Continued.

Table B.3: Longitudinal HVSHGP models - performance metrics - continued.

Week 48

Individual Limb (aggregating contralateral limbs)					Individual Limb - Left					Individual Limb - Right				
Subj No.	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**	
	Train	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*
1	11.560	12.274	-1.471	-1.457	1	4.792	5.234	-1.930	-1.925	1	5.693	5.784	-2.171	-2.160
2	26.881	26.164	-0.943	-0.957	2	2.865	2.647	-2.149	-2.175	2	18.531	18.961	-1.093	-1.096
3	7.341	7.595	-1.506	-1.493	3	3.238	3.378	-2.024	-2.023	3	3.650	3.718	-1.871	-1.860
4	6.112	6.012	-1.519	-1.526	4	1.759	1.923	-2.174	-2.139	4	1.665	1.572	-2.242	-2.258
5	27.923	28.840	-0.944	-0.918	5	4.185	4.755	-1.987	-1.930	5	4.493	4.706	-1.778	-1.739
6	5.246	5.281	-1.721	-1.713	6	1.797	1.734	-2.370	-2.387	6	2.642	2.662	-2.094	-2.079
7	22.328	21.859	-1.196	-1.198	7	8.342	8.380	-1.536	-1.532	7	5.459	5.090	-1.942	-1.957
8	10.000	9.832	-1.494	-1.498	8	9.160	8.418	-1.630	-1.642	8	6.567	6.704	-1.694	-1.687
9	8.362	8.686	-1.442	-1.426	9	3.781	3.938	-1.866	-1.846	9	4.479	4.696	-1.659	-1.637
10	9.528	9.439	-1.567	-1.574	10	1.960	1.910	-2.221	-2.236	10	3.727	3.877	-1.925	-1.906
11	6.476	6.599	-1.647	-1.636	11	2.639	2.787	-2.149	-2.124	11	3.479	3.486	-1.895	-1.882
12	10.723	10.781	-1.395	-1.392	12	2.990	3.148	-1.923	-1.878	12	3.181	3.151	-2.322	-2.366
13	2.679	2.760	-1.997	-1.983	13	2.713	2.762	-2.017	-2.008	13	1.188	1.244	-2.355	-2.333
14	9.976	10.566	-1.603	-1.585	14	7.452	7.629	-1.686	-1.679	14	10.971	12.036	-1.592	-1.558
15	16.724	18.764	-1.474	-1.416	15	10.225	11.974	-1.577	-1.529	15	6.649	7.785	-1.927	-1.824
16	4.286	4.134	-1.823	-1.830	16	1.797	1.708	-2.322	-2.339	16	1.906	1.830	-2.234	-2.232

Week 96

Individual Limb (aggregating contralateral limbs)					Individual Limb - Left					Individual Limb - Right				
Subj No.	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**		*Subj No.*	**NMSE**		**MSLL**	
	Train	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*		*Train*	*Test*	*Train*	*Test*
1	3.126	3.032	-2.010	-2.016	1	1.380	1.426	-2.587	-2.549	1	3.047	2.832	-2.118	-2.139
2	26.240	26.071	-0.935	-0.944	2	5.561	5.629	-1.598	-1.595	2	3.555	3.913	-2.138	-2.123
3	32.006	32.728	-1.003	-0.992	3	13.675	14.285	-1.530	-1.507	3	15.153	15.054	-2.084	-2.079
4	7.033	7.151	-1.513	-1.502	4	5.293	5.383	-1.750	-1.708	4	3.836	3.666	-1.872	-1.875
5	12.368	12.289	-1.373	-1.386	5	5.859	5.491	-1.710	-1.728	5	7.186	7.274	-1.632	-1.637
6	6.069	6.148	-1.563	-1.560	6	1.880	1.880	-2.376	-2.363	6	2.953	3.091	-2.005	-1.995
7	19.388	18.346	-1.159	-1.194	7	9.125	7.528	-1.503	-1.578	7	6.807	6.915	-1.706	-1.712
8	7.627	7.287	-1.612	-1.630	8	8.797	8.187	-1.581	-1.597	8	5.689	5.572	-1.766	-1.784
9	6.414	6.268	-1.552	-1.552	9	3.575	3.550	-1.870	-1.872	9	4.139	4.130	-1.727	-1.717
10	10.586	10.799	-1.479	-1.467	10	4.309	4.551	-1.952	-1.941	10	4.544	4.738	-1.699	-1.681
11	7.307	7.355	-1.583	-1.583	11	1.453	1.477	-2.403	-2.399	11	3.749	3.967	-1.824	-1.812
12	8.688	8.550	-1.453	-1.461	12	4.343	4.273	-1.691	-1.693	12	1.928	1.925	-2.235	-2.239
13	1.958	2.053	-2.139	-2.121	13	2.176	2.307	-2.118	-2.092	13	1.299	1.318	-2.314	-2.307
14	4.947	4.972	-1.866	-1.861	14	3.463	3.261	-1.979	-1.981	14	5.361	5.164	-1.886	-1.892
15	16.195	16.421	-1.220	-1.217	15	4.933	4.952	-1.693	-1.685	15	3.633	3.470	-2.099	-2.118
16	4.973	5.063	-1.727	-1.724	16	2.141	2.131	-2.186	-2.186	16	1.252	1.269	-2.378	-2.374

B.5 Mixed-Effects Models - Implementation Details

The mixed-effects model used in Section 7.2 for investigating the relationship between disease severity and GP predictive performance metrics is given in Equation B.1.

$$EDSS_{i,t} = \beta_0 + NMSE_{i,t}\beta_1 + MSLL_{i,t}\beta_2 + Time_i\beta_3 + u_i + e_{i,t} \tag{B.1}$$

where $EDSS_{i,t}$ is the EDSS score of participant i at time t, β_0 is the mean intercept, $NMSE_{i,t}$ and $MSLL_{i,t}$ are the explanatory variables, or the fixed effects, β_1, β_2, β_3 are the slopes of the fixed effects, $Time_i$ represents the time in weeks since participant i's baseline assessment at week 0, u_i is the subject-specific random effect or intercept and $e_{i,t}$ is the error term. Here, the random effects account for unobserved variables or confounding factors that take on the same value for all observations of the same participant, thereby helping to reduce the potential bias [312].

In this study, the analysis was performed using Statsmodels 0.14.1 Python package [332]. The model was fitted using a restricted maximum likelihood [333] and included fixed effects for the NMSE, MSLL and the time of assessment, as well as a random intercept for each participant, meaning that each participant was effectively modelled with a different baseline EDSS score. A priori alpha value of $\alpha = 0.05$ was used to indicate statistical significance. The summary results are presented in Table B.4. For clarity, the coefficients represent the expected change in EDSS relative to a unit change in the predictor, assuming that all other variables are held constant.

Table B.4: Summary results for the linear mixed-effects model for disability quantification using baseline HC GP model predictive performance metrics.

