COMPUTATIONALLY EFFICIENT STATISTICAL METHODS IN IOT AND HUMAN GAIT ANALYSIS

COMPUTATIONALLY EFFICIENT STATISTICAL METHODS IN IOT AND HUMAN GAIT ANALYSIS

By MANAN MUKHERJEE, M.Sc.,B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial
Fulfillment of the Requirements for
the Degree Doctor of Philisophy

McMaster University

DOCTOR OF PHILISOPHY (2025)

Hamilton, Ontario, Canada (Mathematics and statistics)

TITLE:	Computationally Efficient Statistical Methods in IoT and Human Gait Analysis
AUTHOR:	Manan Mukherjee M.Sc.,B.Sc.(statistics), Calcutta University, Kolkata, India
SUPERVISOR:	N. Balakrishnan M. Jamal Deen Shiva Kumar
NUMBER OF PAGES:	xx, 211

Abstract

In the era of smart systems and wearable technologies, computational efficiency and interpretability are paramount in developing suitable statistical methodologies for real-world applications. This thesis, titled *Computationally Efficient Statistical Methods in IoT and Human Gait Analysis*, presents a trilogy of studies addressing these needs through developments in outlier detection and human gait assessment.

The first part of this thesis focuses on enhancing anomaly detection in Internet of Things (IoT)-based systems, where outliers such as faults or intrusions can compromise data reliability and Quality of Service. We improve upon the widely used Recursive Principal Component Analysis (R-PCA) method by introducing a data-driven Satterthwaite-based approximation to model the distribution of squared prediction error (SPE) scores more accurately. This refinement corrects the theoretical ambiguities of the Gaussian assumptions in traditional R-PCA and provides a reproducible, real-time outlier detection algorithm with superior performance validated through simulations and graphical plots.

The second part of the thesis explores the use of beta regression models to understand how demographic and gait-specific parameters influence the human Gait Index (GI). By analyzing data from healthy individuals, this study identifies key factors such as walking speed, stride length, knee angle, and stance-to-swing phase ratio as significant contributors to gait variability. Importantly, it also reveals notable interaction effects, including those between age and gait features, which underscore the complexity of gait dynamics. We also develop an unified multivariate Beta regression model by using the Gait Index to improve gait stability assessment and for a better understanding of the variability in gait stability. This methodological advancement provides valuable insights for clinical applications, enabling personalized rehabilitation strategies and more accurate evaluations of gait health.

The third study applies an interpretable machine learning framework using Bayesian Additive Regression Trees (BART) to classify gait patterns into healthy, neurological, and orthopedic categories based on data from over 40,000 footsteps across 230 subjects. The developed approach not only demonstrates high predictive performance (in terms of improved AUC and F1 scores), but also identifies physiologically meaningful features—such as loading phase, walking speed, stride length, and asymmetry in single support time—as key discriminators. Through SHAP and permutation-based analyses, we further establish the interpretability and clinical relevance of the model, offering insight into the underlying mechanics of gait abnormalities.

Together, these studies provide a cohesive body of work that advances the statistical and machine learning methodologies for outlier detection and human gait analysis—balancing computational efficiency with interpretability and real-world applicability in both engineering and biomedical domains.

I would like to dedicate my thesis to my mom, dad and to my supervisor, Prof. N. Balakrishnan.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Narayanaswamy Balakrishnan, who introduced me to the topic of my thesis and provided me priceless and infinite guidance, suggestions, patience and care. I would like to thank him for never giving up on me and for being there during the darkest periods of my academic journey. His feedback allowed me to deepen and refine my research, work, and ideas, and the results presented in my thesis would have been impossible without his supervision. He not only taught me how to conduct research, but also how to become a calm, compassionate, and good human being.

I am most grateful to my co-supervisors, Prof. M. Jamal Deen and Dr. Shiva Kumar, for lending me their expertise and intuition in addressing my scientific and technical challenges. I would especially like to thank Prof. Deen for introducing me to cutting-edge biomedical research problems and for his constant motivation to excel. His continuous feedback and guidance helped me publish two first-author papers in high-impact journals during my doctoral journey.

I would like to thank my doctoral supervisory committee member, Prof. Ben Bolker, for being part of my committee. He generously gave his time to offer valuable feedback that helped improve my work. I am especially grateful to him for teaching me how to tackle real-world applied datasets and for encouraging me not to get lost in the labyrinth of the ‘curse of novelty’.

I would like to thank my parents and my grandparents for their love, support and encouragement during my graduate studies. Without them, this day would not have been possible.

I would like to thank all my collaborators—Dr. Manali Mukherjee, and Prof. Rama Singh—who helped me shape my statistical intuition and problem solving skills.

I wish to sincerely thank my friends Dr. Katherine Davies, Greogory Forkutza, Elorm Sowu, Eman Alamer, Dean Hansen, Cameron Roopnarine, Arka Banerjee, Arghya Dutta, Ramakrishna Jyoti Samanta, Abhiroop Chowdhury, Arijit Nandi, Juan Yin, Xuefeng, Qasim, Subhajit Mishra, Abu Ilius Faisal and all other research students in the Department of Mathematics and Statistics for their kind assistance and moral support.

I acknowledge McMaster Univeristy for providing me with scholarship funding, and I thank the Department and the University for providing various learning, enriching, networking, and professional and personal development experiences.

Table of Contents

List of Figures

List of Tables

Declaration of Academic Achievement

Chapter 2 of this thesis was submitted on 15 November 2023; revised on 23 February 2024 and again on 5 May 2024; accepted on 30 May 2024; and published on 6 June 2024 in *IEEE Internet of Things Journal*, Vol. 11, No. 20, 15 October 2024.

Chapter 3 of this thesis has been accepted for publication in the *IEEE Journal of Biomedical and Health Informatics.* It is scheduled to appear in the upcoming issue: JBHI, Vol. 29, Issue 10, October 2025.

Chapter 1

Introduction

1.1 Outlier Detection in IoT: Part A

In the context of Internet of Things, outliers can be thought of as data values that are considerably different from the rest of the data points, do not correspond to the predicted normal behaviour, or conform well to a defined abnormal behaviour. Depending on their position, outlying observations may or may not have a large effect on the results of the analysis. Sometimes, outliers are extremes of the original data points, which lead to erroneous results. In the IoT literature, these anomalies/outliers can be thought of as noise and error (dealing with fault detection), events (event detection), or malicious attacks (intrusion detection). All of these anomalies distort the statistical analysis of the data points. In wireless sensor networks (WSNs), outliers can be defined as measurements that vary considerably from the typical pattern of the sensed data. These networks consist of sensors or sensor nodes that transfer the data to a base station. These sensors collect data on different measurements of different physical characteristics. The presence of anomalies in these data streams greatly affects the decision-making of the system, leading to service degradation and poor quality of service (QoS).

IoT-based systems deal with high-dimensional data that include many attributes or features. Each sensor collects strings of data measurements on different physical characteristics. A more complex system would consist of more of these sensors, resulting in a data matrix with a large number of features. The analysis is thus complex and would require a reduction of dimension. The Principal Component Analysis (PCA) transforms the data space onto a lower-dimensional subspace (a low-rank intrinsic approximation of the data matrix) that captures most of the variability of the data space. It keeps the most essential features of the original data. The newly transformed axes, i.e., the principal components, are uncorrelated among themselves, which also deals with possible multicollinearity among the features.

Data outliers can dramatically influence the residual values (also known as reconstruction errors, and Squared Prediction Errors (SPE) scores) after extracting the desired PCs. An anomaly is detected by the significant change in the SPE scores. Hence, the calculation of SPE scores necessitates a test statistic, accompanied by an assumption about the distribution of scores and the implementation of a threshold-based scheme. The difficulty of identifying outliers, particularly multivariate outliers, arises from the diverse array of types they can exhibit, which hampers efforts to discern their underlying patterns or causes. However, there are good reasons for looking at the directions defined by either the first few or the last few PCs in order to detect outliers [30].

Some of the most popular anomaly detection schemes, used in the IoT industry, suffer from a theoretical ambiguity when implementing these methods in their domain. The improvisations might be helpful in terms of computational cost and real-time applications, but they may affect the accuracy of the measure. One such popular method is Recursive Principal Component Analysis (R-PCA) based outlier detection and data aggregation scheme [22]. The R-PCA algorithm is based on k-means clustering for outlier detection in IoT systems. The algorithm clusters the data and transmits it to the cluster heads. R-PCA is then used to analyze the data, taking into account the spatial correlation and dynamic changes in the IoT data. The parameters of R-PCA are recursively updated to accommodate these changes. Compared to PCA, this algorithm offers superior performance in terms of both low false alarm rate (FAR) and low power consumption. Ref. [22] used the idea of [13] for defining the SPE score thresholds. R-PCA was an improvement of the work of Chan et al. [6], which proposed a robust recursive fault detection technique. Ref. [6] focused mainly on the sensitivity to monitor system changes and robustness to dramatic data faults. As mentioned in [22], the work of [6] focused on only one single node and was too complex for real-time implementation. R-PCA [22] was able to address this issue.