	Coefficient	Standard Error	95 % CI		p-value
Intercept	4.977	0.349	4.293	5.660	<0.001***
NMSE	-0.016	0.007	-0.030	-0.002	0.023*
MSLL	0.185	0.074	0.039	0.330	0.013*
Time	0.004	0.001	0.001	0.007	0.003**
Random Effects	0.014	0.033	-0.052	0.079	0.686

CI: Confidence interval
*p-*value*: statistical significance at <0.05, ** <0.01, and *** <0.001.

Similarly, the mixed-effects model used in Section 7.2 for investigating the relationship between disease severity and gait asymmetry metrics is given in Equation B.2. The summary results are presented in Table B.5.

$$EDSS_{i,t} = \beta_0 + WD_{i,t}\beta_1 + \Delta\ KL_{i,t}\beta_2 + Time_i\beta_3 + u_i + e_{i,t} \tag{B.2}$$

where $EDSS_{i,t}$ is the EDSS score of participant i at time t, β_0 is the mean intercept, $WD_{i,t}$ and $\Delta\ KL_{i,t}$ are the explanatory variables, or the fixed effects, β_1, β_2, β_3 are the slopes of the fixed effects, $Time_i$ represents the time in weeks since participant's i's baseline assessment at week 0, u_i is the subject-specific random effect or intercept and $e_{i,t}$ is the error term.

Table B.5: Summary results for the linear mixed-effects model for disability quantification using longitudinal asymmetry measures.

	Coefficient	Standard Error	95 % CI		p-value
Intercept	5.103	0.463	4.196	6.011	<0.001***
WD	0.877	2.329	-3.688	5.441	0.707
$\Delta\ KL \times 10^{-5}$	-7.312	1.766	-10.773	-3.850	<0.001***
Time	0.005	0.001	0.003	0.008	<0.001***
Random Effects	-0.007	0.038	-0.082	0.068	0.861

CI: Confidence intervals

*p-*value*: statistical significance at <0.05, ** <0.01, and *** <0.001.

Bibliography

[1] A. Polhemus *et al.* Walking on common ground : a cross-disciplinary scoping review on the clinical utility of digital mobility outcomes. *npj Digit. Med.*, 4 (149), 2021.

[2] J. J. Geurts and F. Barkhof. Grey matter pathology in multiple sclerosis. *The Lancet Neurology*, 7(9):841–851, 2008.

[3] M. H. Cameron and J. M. Wagner. Gait abnormalities in multiple sclerosis: Pathogenesis, evaluation, and advances in treatment. *Current Neurology and Neuroscience Reports*, 11:507–515, 2011.

[4] P. Browne *et al.* Atlas of multiple sclerosis 2013: A growing global problem with widespread inequity. *Neurology*, 83:1022–1024, 2014.

[5] N. Koch-Henriksen and P. S. Sorensen. The changing demographic pattern of multiple sclerosis epidemiology. *The Lancet Neurology*, 9:520–532, 2010.

[6] U. Givon, G. Zeilig, and A. Achiron. Gait analysis in multiple sclerosis: Characterization of temporal-spatial parameters using gaitrite functional ambulation system. *Gait and Posture*, 29:138–142, 2009.

[7] N. G. LaRocca. Impact of walking impairment in multiple sclerosis: Perspectives of patients and care partners. *Patient*, 4:189–201, 2011.

[8] M. Psarakis *et al.* Wearable technology reveals gait compensations, unstable walking patterns and fatigue in people with multiple sclerosis. *Physiological Measurement*, 39:075004, 2018.

[9] C. Mazzà *et al.* Technical validation of real-world monitoring of gait: a multicentric observational study. *BMJ open*, 11(12), 2021.

[10] M. E. Morris *et al.* Changes in gait and fatigue from morning to afternoon in people with multiple sclerosis. *Journal of Neurology Neurosurgery and Psychiatry*, 72:361–365, 2002.

[11] K. J. Smith and W. I. McDonald. The pathophysiology of multiple sclerosis: The mechanisms underlying the production of symptoms and the natural history of the disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 354:1649–1673, 1999.

[12] B. Trapp *et al.* Relapses and progression of disability in multiple sclerosis. *N Engl J Med*, 338:278–285, 1998.

[13] E. Fisher *et al.* Gray matter atrophy in multiple sclerosis: a longitudinal study. *Annals of Neurology*, 64(3):255–265, 2008.

[14] M. M. Goldenberg. Multiple sclerosis review. *Pharmacy and therapeutics*, 37 (3):175, 2012.

[15] M. Bendszus *et al.* Multiple sclerosis and other demyelinating diseases. In *Inflammatory Diseases of the Brain*, pages 3–18. Springer, 2013.

[16] A. P. Creagh *et al.* Longitudinal trend monitoring of multiple sclerosis ambulation using smartphones. *IEEE Open Journal of Engineering in Medicine and Biology*, 3:202–210, 2022.

[17] L. Angelini *et al.* A multifactorial model of multiple sclerosis gait and its changes across different disability levels. *IEEE Transactions on Biomedical Engineering*, 68(11):3196–3204, 2021.

[18] A. P. Creagh *et al.* Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones. *Scientific Reports 2021 11:1*, 11:1–14, 2021.

[19] C. L. Martin *et al.* Gait and balance impairment in early multiple sclerosis in the absence of clinical disability. *Multiple Sclerosis*, 12:620–628, 2006.

[20] M. Pau *et al.* Clinical assessment of gait in individuals with multiple sclerosis using wearable inertial sensors: Comparison with patient-based measure. *Multiple Sclerosis and Related Disorders*, 10:187–191, 2016.

[21] J. F. Kurtzke. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (edss). *Neurology*, 33:1444–1452, 1983.

[22] L. Angelini *et al.* Wearable sensors can reliably quantify gait alterations associated with disability in people with progressive multiple sclerosis in a clinical setting. *Journal of Neurology*, 267(10):2897–2909, 2020.

[23] A. M. de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors*, 14:3362–3394, 2 2014.

[24] J. V. McLoughlin *et al.* Fatigue induced changes to kinematic and kinetic gait parameters following six minutes of walking in people with multiple sclerosis. *Disability and Rehabilitation*, 38(6):535–543, 2016.

[25] A. A. Ibrahim *et al.* Inertial sensor-based gait parameters reflect patient-reported fatigue in multiple sclerosis. *Journal of NeuroEngineering and Rehabilitation*, 17:165, 2020.

[26] E. Panero *et al.* Comparison of different motion capture setups for gait analysis : Validation of spatio-temporal parameters estimation. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2018.

[27] A. Cappozzo *et al.* Human movement analysis using stereophotogrammetry. part 1: Theoretical background. *Gait and Posture*, 21:186–196, 2005.

[28] M. Alaqtash *et al.* Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 453–457, 2011.

[29] R. Caldas *et al.* A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait and Posture*, 57:204–210, 2017. ISSN 0966-6362.

[30] M. L. Frechette *et al.* Next steps in wearable technology and community ambulation in multiple sclerosis. *Current Neurology and Neuroscience Reports*, 19, 2019.

[31] M. Pau *et al.* Smoothness of gait detects early alterations of walking in persons with multiple sclerosis without disability. *Gait and Posture*, 58:307–309, 2017.

[32] J. M. Huisinga *et al.* Accelerometry reveals differences in gait variability between patients with multiple sclerosis and healthy controls. *Annals of Biomedical Engineering*, 41:1670–679, 2012.

[33] Y. Moon *et al.* Stride-time variability and fall risk in persons with multiple sclerosis. *Multiple Sclerosis International*, 2015:1–7, 2015.

[34] C. Motta *et al.* Disability and fatigue can be objectively measured in multiple sclerosis. *PLOS ONE*, 11, 2016.

[35] M. M. Engelhard *et al.* Quantifying six-minute walk induced gait deterioration with inertial sensors in multiple sclerosis subjects. *Gait and Posture*, 49:340–345, 2016.

[36] Y. Moon *et al.* Monitoring gait in multiple sclerosis with novel wearable motion sensors. *PLOS ONE*, 12, 2017.

[37] R. I. Spain *et al.* Body-worn motion sensors detect balance and gait deficits in people with multiple sclerosis who have normal walking speed. *Gait and Posture*, 35:573–578, 2012.

[38] J. J. Craig *et al.* The relationship between trunk and foot acceleration variability during walking shows minor changes in persons with multiple sclerosis. *Clinical Biomechanics*, 49:16–21, 2017.

[39] S. H. Corporaal *et al.* Balance control in multiple sclerosis: Correlations of trunk sway during stance and gait tests with disease severity. *Gait and Posture*, 37:55–60, 2013.

[40] M. Pau *et al.* Texting while walking differently alters gait patterns in people with multiple sclerosis and healthy individuals. *Multiple Sclerosis and Related Disorders*, 19:129–133, 1 2018.

[41] I. Carpinella *et al.* Instrumental assessment of stair ascent in people with multiple sclerosis, stroke, and parkinson's disease: A wearable-sensor-based approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26:2324–2332, 2018.

[42] R. Kaur *et al.* Predicting multiple sclerosis from gait dynamics using an instrumented treadmill: A machine learning approach. *IEEE Transactions on Biomedical Engineering*, 68(9):2666–2677, 2021.

[43] L. Angelini *et al.* Is a wearable sensor-based characterisation of gait robust enough to overcome differences between measurement protocols? a multi-centric pragmatic study in patients with multiple sclerosis. *Sensors*, 20, 2020.

[44] A. P. Creagh *et al.* Smartphone-and smartwatch-based remote characterisation of ambulation in multiple sclerosis during the two-minute walk test. *IEEE Journal of Biomedical and Health Informatics*, 25:838–849, 2021.

[45] S. Dreyer-Alster *et al.* Longitudinal relationships between disability and gait characteristics in people with ms. *Scientific Reports*, 12:1–10, 2022.

[46] J. M. Gelfand. Chapter 12 - multiple sclerosis: diagnosis, differential diagnosis, and clinical presentation. In D. S. Goodin, editor, *Multiple Sclerosis and Related Disorders*, volume 122 of *Handbook of Clinical Neurology*, pages 269–290. Elsevier, 2014.

[47] S. Saeb *et al.* The need to approximate the use-case in clinical machine learning. *Giga Science*, 6(5), 2017.

[48] A. Stihi *et al.* On gait consistency quantification through arx residual modelling and kernel two-sample testing. *IEEE Transactions on Biomedical Engineering*, 71(3):720–731, 2024.

[49] O. Y. Chén *et al.* Personalized longitudinal assessment of multiple sclerosis using smartphones. *IEEE Journal of Biomedical and Health Informatics*, 27 (7):3633–3644, 2023.

[50] A. A. Ibrahim *et al.* Association between cognition and gait in multiple sclerosis: A smartphone-based longitudinal analysis. *International Journal of Medical Informatics*, 177:105145, 2023.

[51] C. R. Farrar and K. Worden. *Structural Health Monitoring: A Machine Learning Perspective*. Wiley, 2012.

[52] K. Worden and J. M. Dulieu-Barton. An overview of intelligent fault detection in systems and structures. *Structural Health Monitoring*, 3(1):85–98, 2004.

[53] B. Specht *et al.* Multiple sclerosis in the digital health age: Challenges and opportunities-a systematic review. *medRxiv*, pages 2023–11, 2023.

[54] C. Bock *et al.* *Machine Learning for Biomedical Time Series Classification: From Shapelets to Deep Learning.* Humana, New York, NY, 2020.

[55] M. CocaTapia *et al.* Gait pattern in people with multiple sclerosis: a systematic review. *Diagnostics*, 11(4):584, 2021.

[56] R. I. Spain *et al.* Body-worn sensors capture variability, but not decline, of gait and balance measures in multiple sclerosis over 18 months. *Gait and Posture*, 39:958–964, 2014.

[57] A. Vienne-Jumeau *et al.* Wearable inertial sensors provide reliable biomarkers of disease severity in multiple sclerosis: A systematic review and meta-analysis. *Annals of Physical and Rehabilitation Medicine*, 63:138–147, 2020.

[58] A. M. Sabatini, G. Ligorio, and A. Mannini. Fourier-based integration of quasi-periodic gait accelerations for drift-free displacement estimation using inertial sensors. *BioMedical Engineering Online*, 14:106, 2015.

[59] L. I. Lim *et al.* Measuring gait and gait-related activities in parkinson's patients own home environment: A reliability, responsiveness and feasibility study. *Parkinsonism and Related Disorders*, 11:19–24, 2005.

[60] S. Lord, B. Galna, and L. Rochester. Moving forward on gait measurement: Toward a more refined approach. *Movement Disorders*, 28:1534–1543, 2013.

[61] L. Di Biase *et al.* Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, 20 (12):3529, 2020.

[62] F. A. Storm *et al.* Wearable inertial sensors to assess gait during the 6-minute walk test: A systematic review. *Sensors (Switzerland)*, 20, 2020.

[63] J. N. Chee *et al.* Influence of multiple sclerosis on spatiotemporal gait parameters: A systematic review and meta-regression. *Archives of Physical Medicine and Rehabilitation*, 102(9):1801–1815, 2021.

[64] P. Berg-Hansen *et al.* Sensor-based gait analyses of the six-minute walk test identify qualitative improvement in gait parameters of people with multiple sclerosis after rehabilitation. *Journal of Neurology*, 1, 2022.

[65] B. Fan, Q. Li, and T. Liu. Accurate foot clearance estimation during level and uneven ground walking using inertial sensors. *Measurement Science and Technology*, 31(5):055106, 2020.

[66] A. Godfrey *et al.* Instrumenting gait with an accelerometer: A system and algorithm examination. *Medical Engineering and Physics*, 37:400–407, 2015.

[67] M. Sekine *et al.* A gait abnormality measure based on root mean square of trunk acceleration. *Journal of NeuroEngineering and Rehabilitation*, 10, 12 2013.