The work reported in [22] is seminal in terms of real-time outlier detection schemes. It detects and diagnoses anomalies in different sensor nodes and aggregates the data coming from individual sensors. However, the concern remains on the distributional assumptions of the SPE scores, which is the foundational building block for the algorithm. But, R-PCA [22] assumes Gaussianity for the dataset, and in this regard, SPE scores could never be Gaussian as it is a quadratic form. The authors simplified the threshold based on [13], where the SPE scores were transformed into Gaussian distributed scores [13]. In [13], the authors did a type of Box-Cox transformation of the non-Gaussian SPE scores and made it Gaussian. Also, the authors in [22] used two databases, NDBC-TAO and Intel Laboratory, for applying the algorithm. Missing values were assumed to be anomalies or outliers in [22], and they were generated randomly in the database for testing the algorithm. They also considered continuously missing data (occasionally failed communications) in the NDBC-TAO database as outliers. However, outliers might not be missing values. Sometimes, one might consider both outliers and missing values as abnormality in the system. However, missing values can be imputed and considered as an usual observation in the database. Missing value handling is an important problem that data-based practitioners are interested in. One can find a vast literature on missing data handling in statistics and related domains. EM-algorithm [4] and different missing-value-imputation methodologies are some of the most fundamental tools for such an analysis.

In order to address the challenges mentioned above, we have proposed a general outlier model for simulation purposes. The general outlier model enables the practitioner to gather more information on such values that are significantly different from the rest of the data points. The contributions of this Chapter 2 can be summarized as follows:

1. We have identified a major theoretical ambiguity in the distributional assumption of the SPE scores in the R-PCA-based method [22]. Through simulation-based Q-Q plots, we have illustrated why a Gaussian approximation, as proposed in [22], may not be theoretically reasonable. We have used different test statistics for calculating the SPE scores that are very popular in anomaly detection in the IoT industry;

2. In our work, we present an improved approach to approximating the distributional assumptions of SPE scores across various test statistics. We have incorporated two additional techniques, building upon the framework established in [22]. One involves employing Rao's [19] test statistics, while the other is based on Hawkins' [9] test statistics, for computing the SPE scores. These implementations are distinct approaches based on the Gaussianity outlined in [22], which serves as our benchmark for comparison. The proposed approximation-based algorithm maintains the same computational complexity as the R-PCA-based approach [22] while achieving superior accuracy compared to both techniques;

3. We conduct a comprehensive Monte Carlo simulation study with various sample sizes and dimensions to compare the relative performance of the proposed method with the original R-PCA-based method [22]. We have provided a statistically sound framework for generating outlier-contaminated data in simulation-based environments. Each Monte Carlo run emulates the dynamic nature of the R-PCA framework (Recursiveness) by generating different data subspaces. In all cases, our proposed method demonstrates superiority in performance, as evidenced by average $F1$ and MCC scores calculated over all Monte Carlo runs, compared to the method described in [22].

Additionally, we have conducted a brief systematic literature review of existing statistical methodologies widely utilized in the IoT industry for anomaly detection in sensor data. As mentioned earlier, two of the most prominent techniques identified in the literature were employed in our simulation studies to compare the proposed scheme with the original R-PCA-based method [22].

The rest of Chapter 2 is organized as follows. In Section 2.1, we briefly discuss existing literature on PCA-based outlier detection from both statistical and IoT practitioner perspectives. Section 2.2 discusses the preliminaries of PCA and different techniques for calculating the SPE scores. In Section 2.3, we briefly discuss the RPCA-based outlier detection scheme and the corresponding algorithms. In Section 2.4, we propose a Satterthwaite-based approximation of the SPE scores and propose an implementation concerning the RPCA-based framework. In Sections 2.5 and 2.6, we discuss different outlier models for simulation purposes and the selected model for our implementation. We also demonstrate the non-Gaussianity of the SPE scores using Q-Q plots and compare the simulation-based results and performance evaluations based on two assumptions and the two most popular test statistics. Section 2.6 also discusses the complexity analysis and potential limitations of the proposed scheme. Finally, in Section 2.7, we discuss all the contributions of this Chapter and some possibilities for future work.

1.2 Human Gait Analysis: Part B

Human gait, the unique way we move, is like a symphony of coordinated movements that serves as a profound physiological marker reflecting an individual's health, mobility, and functional strength [32], [33], [34]. The significance of gait analysis in understanding and addressing various health conditions has gained prominence, transcending from a biomechanical curiosity to a pivotal aspect of clinical evaluation [35], [36], [37], [38], [39], [40]. Gait analysis holds a unique position as a non-invasive, dynamic assessment tool that offers valuable insights into musculoskeletal function, cardiovascular health, neurological integrity, and overall mobility [41], [42], [43], [44],

[45], [46], [47], [48]. In this paradigm, the Gait Index (GI) emerges as a transformative milestone, taking us a step further in simplifying the complexities of gait into a succinct, quantifiable metric.

This thesis advances statistical and machine learning methodologies for understanding and classifying human gait patterns, with a focus on interpretability and clinical relevance. The inferential part investigates the influence of demographic and gait-specific parameters on the Human Gait Index (GI) through beta regression models. Using data from healthy individuals, it identifies key predictors such as walking speed, stride length, knee angle, and the stance to swing ratio, and highlights important interaction effects, particularly between age and gait features, underscoring the complexity of gait dynamics. An unified beta regression framework is proposed using the Gait Index to enhance the assessment of gait stability and variability. The predictive modeling study introduces an interpretable machine learning approach using Bayesian Additive Regression Trees (BART) to classify gait patterns into healthy, neurological, and orthopedic categories, based on over 40,000 footsteps from 230 subjects. The model demonstrates strong predictive performance while identifying physiologically meaningful features such as the loading phase, asymmetry in single support time, and stride-related variables as key discriminators. SHAP and permutation-based analyses further validate the model's interpretability and clinical utility. Collectively, the work contributes a robust and interpretable toolkit for gait analysis with applications spanning personalized rehabilitation, biomedical research, and wearable sensor technologies.

Here, we introduce the fundamentals of wearable device–based gait analysis, outline current practices and study protocols, describe the datasets used in our studies, and explain the data pre-processing techniques.

1.2.1 Wearable Devices for Gait Analysis

Wearable devices have emerged as transformative tools in gait analysis, offering non-invasive and real-time monitoring of human mobility. These devices are particularly advantageous for tracking joint movements and physiological parameters during daily activities, making them indispensable in fields like rehabilitation, sports medicine, and elderly care. Among wearable technologies, inertial measurement units (IMUs) stand out due to their compact design, precision, and versatility.

1.2.2 Inertial Measurement Units (IMUs) in Gait Analysis

IMUs are widely used in wearable systems for gait analysis as they provide detailed mechanical data about joint motion and orientation. These sensors measure three-dimensional linear acceleration (via accelerometers), angular velocity (via gyroscopes), and magnetic field vectors (via magnetometers). By combining data from these components, IMUs can calculate joint angles, orientation, and other critical gait parameters such as stride length, gait speed, cadence (steps per minute), and minimum foot clearance (MFC) [79] [84].

A typical IMU-based setup involves placing two calibrated IMUs above and below the knee joint to capture precise movement data. The compact size of IMUs and their ability to wirelessly transmit data make them ideal for long-term monitoring without restricting users to laboratory environments. Additionally, a multi-sensor wearable system can be adapted which combines IMUs with additional sensors to provide a comprehensive analysis of knee joint health and mobility. The device integrates:

- **IMU Sensors:** For measuring acceleration, angular velocity, and orientation;
- **Temperature Sensors:** To monitor skin temperature around the knee;
- **Pressure Sensors:** To assess muscle pressure during movement;
- **Galvanic Skin Response (GSR) Sensors**: To measure sweat gland activity related to stress or exertion.

These sensors work together to collect diverse data points that reflect both mechanical and physiological aspects of knee health. The system uses wireless connectivity to transmit data to a smartphone app for storage and processing [84].

1.2.3 Study Protocol for Gait Analysis

Each participant completed a physician-prepared questionnaire designed to gather key physical information, including sex, age, weight, height, leg length, and knee circumference. Participants then performed specific tasks while wearing a multi-sensor system to comprehensively capture knee joint movements. Two IMUs were strategically positioned: one above and one below the knee joint. Additional sensors measuring temperature, pressure, and galvanic skin response (GSR) were placed around the knee area to record physiological parameters such as skin temperature, muscle pressure, and sweat rate.

Participants walked approximately 200 meters on a well-lit, obstacle-free walkway at their preferred walking speed while wearing a knee brace. The IMUs were oriented such that the x, y, and z axes corresponded to the upright (longitudinal), outward (mediolateral), and forward (anteroposterior) directions, respectively. To ensure measurement consistency across all subjects, the knee brace was always positioned in the same location and orientation, with the knee centered and the IMUs placed 14 cm above and below the joint.

Data collection was facilitated through an Android application capable of synchronously retrieving data from multiple CPro modules via Bluetooth. All sensor data were securely stored in an anonymized format (.csv) on a computer for post-processing and further analysis. It is important to note that our analysis primarily focused on spatio-temporal gait mechanical parameters. Physiological parameters were not considered in this study.