[68] R. P. Young and R. G. Marteniuk. Acquisition of a multi-articular kicking task: Jerk analysis demonstrates movements do not become smoother with learning. *Human Movement Science*, 16:677–701, 1997.

[69] R. Moe-Nilssen and J. L. Helbostad. Estimation of gait cycle characteristics by trunk accelerometry. *Journal of Biomechanics*, 37:121–126, 2004.

[70] C. Buckley *et al.* Gait asymmetry post-stroke: Determining valid and reliable methods using a single accelerometer located on the trunk. *Sensors (Switzerland)*, 20, 2020.

[71] S. Yang *et al.* Estimation of spatio-temporal parameters for post-stroke hemiparetic gait using inertial sensors. *Gait and Posture*, 37:354–358, 2013.

[72] D. Trojaniello *et al.* Estimation of step-by-step spatio-temporal parameters of normal and impaired gait using shank-mounted magneto-inertial sensors: application to elderly, hemiparetic, parkinsonian and choreic gait. *Journal of neuroengineering and rehabilitation*, 11:1–12, 2014.

[73] O. Dehzangi, M. Taherisadr, and R. ChangalVala. Imu-based gait recognition using convolutional neural networks and multi-sensor fusion. *Sensors*, 17:2735, 2017.

[74] G. P. Panebianco *et al.* Analysis of the performance of 17 algorithms from a systematic review: Influence of sensor position, analysed variable and computational approach in gait timing estimation from imu measurements. *Gait and Posture*, 66:76–82, 2018.

[75] A. Supratak *et al.* Remote monitoring in the home validates clinical gait measures for multiple sclerosis. *Frontiers in Neurology*, 9:561, 2018.

[76] T. Chitnis *et al.* Quantifying neurologic disease using biosensor measurements in-clinic and in free-living settings in multiple sclerosis. *npj Digital Medicine*, 2019.

[77] M. Ullrich *et al.* Detection of gait from continuous inertial sensor data using harmonic frequencies. *IEEE Journal of Biomedical and Health Informatics*, pages 2168–2194, 2020.

[78] L. Palmerini *et al.* Mobility recorded by wearable devices and gold standards: the mobilise-d procedure for data standardization. *Scientific Data*, 10(1):38, 2023.

[79] M. E. Micó-Amigo *et al.* Assessing real-world gait with digital technology? validation, insights and recommendations from the mobilise-d consortium. *Journal of neuroengineering and rehabilitation*, 20(1):78, 2023.

[80] F. A. Storm, C. J. Buckley, and C. Mazzà. Gait event detection in laboratory and real life settings: Accuracy of ankle and waist sensor based methods. *Gait & posture*, 50:42–46, 2016.

[81] W. Niswander and K. Kontson. Evaluating the impact of imu sensor location and walking task on accuracy of gait event detection algorithms. *Sensors*, 21 (12):3989, 2021.

[82] R. Romijnders *et al.* Validation of imu-based gait event detection during curved walking and turning in older adults and parkinson's disease patients. *Journal of neuroengineering and rehabilitation*, 18(1):28, 2021.

[83] R. Romijnders *et al.* A deep learning approach for gait event detection from a single shank-worn imu: Validation in healthy and neurological cohorts. *Sensors*, 22(10):3859, 2022.

[84] L. Filli *et al.* Profiling walking dysfunction in multiple sclerosis: characterisation, classification and progression over time. *Scientific Reports*, 8(1):1–13, 2018.

[85] M. Pau *et al.* Kinematic analysis of lower limb joint asymmetry during gait in people with multiple sclerosis. *Symmetry*, 13(4):598, 2021.

[86] D. W. Kim *et al.* A comparison of activity monitor data from devices worn on the wrist and the waist in people with parkinson's disease. *Movement disorders clinical practice*, 6(8):693–699, 2019.

[87] R. Morris *et al.* A model of free-living gait: A factor analysis in parkinson's disease. *Gait and Posture*, 52:68–71, 2017.

[88] S. Lord *et al.* Independent domains of gait in older adults and associated motor and nonmotor attributes: Validation of a factor analysis approach. *J Gerontol A Biol Sci Med Sci*, 68:820–827, 2013.

[89] J. Verghese *et al.* Quantitative gait markers and incident fall risk in older adults. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 64:896–901, 2009.

[90] J. H. Hollman *et al.* Normative spatiotemporal gait parameters in older adults. *Gait and Posture*, 34:111–118, 2011.

[91] R. Z. U. Rehman *et al.* Accelerometry-based digital gait characteristics for classification of parkinson's disease: What counts? *IEEE Open Journal of Engineering in Medicine and Biology*, 1:65–73, 2020.

[92] F. B. Horak *et al.* Balance and gait represent independent domains of mobility in parkinson disease, 2016.

[93] M. Wuehr *et al.* Independent domains of daily mobility in patients with neurological gait disorders. *Journal of Neurology*, 267:292–300, 2020.

[94] P. Thingstad *et al.* Identification of gait domains and key gait variables following hip fracture. *BMC Geriatrics*, 15, 2015.

[95] S. Shema-Shiratzky *et al.* Deterioration of specific aspects of gait during the instrumented 6-min walk test among people with multiple sclerosis. *Journal of Neurology*, 266:3022–3030, 2019.

[96] I. Pasciuto *et al.* Overcoming the limitations of the harmonic ratio for the reliable assessment of gait symmetry. *Journal of Biomechanics*, 53:84–89, 2017.

[97] C. Buckley *et al.* Upper body accelerations as a biomarker of gait impairment in the early stages of parkinson's disease. *Gait and Posture*, 71:289–295, 6 2019.

[98] P. Fazio *et al.* Gait measures with a triaxial accelerometer among patients with neurological impairment. *Neurological Sciences*, 34:435–440, 2013.

[99] A. Caronni *et al.* Local dynamic stability of gait in people with early multiple sclerosis and no-to-mild neurological impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28:1389–1396, 2020.

[100] P. Khera and N. Kumar. Role of machine learning in gait analysis: a review, 2020.

[101] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu. Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[102] C. M. Bishop and H. Bishop. *Deep Learning: Foundations and Concepts.* Springer International Publishing, 2024.

[103] E. Halilaj *et al.* Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, 81:1–11, 2018.

[104] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[105] A. Webb. *Statistical Pattern Recognition.* Wiley InterScience electronic collection. Wiley, 2003.

[106] F. M. H. Oliveira, A. R. P. Machado, and A. O. Andrade. On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with parkinson's disease. *Computational and mathematical methods in medicine*, 2018.

[107] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[108] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, pages 833–840, 2002.

[109] D. Slijepcevic *et al.* Explaining machine learning models for clinical gait analysis. *ACM Transactions on Computing for Healthcare*, 3(2):1–27, 2021.

[110] M.-J. Yang *et al.* Combining feature ranking with pca: An application to gait analysis. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pages 494–499, 2010.

[111] O. Vyšata *et al.* Classification of ataxic gait. *Sensors*, 21(16), 2021.