1.2.4 Gait Index: Due to Abu Ilius Faisal et al. [49]

For calculating the systemic Gait Index, an initial systematic literature review identified key gait parameters. Six spatiotemporal parameters were identified: walking speed (WS), stride length (SL), gait cycle (GC), cadence (Cad), stance phase (StPh), and swing phase (SwPh)—described in either time or percentage. Additionally, three angular parameters were considered: maximum knee flexion angle ($KA_{\max}$), hip angle ($HA_{\max}$), and ankle angle ($AA_{\max}$). A dataset comprising 120 healthy subjects was then analyzed to compute these parameters ([84], [116]) as well as their demographic factors such as age, gender, and BMI. Statistical analyses were conducted to identify the most statistically significant gait parameters: walking speed – WS, maximum knee flexion angle – $\mathrm{KA}_{\max}$, height normalized stride length – $\mathrm{SL}_{\mathrm{norm(h)}}$, and stance-to-swing phase ratio – StPh/SwPh, forming the basis for the GI as

$$\mathrm{GI} = \frac{WS \times KA_{\max} \times SL_{\mathrm{norm(h)}}}{StPh/SwPh} \tag{1.2.11}$$

While analyzing the final four gait features to formulate the GI, it was observed that stride lengths (SLs) were affected by the heights of the subjects. Therefore, it was necessary to normalize the SL by the height of the subject to minimize the effect of height h, as

$$SL_{\text{norm}}(h) = \frac{SL}{h}. \tag{1.2.12}$$

We also converted the unit of KA_{max} from degrees ($^{\circ}$) to radians (rad) to bring the values of all the parameters into a similar range. In the above mentioned formula, walking speed (WS), maximum knee flexion angle (KA_{max}), and normalized stride length ($SL_{\text{norm}}(h)$) are placed in the numerator since they are positively correlated with gait health—higher values indicate better gait performance. Conversely, the stance-to-swing phase ratio ($StPh/SwPh$) is positioned in the denominator due to its inverse relationship with gait health. As all four parameters contribute equally to explaining variability in the gait dataset—demonstrated by their similar influence on the first two principal components-each variable is assigned equal weight in the equation. This formulation provides a comprehensive index for quantifying an individual's overall gait quality.

In this study, we present a comprehensive inferential modeling framework centered on the Gait Index (GI), designed to quantify gait stability and understand its underlying determinants. Our contributions are three-fold. First, we construct a model that captures the main effects and interactions of gait and demographic factors on the GI, providing a detailed assessment of their influence on gait stability. Second, we incorporate a dispersion sub-model to explain variability in the GI across individuals, offering insight into within-cohort differences in gait profiles. Third, we conduct extensive model diagnostics and interpretation, highlighting key parameters driving both the mean and variability of the GI. Together, these components yield a unified, model-based approach for gait assessment that advances beyond descriptive techniques and holds potential for improving clinical evaluation and personalized care strategies.

1.2.5 Patient Level Gait Prediction: The Study of Human Locomotion with Inertial Measurements Units [142]

The study involved 230 subjects (141 males, 89 females) categorized into three groups: healthy individuals (52 subjects), orthopedic patients (53 subjects), and neurological patients (125 subjects). Participants were recruited from various medical departments in Paris, France, between April 2014 and October 2015. The study was approved by an ethics committee, and informed consent was obtained from all participants. Additionally, more than 40,000 footsteps have been annotated with precise timestamps. This dataset supports clinical research on gait abnormalities and the development of algorithms for quantitative gait analysis.

Study Protocol

Participants performed a fixed sequence of activities:

1. Standing still for 6 seconds;

2. Walking 10 meters at a comfortable pace on a level surface;

3. Turning around without specific instructions on direction;

4. Walking back to the starting point;

5. Standing still for 2 seconds.

Two inertial measurement units (IMUs) were attached to the dorsal face of each foot to record accelerations and angular velocities at a sampling rate of 100 Hz. The IMUs used were XSens™ and Technoconcept® devices.

Data Pre-Processing

In the signal acquisition phase, each IMU captured multivariate time series data comprising accelerations and angular velocities along four axes: X, Y, Z, and V (vertical axis). This resulted in 16-dimensional signals per trial (8 dimensions per foot). The dataset includes over 8.5 hours of gait signals distributed across 1020 trials.

The metadata is enriched with extensive metadata, including: participant demographics (age, gender, height, weight), trial-specific annotations such as timestamps for each footstep event (heel-strike, toe-strike, heel-off, toe-off) and pathological conditions categorized into orthopedic or neurological disorders. Footstep annotations were manually provided by medical experts. These annotations include timestamps marking the start and end of gait cycles, stance phases, and swing phases.

Extracted parameters

A gait cycle is defined as the period between two consecutive heel-strikes of the same foot. It is further divided into:

- Stance Phase: Foot in contact with the ground; includes events such as heel-strike (HS), toe-strike (TS), heel-off (HO), and toe-off (TO);

- Swing Phase: Foot not in contact with the ground.

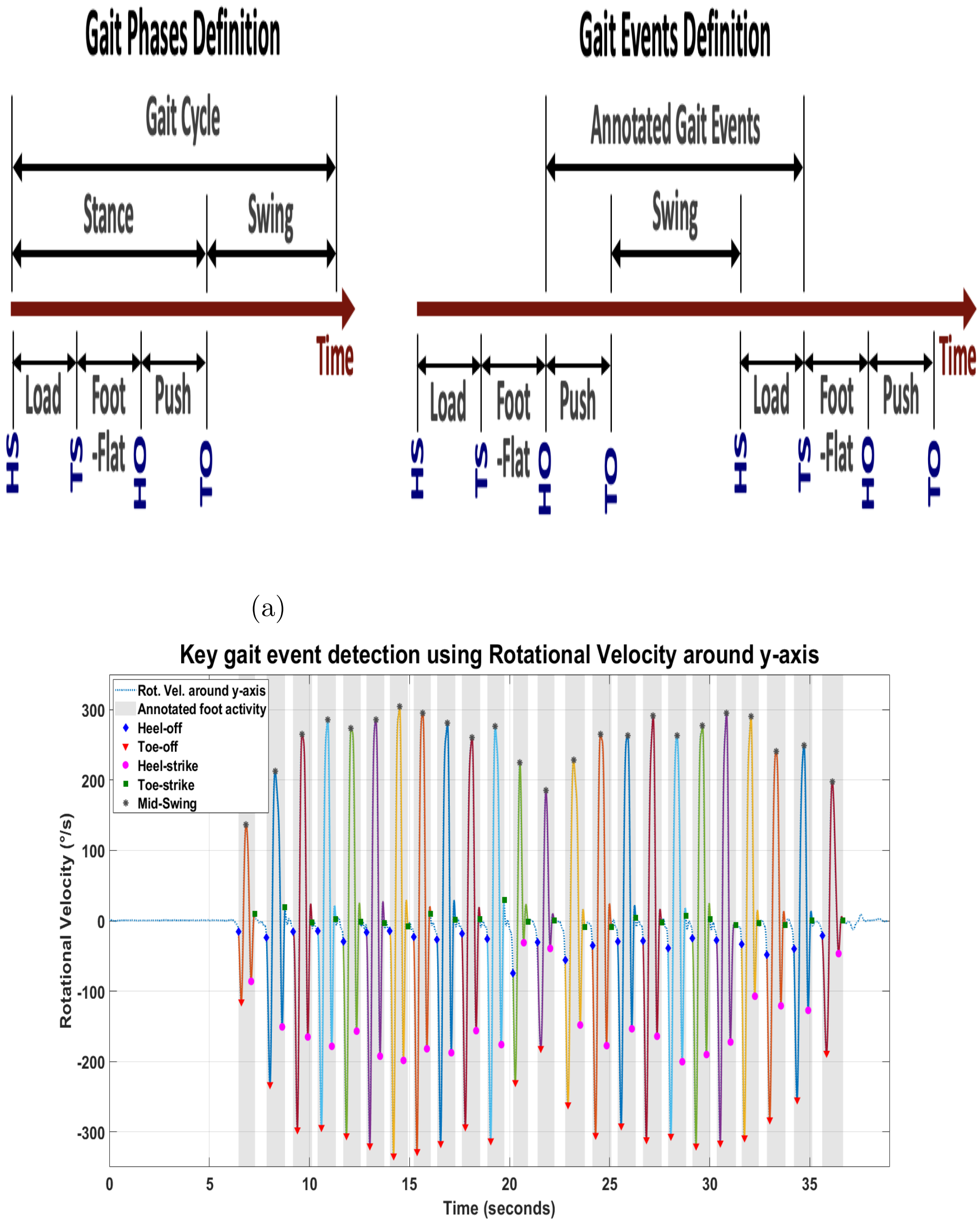
Additionally, features such as acceleration magnitudes, angular velocities, and temporal boundaries of gait events are extracted from the raw time-series data. These features are crucial for analyzing walking patterns and identifying abnormalities associated with specific pathologies.