[112] M. Gavrilović and D. B. Popović. A principal component analysis (pca) based assessment of the gait performance. *Biomedical Engineering*, 66(5):449–457, 2021.

[113] D. Thakur, A. Guzzo, and G. Fortino. t-sne and pca in ensemble learning based human activity recognition with smartwatch. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–6, 2021.

[114] D. Martínez-Pascual *et al.* Gait activity classification with convolutional neural network using lower limb angle measurement from inertial sensors. *IEEE Sensors Journal*, 24(13):21479–21489, 2024.

[115] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517, 10 2007.

[116] A. Tsanas. Relevance, redundancy, and complementarity trade-off (rrct): A principled, generic, robust feature-selection tool. *Patterns*, 3(5), 2022.

[117] B. Caby *et al.* Feature extraction and selection for objective gait analysis and fall risk assessment by accelerometry. *BioMedical Engineering Online*, 10, 2011.

[118] T. W. Yeoh *et al.* On the effectiveness of feature selection methods for gait classification under different covariate factors. *Applied Soft Computing*, 61: 42–57, 2017.

[119] A. Phinyomark *et al.* Analysis of big data in gait biomechanics: Current trends and future directions. *Journal of medical and biological engineering*, 38:244–260, 2018.

[120] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán. Filter methods for feature selection - a comparative study. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4881 LNCS, pages 178–187, 2007.

[121] M. A. Hall and L. A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, page 235–239, 1999.

[122] Kurniabudi *et al.* Cicids-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access*, 8, 2020.

[123] J. Howcroft, J. Kofman, and E. D. Lemaire. Feature selection for elderly faller classification based on wearable sensors. *Journal of NeuroEngineering and Rehabilitation*, 14, 2017.

[124] K. Worden *et al.* The fundamental axioms of structural health monitoring. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463:1639–1664, 2007.

[125] K. Worden, G. Manson, and N. R. Fieller. Damage detection using outlier analysis. *Journal of Sound and Vibration*, 229:647–667, 2000.

[126] G. Y. Mclachlan. Mahalanobis distance. *Resonance*, pages 20–26, 1999.

[127] P. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences (Calcutta)*, pages 49–55, 1936.

[128] N. Dervilis, K. Worden, and E. J. Cross. On robust regression analysis as a means of exploring environmental and operational conditions for shm data. *Journal of Sound and Vibration*, 347:279–296, 7 2015.

[129] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

[130] N. Dervilis *et al.* Robust methods of inclusive outlier analysis for structural health monitoring. *Journal of Sound and Vibration*, 333:5181–5195, 2014.

[131] J. J. Moughty and J. R. Casas. Performance assessment of vibration parameters as damage indicators for bridge structures under ambient excitation. *Procedia engineering*, 199:1970–1975, 2017.

[132] J. Gong, M. D. Goldman, and J. Lach. Deepmotion: A deep convolutional neural network on inertial body sensors for gait assessment in multiple sclerosis. In *IEEE Wireless Health*, pages 164–171. Institute of Electrical and Electronics Engineers Inc., 2016.

[133] M. Gadaleta, L. Merelli, and M. Rossi. Human authentication from ankle motion data using convolutional neural networks. *IEEE Workshop on Statistical Signal Processing Proceedings*, 2016.

[134] S. Münzner *et al.* Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM international symposium on wearable computers*, pages 158–165, 2017.

[135] F. Moya Rueda *et al.* Convolutional neural networks for human activity recognition using body-worn sensors. In *Informatics*, volume 5, page 26. MDPI, 2018.

[136] F. Horst, S. Lapuschkin, W. Samek, K. R. Müller, and W. I. Schöllhorn. Explaining the unique nature of individual gait patterns with deep learning. *Scientific Reports*, 9:1–13, 2019.

[137] K. Xia, J. Huang, and H. Wang. Lstm-cnn architecture for human activity recognition. *IEEE Access*, 8:56855–56866, 2020.

[138] S. Kiranyaz *et al.* 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021.

[139] O. Adel, M. Soliman, and W. Gomaa. Inertial gait-based person authentication using siamese networks. In *2021 International joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2021.

[140] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[141] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[142] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[143] T. Chen *et al.* A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[144] D. Baehrens *et al.* How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010.

[145] S. Wolf *et al.* Automated feature assessment in instrumented gait analysis. *Gait and Posture*, 23:331–338, 2006.

[146] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4793–4813, 2021.

[147] A. Holzinger *et al.* What do we need to build explainable ai systems for the medical domain?, 2017.

[148] L. A. Hendricks *et al.* Generating visual explanations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 3–19. Springer Verlag, 2016.

[149] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 818–833, 2014.

[150] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3449–3457, 2017.

[151] A. Shrikumar *et al.* Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[152] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[153] D. Smilkov *et al.* Smoothgrad: removing noise by adding noise, 2017.

[154] S. Bach *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 7 2015.

[155] G. Montavon *et al.* Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[156] J. Zhang *et al.* Top-down neural attention by excitation backprop. *Int. J. Comput. Vision*, 126:1084–1102, 2018.

[157] J. Jenkins and C. Ellis. Using ground reaction forces from gait analysis: Body mass as a weak biometric. In *Pervasive Computing: 5th International Conference, PERVASIVE 2007, Toronto, Canada, May 13-16, 2007. Proceedings 5*, pages 251–267. Springer, 2007.

[158] C. R. Farrar and K. Worden. An introduction to structural health. *New Trends in Vibration Based Structural Health Monitoring*, 520(1), 2012.

[159] L. A. C. Nogueira *et al.* Gait characteristics of multiple sclerosis patients in the absence of clinical disability. *Disability and Rehabilitation*, 35(17):1472–1478, 2013.

[160] H. Inojosa *et al.* A focus on secondary progressive multiple sclerosis (spms): challenges in diagnosis and definition, 2019.

[161] S. Da Silva, M. Dias, and V. Lopes. Damage detection in a benchmark structure using ar-arx models and statistical pattern recognition. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 29(2):174–184, 2007.

[162] H. Sohn *et al.* Structural health monitoring using statistical pattern recognition techniques. *Journal of Dynamic Systems, Measurement, and Control*, 123(4): 706–711, 2001.

[163] H. Sohn, J. A. Czarnecki, and C. R. Farrar. Structural health monitoring using statistical process control. *Journal of Structural Engineering*, 126(11): 1356–1363, 2000.

[164] A. Entezami, H. Shariatmadar, and S. Mariani. Early damage assessment in large-scale structures by innovative statistical pattern recognition methods based on time series modeling and novelty detection. *Advances in Engineering Software*, 150, 2020.

[165] H. Mahzoun Alzakerin, Y. Halkiadakis, and K. D. Morgan. Characterizing gait pattern dynamics during symmetric and asymmetric walking using autoregressive modeling. *PloS one*, 15(12):e0243221, 2020.

[166] C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[167] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:283–298, 2008.