In Table 1.1, we have presented the extracted gait parameters and demographic parameters used in our study. We handled the missing values where necessary and estimated all the gait phases and calculated the mean, Standard Deviation (SD) and Coefficient of Variation (CV) value, and Asymmetry for the total 82 Features. We have outlined the pre-processing techniques for identifying gait phases and key spatiotemporal gait parameters from IMU-based sensor data, as shown in Fig. 1.1. Pre-processing techniques for identifying gait phases and key gait parameters from IMU-based sensor data: (a) **Gait Phases Definition**: The gait cycle is divided into stance and swing phases. The stance phase consists of Load, Foot-Flat, and Push subphases, while the swing phase follows after Toe-Off (TO); (b) **Gait Events Definition**: Key gait events, including Heel-Strike (HS), Toe-Strike (TS), Heel-Off (HO), and Toe-Off (TO), are illustrated along the gait cycle timeline. These events are used for identifying critical transition points in walking patterns; (c) **Key gait event detection using rotational velocity around the y-axis**: Rotational velocity profiles from the IMU gyroscope data are shown with different colors representing individual gait cycles. Key gait events are marked using different symbols: Heel-Off (HO) (blue diamond), Toe-Off (TO) (red downward triangle), Heel-Strike (HS) (pink circle), Toe-Strike (TS) (green square), and Mid-Swing (gray asterisk). Manually annotated events (HO and TS) and algorithmically detected events (TO and HS) are indicated. The periodic pattern of detected events corresponds to distinct gait cycles, demonstrating the effectiveness of the peak detection approach in identifying key gait transitions. The background shading highlights alternating gait cycles for better visualization.

In this part of the thesis, we present our contributions to predictive modeling using wearable sensor-based gait data. We employed the IOPL dataset, mentioned earlier, which includes over 1,000 time series collected from 230 participants performing structured movement tasks. After manual annotation of more than 40,000 individual footsteps, we extracted temporal, spatial, and spatiotemporal gait features (as given in Table 1.1) and transformed the raw signals into a structured tabular format. To enable clinical interpretation, we categorized participants into healthy and pathological cohorts, and further separated the pathological group into orthopedic and neurological subgroups. This dichotomization allowed us to characterize gait differences across health conditions and set the stage for predictive modeling. We evaluated the predictive performance of Bayesian Additive Regression Trees (BART) in distinguishing between these groups. Various performance metrics and model diagnostics were employed to ensure generalizability and convergence. We benchmarked BART against several traditional machine learning models, including support vector machines, decision tree ensembles, and logistic regression. A key contribution of this work is the interpretability of BART's predictions. We generated cohort-specific feature importance estimates and examined individual predictor effects using Accumulated Local Effects plots. These analyses provided insight into how gait features influence predictions. To strengthen confidence in feature importance, we also conducted a SHAP-based interpretation across competing models. Our findings reveal that BART not only offers competitive predictive performance but also aligns with known biomechanical principles reported in prior studies. Although feature importance alone does not imply causality, the interpretability framework developed here provides clinicians with an enhanced understanding of gait signatures that distinguish patient groups, thereby supporting more informed diagnostic decisions.

Category	**Features**
Gait Features	• Walking speed • Swing and Stance phase (left and right) • Stride time and length • Step time and length (left and right) • Load phase • Push phase • Flat foot phase • Asymmetry: $\text{Asymmetry} = 100 \times \left\| \ln \left(\frac{X_{\text{left}}}{X_{\text{right}}} \right) \right\|$ where X_{left} and X_{right} are the values of a specific gait parameter for the left and right sides, respectively.
Demographic Features	• Age • Gender • BMI (Height, Weight) • Laterality • Pathology group: Neurological, Orthopedic, and Healthy

Table 1.1: Gait and demographic features



(a)

(b)

Figure 1.1: Pre-Processing techniques for identifying gait phases and key gait parameters from IMU-based sensor data.

Chapter 2

An R-simulation-based Improvement of the R-PCA-based Outlier Detection Method

2.1 Introduction

Most of the PCA-based outlier detection schemes depend heavily on SPE scores (residuals or also known as reconstruction errors) calculations and distributional assumption on them. Rao [19] defined the test statistic for calculating the SPE scores based on the last few Principal Components (PCs), which was discussed further by Gnanadesikan and Kettenring [7], as the sum of squares of the values of the last few PCs. Hawkins [9] proposed a revision of this statistic for providing equal weights to the last few PCs which have decreasing order of variance to the end. Gnanadesikan and Kettenring [7] suggested another statistic that focuses on the observations that have an influence over the first few PCs. They target those outliers which inflate the variance of one or more of the original variables. Here, multivariate normality of the selected data is assumed, and the mean and the covariance matrix of the data are known. The distributions of the three statistics mentioned above follow a Gamma distribution. In general, SPE scores are not Gaussian distributed. Jackson and Mudholkar [13] proposed a mapping of SPE that approximately follows a normal distribution. The transformed random variable follows a standard normal distribution (with zero mean and unit variance). There are some other forms of test statistics that have been considered as further improvements of [7], [9] and [19], such as those proposed by Hawkins and Fati [10], and Mertens et al. [17].

All of these test statistics are extensively used in the IoT industry. SPE scores of the form [7], [9] and [19] have been used in [3], [14], [16], and [21], for example. Most of these works have been proposed since 2000. Several works based on the data-subspace reduction using PCA utilize the ideas of [7], [9], and [19] for detecting different kinds of outliers [5], [11], and [15]. Different forms of distributional assumptions have been taken for all such test statistics. Shyu [21] and Lakhina [15] were the first to successfully implement PCA for anomaly detection in IoT. In [21], the use of the F-distribution for D_i^2 is suggested. In [8], the authors have given a framework that combines PCA, Hotelling's T^2, and Q statistics in the context of IoT. In [18], the authors have discussed using the Chi-square and Mixture-Gaussian model in the context of a PCA-based anomaly detection scheme, while in [1], the authors have discussed the method proposed by Jackson and Mudholkar [13] along with other statistical techniques, for detecting anomalies in data. Recently, [31] has employed complex formulations from [13] to calculate the SPE scores and the cut-off value for detecting online fault monitoring in the R-PCA setup. We have observed that some of the most popular anomaly detection schemes, used in the IoT industry, suffer from a theoretical ambiguity when implementing these methods in their domain. The improvisations might be helpful in terms of computational cost and real-time applications, but they may affect the accuracy of the measure.

2.2 PCA Preliminaries

Let X be the $p \times n$ original data matrix. This data matrix consists of p features and n samples. $\hat{X}_l$ is a lower rank approximated form of the original data matrix X, which is of order $l \times n$, where $l < p$. Y is the projection of X in the subspace termed as the score matrix. If P, a $p \times p$ orthogonal matrix, is the transformation basis, then the principal components would be

$$Y_{[p\times n]} = P_{[p\times p]} X_{[p\times n]}. \tag{2.2.1}$$

Now, X should be mean-centered before performing the PCA. Sometimes, the mean-centered data is denoted by $\bar{X}$. The aim of PCA is to derive a transformation basis P that can make the projection of X, i.e., $Y = PX$, linearly uncorrelated and less-dimensional. We choose only the first few, say l, principal components that capture around $80 - 90\%$ of the variability of the data-space. Then, Eq. (2.2.1) becomes

$$Y_{[l\times n]} = P_{[l\times p]} X_{[p\times n]}, \tag{2.2.2}$$

where $P_{[l\times p]}$ (also known as projection matrix, is set to the transposition of the re-ordered and reduced eigenvector matrix) consisting of the first l eigenvectors of the sample covariance matrix $\Sigma_{p\times p}$, as its columns. After selecting the first l Principal

Components (PCs) that explain most of the variability in the original data, we get $\hat{X}_l$. Thus, the error of approximation (also, known as re-construction error) would be

$$\epsilon = (X - \hat{X}_l), \tag{2.2.3}$$

and the squared prediction error (SPE) score would then be

$$SPE(l) = \|X - \hat{X}_l\|_2^2. \tag{2.2.4}$$

This represents the sum of squares of the deviation of $\hat{X}_l$ from X.

Let us now describe each of the three test statistics that we have mentioned in Section 2.1. Rao [19] defined the test statistic SPE_{1i} based on the last few PCs as the sum of squares of the values of the last few PCs, that is,

$$SPE(l)_{1i} = \sum_{k=p-l+1}^{p} z_{ik}^2, \tag{2.2.5}$$

where z_{ik} is the value of the k^{th} PC for the i^{th} observation. The test statistics $SPE(l)_{1i}$, $i = 1, 2, ..., n$, should be approximately independent observations from a gamma distribution if there are no outliers, so that a gamma probability plot with a suitably estimated shape parameter may expose outliers [7].

Hawkins' [9] test statistic is denoted by $SPE(l)_{2i}$. The adjusted z_{ik} is $z_{ik}/\lambda_k^{1/2}$, where $\lambda_k^{1/2}$ is the standard deviation of the k^{th} sample PC. Thus, the new statistic is

$$SPE(l)_{2i} = \sum_{k=p-l+1}^{p} z_{ik}^2/\lambda_k. \tag{2.2.6}$$

Note: When $l = p$, $SPE_{2i}(l)$ simply becomes the Mahalanobis distance D_i^2 between the i^{th} observation and the sample mean, defined as $D_i^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x})$, where S is the sample variance-covariance matrix.

Gnanadesikan and Kettenring [7] suggested another statistic of the form

$$SPE(l)_{3i} = \sum_{k=1}^{l} z_{ik}^2 \lambda_k. \tag{2.2.7}$$

The distributions of the three statistics (Eqs. (2.2.5)-(2.2.7)) follow a gamma distribution. Thus, the plots of $SPE(l)_{1i}$, $SPE(l)_{2i}$, and $SPE(l)_{3i}$ can be used to identify potential outliers. However, in most practical scenarios, the data do not come from a multivariate normal distribution with known mean μ and covariance matrix Σ. Therefore, these distributional results become only approximations.

Jackson and Mudholkar [13] proposed the following mapping of SPE that approximately follows a normal distribution:

$$z = f(SPE) = \theta_1 \frac{[(\frac{SPE}{\theta_1})^{h_0} - \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} - 1]}{\sqrt{2\theta_2 h_0^2}}, \tag{2.2.8}$$

where

$\theta_1 = \sum\limits_{i=k+1}^{p} l_i,$

$\theta_2 = \sum\limits_{i=k+1}^{p} l_i^2,$

$\theta_3 = \sum\limits_{i=k+1}^{p} l_i^3$, and

$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2},$

so that z has a standard normal distribution (with zero mean and unit variance). They [13] also proposed a cutoff value for z, denoted by z_α, with a pre-fixed significance level α, which has a complicated form. Sometimes, the threshold is also denoted by δ_α^2 [1]. We have employed both $SPE(l)_{1i}$ and $SPE(l)_{2i}$ to compare our proposed method with the approach outlined in [22].

2.3 Recursive PCA-based Outlier Detection

The proposed Recursive Principal Component Analysis (R-PCA) algorithm involves two phases: aggregating the data points from the sensor nodes to the cluster heads (which use spatial correlations among the adjacent sensor nodes) and then detecting the outliers based on the aggregated data matrix at the cluster heads. Then, the IoT data center records those outliers along with the aggregated data. Our interest here lies specifically in the outlier detection phase. The outlier detection algorithm has two phases: (1) Initialization Phase and (2) Recursion Phase. In this section, we review the method proposed in [22].

The method in [22] assumes Gaussianity of the data-points. Data matrices are considered real-time, for a particular time instant, say, t. The work-flow consists of two algorithms. **Algorithm 1** detects the outlier in the raw sensor data matrix $\mathbf{X} = \left(X_1^T, \ X_2^T, \ \cdots \ , X_k^T\right)^T$, where k is the number of nodes. Then, $\mathbf{X}$ is normalized to a zero mean and unit variance matrix $\bar{\mathbf{X}}$ through

$$\bar{x}_i(j) = \frac{x_i(j) - \mu_i}{\sigma_i}, \quad j = 1, 2, ..., k, \tag{2.3.1}$$

where μ_i and σ_i are the mean and standard deviation of x_i. So, basically, $\mathbf{X}$ and $\bar{\mathbf{X}}$ are two $n \times k$ matrices. Based on $\bar{\mathbf{X}}$, the sample covariance matrix $\mathbf{\Sigma_X}$ is calculated (see the expression in Step 2 of **Algorithm 1**).

In the next phase, the algorithm calculates the eigenvector $\mathbf{E}$ and eigenvalue (diagonal) matrix Λ from $\mathbf{\Sigma_X}$. Then, it calculates the SPE scores based on the reduced eigenvector space, i.e., based on the number of principal components chosen. Specifically, the SPE scores, for $j = 1, 2, ..., n$, are calculated as

$$SPE(j) = \|\bar{x}(j) - \tilde{E}_l \tilde{E}_l^T \bar{x}(j)\|_2^2, \quad j = 1, 2, ..., n, \tag{2.3.2}$$

where $\mathbf{\bar{X}(j)} = \begin{pmatrix}\bar{X}_1(j), & \bar{X}_2(j), & \cdots & ,\bar{X}_k(j)\end{pmatrix}^T$ and $\tilde{E}_l$ is the reduced eigenvector matrix that corresponds to the first l eigenvalues. Then, in the Recursion phase, the new data points $\mathbf{\bar{X}(t)} = \begin{pmatrix}\bar{X}_1(t), & \bar{X}_2(t), & \cdots & ,\bar{X}_k(t)\end{pmatrix}^T$ are collected at a time t. After normalizing the data, **Algorithm 1** recursively updates the means and standard deviations of each nodes as

$$\mu_i(t) = (1 - \beta)\mu_i(t) + \beta x_i(t), \tag{2.3.3}$$

$$\sigma_i^2(t) = (1 - \beta)\sigma_i^2(t - 1) + \beta(x_i(t) - \mu_i(t))^2, \tag{2.3.4}$$

where $\beta = (1/t)$ is the forgetting factor. Finally, it detects outliers based on the Gaussian distribution assumption on the SPE scores. From Eq. (2.3.2), the mean μ_{SPE} and standard deviation σ_{SPE} of the SPE scores are calculated. A 3σ limit on the $SPE(t)$ scores is introduced, i.e., those SPE scores which fall in the interval $[\boldsymbol{\mu_{SPE} - 3.\sigma_{SPE}, \mu_{SPE} + 3.\sigma_{SPE}}]$ are considered as inliers while those falling outside this interval are considered as outliers.

After **Algorithm 1** determines the outliers, then **Algorithm 2** is used to diagnose the outliers. **Algorithm 2** calculates the SPE scores for each sensor node of

Algorithm 1 R-PCA based Outlier Detection Algorithm: [22]

```
Step 1: Initialization
Step 2: normalize $\mathbf{X} \implies \bar{X} \hookrightarrow \mathcal{N}(0,\ 1)$
Step 3: calculate $\mathbf{E}$ and $\Lambda$ of $\frac{\bar{X}\bar{X}^T}{n-1}$
Step 4: initialize $\mu_{SPE}$ and $\sigma_{SPE}$
Step 5: Recursions
Step 6: update $\mu$, $\sigma$ and normalize $\mathbf{x(t)}$
Step 7: rank $\Lambda$, $\mathbf{E}$ and calculate the number of PCs, $l$
Step 8: reduce $\tilde{E}$ to $\tilde{E}_l$ and $\tilde{\Lambda}$ to $\tilde{\Lambda}_l$
Step 9: calculate $\mathbf{SPE(t)}$
Step 10: if $|SPE(t) - \mu_{SPE}| > \xi_\alpha \sigma_{SPE}$ then
Step 11: outlier detected
Step 12: call Algorithm2
Step 13: else
Step 14: update $\mathbf{E}$ and $\Lambda$
Step 15: update $\mu_{SPE}$ and $\sigma_{SPE}$
Step 16: end if
```

$\bar{\mathbf{X}}(\mathbf{t}) = \Bigg(\bar{X}_1(t),\quad \bar{X}_2(t),\quad \cdots\quad , \bar{X}_k(t)\Bigg)^T$ at time t as

$$SPE_i(t) = \|\bar{x}_i(t) - \tilde{E}_{l,i}\tilde{E}_{l,i}^T\bar{x}_i(t)\|_2^2, \quad i = 1, 2, ..., k, \tag{2.3.5}$$

where k is the number of sensor nodes and $\tilde{E}_{l,i}$ is the eigenvector space for node i. Then, **Algorithm 2** calculates the following ratio:

$$\eta_i = SPE_i(t)/SPE(t) \quad i = 1, 2, ..., k. \tag{2.3.6}$$

Finally, it compares the value of η_i to a prefixed quantity as mentioned in Step 3 of **Algorithm 2**. Based on this comparison, the algorithm finds the outlier and records the time t along with sensor node i. Note that R-PCA [22] uses first-order perturbation theory to calculate the updated $\mathbf{E}(t)$ and $\Lambda(t)$.

Algorithm 2 Outlier Diagnosis Algorithm: [22]

```
Step 1: *for* $i = 1 : k$ **do**
Step 2: calculate $SPE_i(t)$ and $\eta_i$
Step 3: **if** $\eta_i > \xi \sum_{j=1}^{k} \eta_j$ **then**
Step 4: outlier detected
Step 5: record outlier with time $t$ and label $i$
Step 6: **end if**
Step 7: **end for**
```

2.4 Proposed Improvement of the Existing RPCA Algorithm

In the Section 2.1, we have detailed the different forms of test statistics that calculate the SPE score. The one proposed by Rao [19] is the first test statistic that we can directly derive from Eq. (2.2.4) of SPE scores. Eq. (2.2.4) is similar to Eq. (2.3.2). The $\hat{X}_l$ is the reconstructed data using the principal components as indicated by $\tilde{E}_l \tilde{E}_l^T \bar{x}(j)$ in Eq. (**10**), where $\bar{x}(j)$ is simply the original data X (with mean subtracted). Based on Eq. (2.2.5) (due to Rao [19]), we calculate the SPE scores, $SPE(l)_{1i}$. Both Rao [19] and Gnanadesikan Kettenring [9] have stated that the SPE scores follow approximately a gamma distribution if the Gaussian assumptions of the data hold true. We are proposing here a data-driven approximation of the SPE scores, i.e., we are approximating $SPE(l)_{1i}$ by $c(l)\chi^2_{\nu(l)}$, where $c(l)$ is a constant and $\chi^2_{\nu(l)}$ represents a chi-square distribution with $\nu(l)$ degrees of freedom, which is known to be related to a gamma distribution. In the literature on IoT [15][21], $SPE(l)_{1i}$, $SPE(l)_{2i}$ and $SPE(l)_{3i}$ have been approximated by a χ^2_{df} distribution, where *df* is the degrees of freedom, taken to be simply $(p - l)$. We are using Satterthwaite approximation [20] to estimate the constant $c(l)$ and the degrees of freedom $\nu(l)$ based on the SPE scores which are directly linked to the given data points.