[168] C. Chun *et al.* Gaussian process learning and interpolation of gait motion for rehabilitation robots. In *2015 6th International Conference on Automation, Robotics and Applications (ICARA)*, pages 198–203, 2015.

[169] J. Hong *et al.* Gaussian process trajectory learning and synthesis of individualized gait motions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1236–1245, 2019.

[170] J. Hong *et al.* Gaussian process trajectory learning and synthesis of individualized gait motions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1236–1245, 2019.

[171] Y. Yun *et al.* Statistical method for prediction of gait kinematics with gaussian process regression. *Journal of Biomechanics*, 47(1):186–192, 2014.

[172] C. Glackin *et al.* Gait trajectory prediction using gaussian process ensembles. In *IEEE-RAS International Conference on Humanoid Robots*, pages 693–633, 2014.

[173] X. Wu *et al.* Individualized gait pattern generation for sharing lower limb exoskeleton robot. *IEEE Transactions on Automation Science and Engineering*, 15(4):1459–1470, 2018.

[174] Z. Chen *et al.* Gait prediction and variable admittance control for lower limb exoskeleton with measurement delay and extended-state-observer. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8693–8706, 2023.

[175] I. Benemerito *et al.* Reducing the complexity of musculoskeletal models using gaussian process emulators. *Applied Sciences (Switzerland)*, 12, 2022.

[176] M. Lázaro-Gredilla and M. K. Titsias. Variational heteroscedastic gaussian process regression. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 841–848, Madison, WI, USA, 2011. Omnipress.

[177] R. Domingues *et al.* A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern recognition*, 74:406–421, 2018.

[178] E. Morel *et al.* Gait profile score in multiple sclerosis patients with low disability. *Gait and Posture*, 51:169–173, 2017.

[179] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley, 1987.

[180] P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[181] P. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications, B*, pages 283–297, 1985.

[182] S. Verboven and M. Hubert. Libra: A matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75:127–136, 2005.

[183] D. W. Aha and R. L. Bankert. *A Comparative Evaluation of Sequential Feature Selection Algorithms*, pages 199–206. Springer, New York, NY, 1996.

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

[185] A. Salarian *et al.* Gait assessment in parkinson's disease: toward an ambulatory system for long-term monitoring. *IEEE Transactions on Biomedical Engineering*, 51(8):1434–1443, 2004.

[186] B. M. Eskofier *et al.* Marker-based classification of young-elderly gait pattern differences via direct pca feature extraction and svms. *Computer Methods in Biomechanics and Biomedical Engineering*, 16:435–442, 2003.

[187] G. Yogev *et al.* Gait asymmetry in patients with parkinson's disease and elderly fallers: When does the bilateral coordination of gait require attention? *Experimental Brain Research*, 177:336–346, 2007.

[188] T. Chau. A review of analytical techniques for gait data. part 1: Fuzzy, statistical and fractal methods. *Gait and Posture*, 13:49–66, 2001.

[189] D. Drover *et al.* Faller classification in older adults using wearable sensors based on turn and straight-walking accelerometer-based features. *Sensors (Switzerland)*, 17, 2017.

[190] J. Howcroft, J. Kofman, and E. D. Lemaire. Prospective fall-risk prediction models for older adults based on wearable sensors. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25:1812–1820, 2017.

[191] A. Mannini *et al.* A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington's disease patients. *Sensors (Switzerland)*, 16, 2016.

[192] E. Sejdic *et al.* A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22:603–612, 2014.

[193] G. Corder and D. Foreman. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley, 2011.

[194] M. J. Socie and J. J. Sosnoff. Gait variability and multiple sclerosis. *Multiple Sclerosis International*, 2013, 2013.

[195] S. R. Dandu *et al.* Understanding the physiological significance of four inertial gait features in multiple sclerosis. *IEEE Journal of Biomedical and Health Informatics*, 22:40–46, 2018.

[196] C. Chotiyarnwong *et al.* Effect of remote ischaemic preconditioning on walking in people with multiple sclerosis: double-blind randomised controlled trial. *BMJ Neurology Open*, 2(1), 2020.

[197] I. Arpan *et al.* Fall prediction based on instrumented measures of gait and turning in daily life in people with multiple sclerosis. *Sensors*, 22(16), 2022.

[198] G. Adusumilli *et al.* Turning is an important marker of balance confidence and walking limitation in persons with multiple sclerosis. *PLoS ONE*, 13, 2018.

[199] M. J. Socie *et al.* Footfall placement variability and falls in multiple sclerosis. *Annals of Biomedical Engineering*, 41(8):1740–1747, 2013.

[200] S. Aich *et al.* A validation study of freezing of gait (fog) detection and machine-learning-based fog prediction using estimated gait characteristics with a wearable accelerometer. *Sensors (Switzerland)*, 18, 2018.

[201] B. T. Cole *et al.* Dynamic neural network detection of tremor and dyskinesia from wearable sensor data. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pages 6062–6065, 2010.

[202] J. G. Remelius *et al.* Gait impairments in persons with multiple sclerosis across preferred and fixed walking speeds. *Archives of Physical Medicine and Rehabilitation*, 93:1637–1642, 2012.

[203] L. Comber, R. Galvin, and S. Coote. Gait deficits in people with multiple sclerosis: A systematic review and meta-analysis. *Gait and Posture*, 51:25–35, 2017.

[204] C. Kirk *et al.* Mobilise-ịnsights to estimate real-world walking speed in multiple conditions with a wearable device. *Scientific Reports*, 14(1):1754, 2024.

[205] T. Woelfle *et al.* Wearable sensor technologies to assess motor functions in people with multiple sclerosis: systematic scoping review and perspective. *Journal of Medical Internet Research*, 25, 2023.

[206] A. Keogh *et al.* Acceptability of wearable devices for measuring mobility remotely: observations from the mobilise-d technical validation study. *Digital Health*, 9:20552076221150745, 2023.

[207] S. Del Din *et al.* Free-living monitoring of parkinson's disease: Lessons from the field. *Movement Disorders*, 31(9):1293–1313, 2016.

[208] H. Qian, T. Tian, and C. Miao. What makes good contrastive learning on small-scale wearable-based tasks? In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3761–3771, 2022.

[209] J.-B. Grill *et al.* Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[210] W. Samek *et al. Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[211] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[212] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[213] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[214] M. Andrychowicz *et al.* Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[215] Y. Nesterov. A method of solving a convex programming problem with convergence rate o $(1/\hat{k2})$. *Doklady Akademii Nauk SSSR*, 269(3):543, 1983.

[216] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12 (7), 2011.

[217] T. Tieleman and G. Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical report*, 2017.

[218] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[219] T. Dozat. Incorporating Nesterov Momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–4, 2016.

[220] J. Redmon *et al.* You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[221] C. Szegedy *et al.* Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[222] C. Esteban, S. L. Hyland, and G. Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[223] G. Iglesias *et al.* Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, 35:10123–10145, 2023.

[224] A. Desai *et al.* Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

[225] X. Li *et al.* Tts-gan: A transformer-based time-series generative adversarial network. In *International Conference on Artificial Intelligence in Medicine*, pages 133–143. Springer, 2022.