In many practical problems, the most efficient estimate of variance available is a linear function of two or more independent mean squares. The exact distribution of such estimates is quite complicated and often intractable. Satterthwaite approximation [20] addresses this issue by approximating the exact distribution by a scaled chi-square distribution with the scaling factor and the degrees of freedom estimated by equating the mean and variance of the SPE scores with those of the scaled chi-square distribution.

After calculating the SPE scores (i.e., $SPE(l)_{1i}$, where $i = 1, 2, ..., n$), we will simply equate the means and variances of $SPE(l)_{1i}$ with the theoretical mean and variance of $c(l)\chi^2_{\nu(l)}$:

$$\overline{SPE(l)_1} = \frac{\sum\limits_{i=1}^{n} SPE(l)_{1i}}{n} = c(l)\nu(l), \tag{2.4.1}$$

$$S^2_{SPE(l)_1} = \frac{\sum\limits_{i=1}^{n}(SPE(l)_{1i} - \overline{SPE(l)_1})^2}{n-1} = 2c(l)^2\nu(l). \tag{2.4.2}$$

Upon solving these two equations for $c(l)$ and $\nu(l)$, we get the estimates of $c(l)$ and $\nu(l)$ as follows:

$$\hat{c}(l) = \frac{S^2_{SPE(l)_1}}{2\overline{SPE(l)_1}}, \tag{2.4.3}$$

$$\hat{\nu}(l) = \frac{(\overline{SPE(l)_1})^2}{2S^2_{SPE(l)_1}}. \tag{2.4.4}$$

Then, we can just set an upper control limit $\alpha = 0.05$ or $\alpha = 0.01$ for $c(l)\chi^2_{\nu(l)}(\alpha)$.

The points that fall beyond this cut-off value should be considered as potential outliers. In the simulation section, we have shown this both pictorially and numerically. The upper control limit determines the trade-off between Type I and Type II errors in hypothesis testing. Lower values of α result in a lower rate of false positives (Type I errors) but may lead to more false negatives (Type II errors), and vice versa. Instead of relying solely on TPR (True Positive Rate) or FPR (False Positive Rate), we have considered using evaluation metrics like the $F1$ and MCC scores. These metrics provide a more comprehensive evaluation of the algorithm's performance across different α values. We have chosen the upper control limit α or the significance level that leads to higher F1 and MCC scores while classifying between outliers and inliers. We have used the two most general significance levels: $\alpha = 0.05$ and $\alpha = 0.01$ and reran the algorithm for the desired Monte Carlo runs for different contamination rates. The choice of $\alpha = 0.05$ yielded to the highest average $F1$ and MCC scores.

The computational cost of our approach is lower than that of Jackson and Mudholkar [13]. The authors of [22] have attempted to optimize the computational cost without using the transformation of [13] (i.e., Eq. (2.2.8)). In the R-PCA paper [22], the authors discussed the computational complexity of using statistics [22] over their oversimplified Gaussian approximation in their paper (see Section V: Performance Evaluation, 4) Threshold: subsection) [22]. The whole point of R-PCA was to use a simplified, less complex threshold-based scheme for the SPE scores that helps to detect outliers more efficiently. The proposed threshold-based scheme takes away the theoretical ambiguity due to the oversimplified Gaussian approximation and keeps the computational complexity the same as that of R-PCA. Our approximation and threshold-based scheme also results in better performance of the algorithm than that of a Gaussian approximated R-PCA [22]. One could implement this onto any of their PCA-based outlier detection methods. We provide a sketch of the revised version of **Algorithm 1** (**Algorithm 3**) with the proposed approximation of the SPE scores. **Algorithm 2** would remain exactly the same as before.

The basic outlier detection framework is simple and can be implemented into any real-time based approach. The work-flow is as follows:

- First, initialize the data matrix $X_{[p\times n]}$ that consists of n observations and p features;
- Next, perform PCA on the data: $Y_{[p\times n]} = P_{[p\times p]}X_{[p\times n]}$, where $Y_{[p\times n]}$ is the principal component matrix;
- Carefully choose the number of PCs, say l, to capture $80-90\%$ of the variability of the data-space, and so we get $Y_{[l\times n]} = P_{[l\times p]}X_{[p\times n]}$;
- Based on the reduced $Y_{[l\times n]}$, reconstruct the data matrix $\hat{X}_l$, which is a lower rank intrinsic approximation of $X_{[p\times n]}$;
- Then, calculate the reconstruction errors: $\epsilon = (X - \hat{X}_l)$;
- Finally, obtain the SPE scores as $SPE(l) = \|X - \hat{X}_l\|_2^2$.

Based on the so-calculated SPE scores, i.e., $SPE(l)$, we perform the Satterthwaite approximation [20] for its distribution. Also, the parameter updates of $\hat{c}(l)_{SPE}$ and $\hat{\nu}(l)_{SPE}$ in **Algorithm 3** would directly follow from the updated forms of μ_{SPE} and σ_{SPE}, as indicated in [22].

Additionally, we compared our scheme with two classical PCA-based benchmark approaches that have been used by practitioners in recent times. These approaches use Hotelling's T^2 and Mahalanobis distances, i.e., D_i^2, and the corresponding thresholds for detecting outliers in the projected subspace [8], [21]. The former classical approach uses Hawkins' [9] test statistics as defined in Eq. (2.2.6) for calculating the SPE scores. Then the distribution of the test statistics is approximated by an F-distribution. Whereas, the latter one uses the Mahalanobis Distances (D_i^2) on the projected space for calculating the SPE scores. In [21], the use of the F-distribution for D_i^2 is suggested. The distribution is approximated by a Chi-Square distribution with the degrees of freedom the same as that of the dimension of the data subspace for large sample sizes. Here, the degrees of freedom depend only on the dimension of the dataspace and do not include any scaling for each real-time feeding of the data. Thus, it does not depend on the dynamic nature of the dataspace. An alternative approximation of D_i^2 is proposed in [21], suggesting that the degrees of freedom should be taken as the number of selected principal components. The connection between Hotelling's T^2 and Mahalanobis distance is explained in Section 2.2. In Section 2.6, we have compared these two classical methods as our benchmarks along with ours to see the performance of our proposed schemes.

2.5 Simulation Models and Performance Evaluation Measures

We want to identify those outliers that influence the correlation structure of the data set. These outliers are more difficult to identify and handle, especially if we have large streams of data-points. We have simulated these outliers from a joint multivariate Gaussian outlier model.

Algorithm 3 Satterthwaite based R-PCA based Outlier Detection Algorithm

```
Step 1: Initialization
Step 2: normalize $\mathbf{X} \implies \bar{X} \hookrightarrow \mathcal{N}(0,\, 1)$
Step 3: calculate $\mathbf{E}$ and $\Lambda$ of $\frac{\bar{X}\bar{X}^T}{n-1}$
Step 4: initialize $\mu_{SPE}$ and $\sigma_{SPE}$
Step 5: initialize $c(\hat{l})_{SPE}$ and $\nu(\hat{l})_{SPE}$.
Step 6: Recursions
Step 7: update $\mu$, $\sigma$ and normalize $\mathbf{x(t)}$
Step 8: rank $\Lambda$, $\mathbf{E}$ and calculate the number of PCs, $l$
Step 9: reduce $\tilde{E}$ to $\tilde{E}_l$ and $\tilde{\Lambda}$ to $\tilde{\Lambda}_l$
Step 10: calculate $\mathbf{SPE(t)}$
Step 11: if $SPE(t) > c(l)\chi^2_{\nu(l)}(\alpha)$ then
Step 12: outlier detected
Step 13: call Algorithm2
Step 14: else
Step 15: update $\mathbf{E}$ and $\Lambda$
Step 16: update $\mu_{SPE}$ and $\sigma_{SPE}$
Step 17: end if
```

For our simulation studies, we define the outlier model as follows:

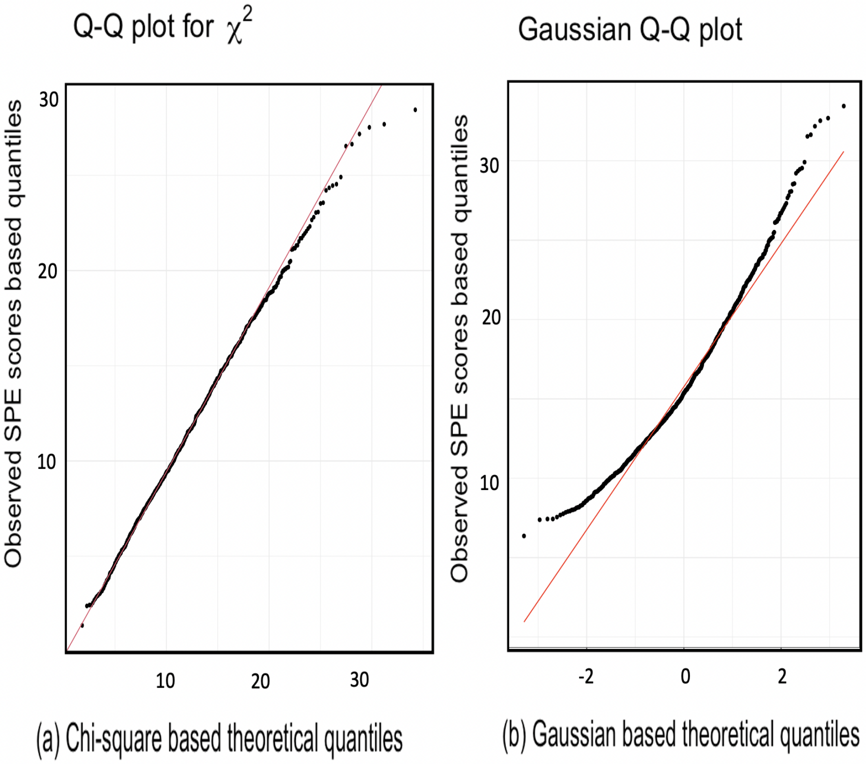
- We sample a core set of $(n - q)$ observations from $N(\mu_p, \Sigma_{p\times p})$ distribution corresponding to regular observations (inliers);
- Then, we add q observations from $N(\mu'_p, \Sigma'_{p\times p})$ as outliers, thus generating a total of n observations with p features;
- Thus, $\delta = q/n$ is the proportion of contamination or outliers, present in the simulated dataset.

There are potentially three kinds of outlier models one could define. First, one could obtain a variance outlier model by setting the covariance matrix $\Sigma'_{p\times p} = \alpha\Sigma_{p\times p}$, where $\alpha > 1$ is a constant and the location parameter $\mu'_p = \mu_p$. Here, the amount of variance in the outlier observations is abnormally large. Second, one could construct a multivariate marginal outlier model by simply changing the location parameter, i.e. $\mu'_p = \mu_p + \Sigma^{1/2}_{p\times p}a$ and the same covariance matrix $\Sigma'_{p\times p} = \Sigma_{p\times p}$. This model simulates outliers as sets of points having potentially high values towards the location parameter. This type of outlier is very easy to detect and requires less complex algorithms. Third, we could obtain a multivariate joint outlier model by setting $\mu'_p = \mu_p + \Sigma^{1/2}_{p\times p}a$ and $\Sigma'_{p\times p} = \Sigma_{p\times p} + \alpha aa^T$, where a is generally a $p \times 1$ vector of constant value and $\alpha > 0$ is a constant. This model simulates outliers as sets of points having potentially high values in some random directions. This type of outlier is common in practical scenarios. Most of the outlier detection schemes are developed for detecting this third kind of outliers (Eq. (2.2.5) and Eq. (2.2.6).

For simulations, we generate data sets of three sample sizes, namely, $n = 1000$, 2000 and 10000. For each of these sample sizes, we contaminate the data sets with outliers at the proportions of $\delta = 5\%$ and $\delta = 10\%$. For the first set of simulations, we generate the data from a $p = 9$ and $p = 12$ variate Gaussian distribution. The choices for the location parameters were taken as $\mu_p = 0$ and $\mu'_p = 8$ for the inlier and outlier models, respectively.

We want to verify our claim that the Satterthwaite approximated scores are essentially an approximated chi-square distribution by creating Q-Q plots based on both chi-square and Gaussian approximations. For the first, we created a Q-Q plot based on the observed SPE sample quantiles and the theoretical chi-square quantiles, and then compared it with the theoretical Gaussian quantiles. In this way, we can observe that the Gaussian assumption on the SPE scores is not reasonable.

Fig 2.1 and Fig 2.2 are based on $n = 1000$ and 2000 samples respectively, generated from Gaussian data, i.e., with $\mu_p = 0_{p\times 1}$ and $\Sigma_{p\times p}$. Then, we generated two Q-Q plots,



(a) Chi-square based theoretical quantiles

(b) Gaussian based theoretical quantiles

Figure 2.1: Q-Q plot based on $n = 1000$ samples without considering any outliers. One can clearly see the departure from the normality in the right.

one based on chi-square distribution and another based on Gaussian. In each set of chi-square-based Q-Q plots, we can see that the sample points mostly lie on a straight line, thereby corroborating our claim. In the Q-Q plots based on Gaussian assumption (on SPE scores), one can observe a clear deviation from the assumption of Gaussianity as the points do not fall on a straight line. One could also observe that at the end of each chi-square-based Q-Q plot, a few points are irregularly scattered and do not coincide with the straight line. The R software uses incomplete gamma functions for approximating the quantiles of chi-square distribution. Hence, these irregularities occur due to approximation errors. This irregularity is more visible when the sample size is large (see Fig 2.2) which is due to the approximation error. Since Chi-Square distribution has heavier tails, it leads to more variability in the observations residing in the tail region which contributes to the tail irregularities that we observe in the chi-square Q-Q plots.

Next, we generated Fig 2.3 and Fig 2.4 based on $n = 1000$ and 2000, from the aforementioned contaminated Gaussian data. Then, we generated two Q-Q plots, one based on chi-square distribution and another based on Gaussian. In each set of chi-square-based Q-Q plots, we can see that the sample points mostly lie on a straight line thereby corroborating our claim. Also, the Satterthwaite approximated chi-square Q-Q plots nicely capture the outliers and shift those values away from the straight lines. On the other hand, in the Q-Q plots based on the Gaussian assumption (on SPE scores), one can observe a clear deviation from the assumption of Gaussianity as fewer points fall on the straight lines. As a result, outliers detected using the Gaussian assumption generate more false positives than the method proposed here. The non-linear structure is visible in the Gaussian Q-Q plot. Approximation errors

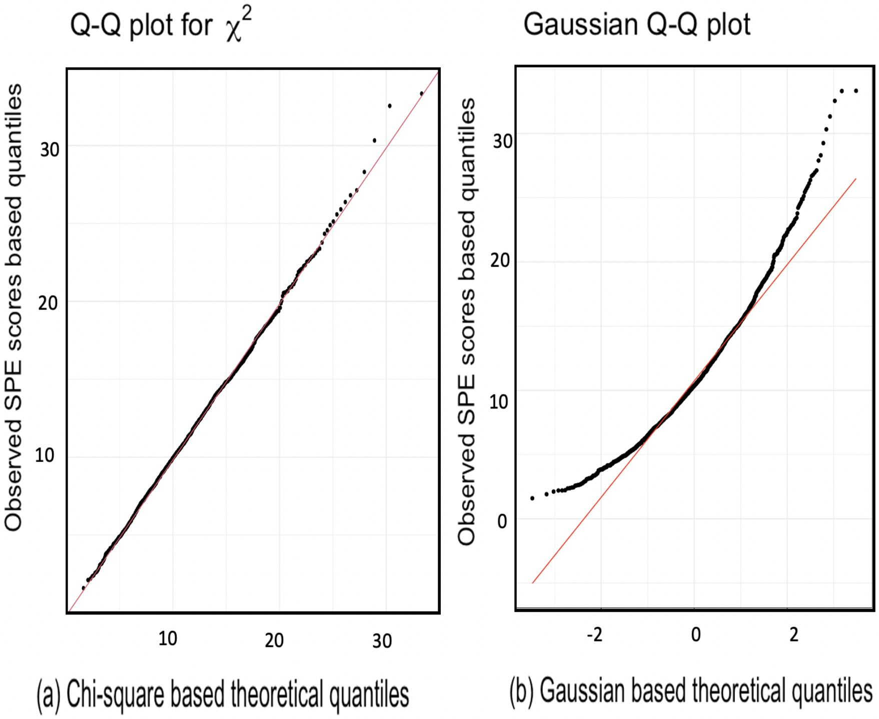


Figure 2.2: Q-Q plot based on $n = 2000$ samples without considering any outliers. One can clearly see the departure from the normality in the right.

and heavy-tailed characteristics still contribute to the irregularities mentioned above.

Our proposed algorithm works as a binary classifier that distinguishes outliers from inliers. However, by selecting specific proportions, we contaminate the data, leading to a class imbalance problem. In IoT, most of the outlier detection literature (e.g. [15], [22]) uses True Positive Rates (TPR) and False Positive Rates (FPR) as measures of accuracy. However, in cases of class-imbalanced data, we already know that one class is more likely to occur than the other. In these situations, accuracy or TPR/FPR cannot be regarded as reliable measures as they overestimate the classifier's performance ability towards the majority class (inliers). In such cases, TPR and FPR can be misleading because a model can achieve a high TPR by marking many inliers as outliers (increasing the FPR). In this regard, $F1$ and Mathews' Correlation Coefficient (MCC) consider both precision and recall, and take into account the class imbalance and facilitate a more accurate assessment of the method's performance.