[226] L. Lin *et al.* Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, pages 1–23, 2023.

[227] X. Yuan and Y. Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

[228] P. Khosla *et al.* Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[229] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[230] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20: 273–297, 1995.

[231] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[232] G. Montavon *et al.* *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. 2019.

[233] M. Kohlbrenner *et al.* Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[234] M. Alber *et al.* innvestigate neural networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019.

[235] M. Cohen *et al.* Should we still only rely on edss to evaluate disability in multiple sclerosis patients? a study of inter and intra rater reliability. *Multiple Sclerosis and Related Disorders*, 54:103144, 2021.

[236] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016.

[237] H. I. Fawaz *et al.* Transfer learning for time series classification. In *2018 IEEE international conference on big data (Big Data)*, pages 1367–1376. IEEE, 2018.

[238] C. o. Huang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[239] R. R. Neptune *et al.* Contributions of the individual ankle plantar flexors to support, forward progression and swing initiation during walking. 34(11): 1387–1398, 2001.

[240] K. Kelleher *et al.* The effect of textured insoles on gait patterns of people with multiple sclerosis. *Gait and Posture*, 32(1):67–71, 2010.

[241] J. M. Huisinga *et al.* Gait mechanics are different between healthy controls and patients with multiple sclerosis. *Journal of applied biomechanics*, 29(3): 303–311, 2013.

[242] R. W. Motl *et al.* Evidence for the different physiological significance of the 6-and 2-minute walk tests in multiple sclerosis. *BMC neurology*, 12:1–7, 2012.

[243] A. D. Kuo. The six determinants of gait and the inverted pendulum analogy: A dynamic walking perspective. *Human Movement Science*, 26(4):617–656, 2007.

[244] F. Pecoraro *et al.* Assessment of level-walking aperiodicity. *Journal of NeuroEngineering and Rehabilitation*, 3, 2006.

[245] J. Chakraborty and A. Nandy. Periodicity detection of quasi-periodic slow-speed gait signal using imu sensor. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body and Motion*, pages 140–152. Springer International Publishing, 2019.

[246] L. Ljung. *System Identification, Theory for the User*. Prentice Hal PTR, 1999.

[247] A. Gretton *et al.* A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[248] F. Kluge *et al.* Towards mobile gait analysis: Concurrent validity and test-retest reliability of an inertial measurement system for the assessment of spatio-temporal gait parameters. *Sensors*, 17(7), 2017.

[249] P. Decavel, T. Moulin, and Y. Sagawa. Gait tests in multiple sclerosis: Reliability and cut-off values. *Gait and Posture*, 67:37–42, 2019.

[250] D. M. Urquhart, M. E. Morris, and R. Iansek. Gait consistency over a 7-day interval in people with parkinson's disease. *Archives of Physical Medicine and Rehabilitation*, 80(6):696–701, 1999.

[251] D. A. Winter. *Biomechanics and Motor Control of Human Movement*. John Wiley & Sons, 4th ed. edition, 2009.

[252] A. Sidiropoulos, R. Magill, and A. Gordon. Coordination of the upper and lower extremities during walking in children with cerebral palsy. *Gait and Posture*, 86:251–255, 2021. ISSN 0966-6362.

[253] M. G. Pandy, Y.-c. Lin, and H. J. Kim. Muscle coordination of mediolateral balance in normal walking. *Journal of Biomechanics*, 43(11):2055–2064, 2010.

[254] J. L. Allen and R. R. Neptune. Three-dimensional modular control of human walking. *Journal of Biomechanics*, 45(12):2157–2163, 2012.

[255] W. Zijlstra and A. L. Hof. Displacement of the pelvis during human walking: experimental data and model predictions. *Gait & Posture*, 6(3):249–262, 1997.

[256] J. M. Hausdorff, D. A. Rios, and H. K. Edelberg. Gait variability and fall risk in community-living older adults: A 1-year prospective study. *Archives of Physical Medicine and Rehabilitation*, 82(8):1050–1056, 2001.

[257] S. B. Richmond *et al.* A temporal analysis of bilateral gait coordination in people with multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 45, 2020.

[258] Y. Mao *et al.* Estimation of stride-by-stride spatial gait parameters using inertial measurement unit attached to the shank with inverted pendulum model. *Scientific Reports*, 11(1), 2021.

[259] G. A. Cavagna, H. Thys, and A. Zamboni. The sources of external work in level walking and running. *The Journal of Physiology*, 262(3):639–657, 1976.

[260] F. Rasouli and K. B. Reed. Identical limb dynamics for unilateral impairments through biomechanical equivalence. *Symmetry*, 13(4), 2021.

[261] E. Peter Carden and J. M. Brownjohn. Rma modelled time-series classification for structural health monitoring of civil infrastructure. *Mechanical Systems and Signal Processing*, 22(2):295–314, 2008.

[262] B. K. Sriperumbudur *et al.* Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2009.

[263] G. E. P. Box *et al.* *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[264] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

[265] P. Stoica and Y. Selén. A review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

[266] A. Smola *et al.* A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer Berlin Heidelberg, 2007.

[267] K. Fukumizu *et al.* Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, 20, 2007.

[268] R. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, Ltd, 1980.

[269] J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[270] A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018.

[271] A. Gretton *et al.* A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20, 2007.

[272] D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic, 2018.

[273] A. Gretton *et al.* Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 25, 2012.

[274] R. H. Byrd, J. C. Gilbert, and J. Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, 89:149–185, 2000.

[275] R. A. Waltz *et al.* An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical Programming*, 107: 391–408, 2006.

[276] R. H. Byrd, M. E. Hribar, and J. Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4): 877–900, 2006.

[277] G. E. Forsythe, M. A. Malcolm, and C. B. Moler. *Computer Methods for Mathematical Computations*. Prentice Hall PTR, Englewood Cliffs, New Jersey, 1979.

[278] R. P. Brent. *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, New Jersey, 1st edition, 1973.

[279] J. Cohen. Chapter 3 - the significance of a product moment rs. In *Statistical Power Analysis for the Behavioral Sciences*, pages 75–107. Academic Press, 1977.

[280] H. Sohn, K. Worden, and C. R. Farrar. Statistical damage classification under changing environmental and operational conditions. *Journal of Intelligent Material Systems and Structures*, 13(9):561–574, 2002.

[281] H. Sohn. Effects of environmental and operational variability on structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1851):539–560, 2006.

[282] E. J. Cross, K. Worden, and Q. Chen. Cointegration: a novel approach for the removal of environmental trends in structural health monitoring data. *Proc. R. Soc. A*, 467:2712–2732, 2011.

[283] K. Worden *et al.* A multiresolution approach to cointegration for enhanced shm of structures under varying conditions - an exploratory study. *Mechanical Systems and Signal Processing*, 47(1):243–262, 2014.

[284] P. Feys *et al.* Within-day variability on short and long walking tests in persons with multiple sclerosis. *Journal of the Neurological Sciences*, 338(1):183–187, 2014.

[285] W. Zaremba, A. Gretton, and M. Blaschko. B-tests: Low variance kernel two-sample tests. *Advances in Neural Information Processing Systems*, 2013.

[286] M. Pau *et al.* Kinematic analysis of lower limb joint asymmetry during gait in people with multiple sclerosis. *Symmetry*, 13, 4 2021.

[287] Y. Moon *et al.* Gait variability in people with neurological disorders: A systematic review and meta-analysis. *Human Movement Science*, 47:197–208, 2016.

[288] S. J. Crenshaw and J. G. Richards. A method for analyzing joint symmetry and normalcy, with an application to analyzing gait. *Gait and Posture*, 24(4): 515–521, 2006.

[289] G. Severini *et al.* Evaluation of clinical gait analysis parameters in patients affected by multiple sclerosis: Analysis of kinematics. *Clinical Biomechanics*, 45:1–8, 6 2017.

[290] R. Salehi *et al.* Comparison of the lower limb inter-segmental coordination during walking between healthy controls and people with multiple sclerosis with and without fall history. *Multiple Sclerosis and Related Disorders*, 41, 2020.

[291] J. Gil-Castillo *et al.* Advances in neuroprosthetic management of foot drop: A review. *Journal of NeuroEngineering and Rehabilitation*, 17, 2020.

[292] D. Rodríguez-Martín *et al.* Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. *PLOS ONE*, 12(2):1–26, 2017.

[293] L. Ingelse *et al.* Personalised gait recognition for people with neurological conditions. *Sensors*, 22(11), 2022.

[294] J. Hensman, N. D. Lawrence, and M. Rattray. Hierarchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, 14(1):1–12, 2013.

[295] M. K. Titsias. Variational learning of inducing variables in sparse gaussian processes. *Journal of Machine Learning Research*, 5:567–574, 2009.

[296] T. Rogers *et al.* Gaussian processes. In *Machine Learning in Modeling and Simulation: Methods and Applications*, pages 121–147. Springer, 2023.

[297] A. B. Abdessalem *et al.* Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential monte carlo. *Frontiers in Built Environment*, 3, 2017.

[298] C. E. Rasmussen and Z. Ghahramani. Occam's razor. In *Advances in neural information processing systems.*, pages 294–300, 2001.

[299] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005.

[300] T. D. Bui, J. Yan, and R. E. Turner. A unifying framework for gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18:1–72, 2017.

[301] J. Hensman, N. Durrande, and A. Solin. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

[302] T. Michalis. Variational model selection for sparse gaussian process regression. Technical report, School of Computer Science, University of Manchester, 2009.

[303] H. Liu, Y. S. Ong, and J. Cai. Large-scale heteroscedastic regression via gaussian process. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 708–721, 2018.

[304] N. Haji Ghassemi *et al.* Segmentation of gait sequences in sensor-based movement analysis: A comparison of methods in parkinson's disease. *Sensors*, 18, 2018.

[305] A. G. de G. Matthews *et al.* Gpflow: A gaussian process library using tensorflow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

[306] D. A. Neumann. *Kinesiology of the musculoskeletal system: foundations for rehabilitation.* Elsevier Health Sciences, 2016.

[307] C. Villani. *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[308] F. Horst *et al.* Daily changes of individual gait patterns identified by means of support vector machines. *Gait and Posture*, 49:309–314, 2016.

[309] B. A. Cree *et al.* Secondary progressive multiple sclerosis: new insights. *Neurology*, 97(8):378–388, 2021.

[310] M. J. Lindstrom and D. M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014–1022, 1988.

[311] A. Cnaan, N. M. Laird, and P. Slasor. *Mixed Models: Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data*, chapter 1, pages 127–158. 2004.

[312] C. Hsiao. Panel data analysis—advantages and challenges. *Test*, 16(1):1–22, 2007.

[313] H. Schielzeth *et al.* Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, 11(9):1141–1152, 2020.

[314] J. Noseworthy *et al.* Interrater variability with the expanded disability status scale (edss) and functional systems (fs) in a multiple sclerosis clinical trial. *Neurology*, 40(6):971–971, 1990.

[315] S. Meyer-Moock *et al.* Systematic literature review and validity evaluation of the expanded disability status scale (edss) and the multiple sclerosis functional composite (msfc) in patients with multiple sclerosis. *BMC neurology*, 14:1–10, 2014.

[316] M. P. Galea *et al.* Gait and balance deterioration over a 12-month period in multiple sclerosis patients with edss scores $\leq$3.0. *NeuroRehabilitation*, 40(2): 277–284, 2017.

[317] A. Bois *et al.* A topological data analysis-based method for gait signals with an application to the study of multiple sclerosis. *PLOS ONE*, 17(5):1–23, 2022.

[318] D. J. Pitchforth *et al.* Grey-box models for wave loading prediction. *Mechanical Systems and Signal Processing*, 159, 2021.

[319] E. J. Cross *et al.* Physics-informed machine learning for structural health monitoring. *Structural Health Monitoring Based on Data Science Techniques*, pages 347–367, 2022.

[320] A. Soltani *et al.* Algorithms for walking speed estimation using a lower-back-worn inertial sensor: A cross-validation on speed ranges. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1955–1964, 2021.

[321] A. Vaswani *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[322] P. Schwab and W. Karlen. Phonemd: Learning to diagnose parkinson's disease from smartphone data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1118–1125, 2019.

[323] W.-Y. Cheng *et al.* U-turn speed is a valid and reliable smartphone-based measure of multiple sclerosis-related gait and balance impairment. *Gait & Posture*, 84:120–126, 2021.

[324] K. Scott *et al.* Design and validation of a multi-task, multi-context protocol for real-world gait simulation. *Journal of NeuroEngineering and Rehabilitation*, 19(1):141, 2022.

[325] L. Bull *et al.* Probabilistic active learning: an online framework for structural health monitoring. *Mechanical Systems and Signal Processing*, 134:106294, 2019.

[326] F. Schwenker and E. Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37: 4–14, 2014.

[327] M. Wang *et al.* Active learning through density clustering. *Expert systems with applications*, 85:305–317, 2017.

[328] A. Vaith, B. Taetz, and G. Bleser. Uncertainty based active learning with deep neural networks for inertial gait analysis. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–8, 2020.

[329] H. B. Menz, S. R. Lord, and R. C. Fitzpatrick. Acceleration patterns of the head and pelvis when walking on level and irregular surfaces. *Gait & posture*, 18(1):35–46, 2003.

[330] M. A. Brodie, H. B. Menz, and S. R. Lord. Age-associated changes in head jerk while walking reveal altered dynamic stability in older people. *Experimental brain research*, 232:51–60, 2014.

[331] K. Dylan *et al.* Evaluation of age-related differences in the stride-to-stride fluctuations, regularity and symmetry of gait using a waist-mounted tri-axial accelerometer. *Gait and posture*, 39, 2014.

[332] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[333] R. R. Corbeil and S. R. Searle. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976.