$F1$ score and Mathews' Correlation Coefficient (MCC) are two evaluation schemes that have gained a considerable amount of attention among the machine learning community. There are many outlier detection works in IoT that use $F1$ scores for performance evaluation. Additionally, MCC scores are somewhat new for the aforementioned purposes. These measures are as follows:

$$F1 = 2 * \frac{Precision \times Recall}{\text{Precision} + Recall} = \frac{2TP}{2TP + FP + FN}, \tag{2.5.1}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{2.5.2}$$

where TP is true positives, FN is false negatives, FP is false positives, and TN is true negatives. The total number of outliers is $TP + FN$ and normal events is $TN + FP$. So, the $F1$ score is essentially the harmonic mean of $Precision$ and $Recall$, where

$$Precision = \frac{TP}{TP + FP}, \tag{2.5.3}$$

$$Recall = \frac{TP}{TP + FN}. \tag{2.5.4}$$

The ranges of $F1$ score and MCC are $[0, 1]$ and $[-1, 1]$, respectively. For the $F1$ score, the minimum is reached for $TP = 0$, that is, when all the positive samples are misclassified and the maximum is achieved for $FN = FP = 0$, i.e., for perfect classification. For MCC, we can interpret it the same way when either of the extremes is achieved. For example,

$MCC = 0$ indicates that the prediction is randomly guessed according to the actual class. $F1$ and MCC scores were chosen as evaluation metrics for our outlier detection algorithm due to their suitability in assessing its overall performance. $F1$ is a metric that strikes a balance between precision and recall, providing a valuable measure for the trade-off between detecting outliers and minimizing false alarms. MCC, on the other hand, takes into account true positives, true negatives, false positives, and false negatives, offering a comprehensive summary of the method's performance. In the context of simulated data, where noise and uncertainty are prevalent, $F1$ and MCC are robust metrics. They consider both true and false classifications, making them less susceptible to fluctuations caused by noise, unlike TPR and FPR. False positives can be particularly costly in simulation-based scenarios, and $F1$ and MCC are well-suited for evaluating the method's ability to reduce false positives while effectively identifying outliers. This balance is essential for maintaining the integrity of simulation results. MCC stands out as a binary-classifier performance evaluation score that rewards models predicting both positive and negative data instances correctly. It condenses the information in the confusion matrix into a single value, facilitating easy performance assessment [24]. Moreover, when comparing different statistics-based outlier detection methods across various samples and contamination levels, using TPR or FPR can be confusing. MCC, being threshold-independent, is a valuable choice in this regard as it is not influenced by the specific decision threshold used to classify instances as outliers or inliers, making it ideal for comparing methods when the optimal threshold is unknown [23].

However, we have compared the FPR when comparing the proposed method with the benchmarks, as mentioned in Section 2.4. The benchmark approaches are well-established outlier detection schemes that perform well in detecting outliers. It's typically expected to have high TPR for such classical benchmarks. However, we are interested in seeing how the FPRs are affected by their non-dynamic nature. So, we have kept the significance level at the same while comparing these methods.

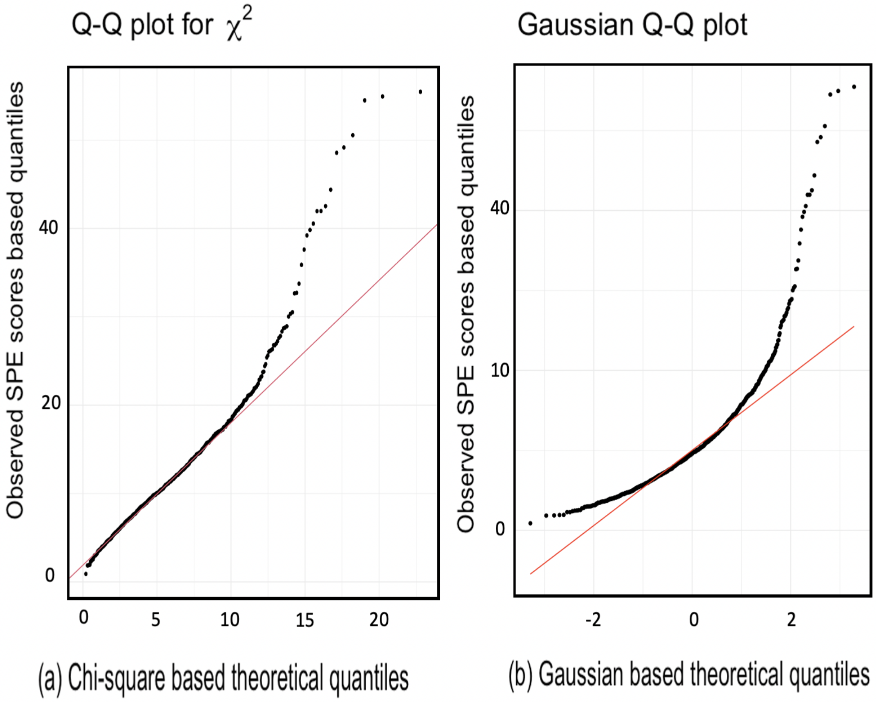
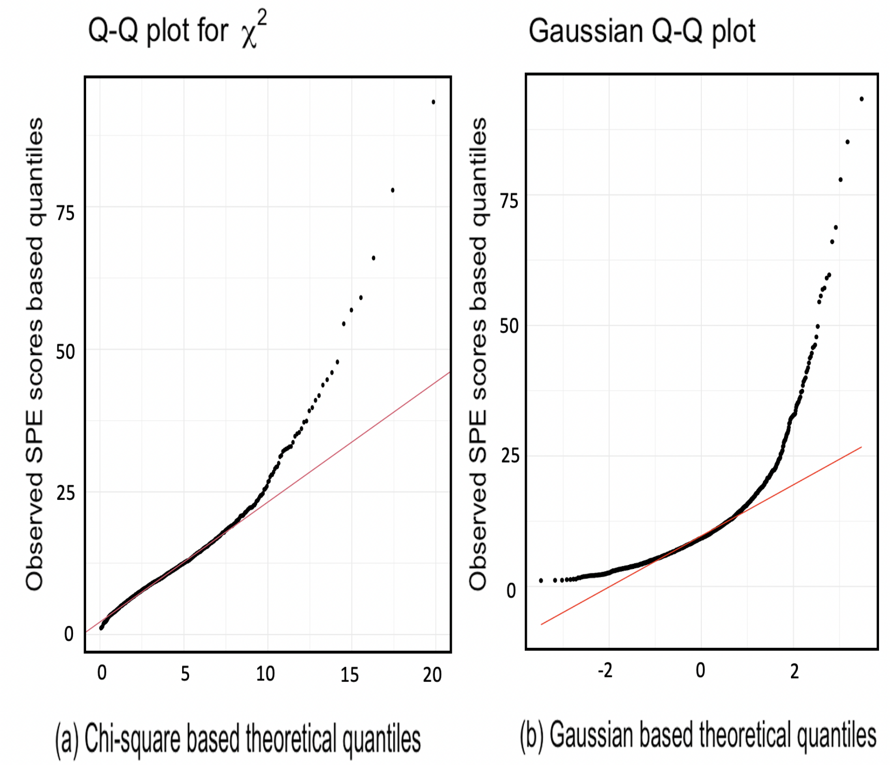


Figure 2.3: Q-Q plot based on outlier model with $n = 1000$ samples. Here, one can clearly see the departure from the normality in the right as well.

We have written a function in R in which the user can plug in the inlier/outlier sample sizes, choices of $\mu_p = 0_{12\times 1}$, $\mu'_p = 8_{12\times 1}$ (or it could be $\mu_p = 0_{9\times 1}$, $\mu'_p = 8_{9\times 1}$, depending on the p itself), and the covariance structures for the two Gaussian set-ups. Based on this, we can easily generate samples of desired sizes, such as $n = 1000, 2000$, and 10000. Next, we run the function 250 Monte Carlo times. For each run, the function generates a p-variate Gaussian data matrix with the desired sample size, say $n = 1000$, from the outlier model, performs PCA on it, and calculates the Squared Prediction Errors (SPE) corresponding to each observation. Finally, it classifies the outliers away from the inliers.

The function also calculates the $F1$ and the MCC scores for each Monte Carlo run.



(a) Chi-square based theoretical quantiles

(b) Gaussian based theoretical quantiles

Figure 2.4: Q-Q plot based on outlier model with $n = 2000$ samples. Here, one can clearly see the departure from the normality in the right as well.

Then, it calculates the average $F1$ scores, MCC scores, and their corresponding standard errors (se) based on these 250 Monte Carlo runs. In this manner, each Monte Carlo run mimics the dynamic environment of the R-PCA scheme by varying the data subspace. In addition, the function also calculates the average TPR and FPR based on the desired number of Monte Carlo runs. We have modified the function in two ways: one uses the usual Rao's statistic for calculating the SPE scores, and the other one uses Hawkins' statistic for the same. In each of these cases, the function approximates the SPE scores based on the Satterthwaite and the Gaussian approaches for detecting outliers. One can easily update the run as per their choices and update the function with new streams of data points with the help of enhanced computational system architecture.

2.6 Performance Evaluation

Table 2.1 and Table 2.2 present simulation-based results from 250 Monte Carlo runs when $p = 9$. We generated samples of sizes $n = 1000, 2000$ and 10000 from the previously described outlier model at contamination rates of $\delta = 5\%$ and 10%. Here, we used both Rao's and Hawkins' test statistics for calculating the SPE scores. Then, we applied both the proposed Satterthwaite and Gaussian approximations used in [22] on these SPE scores to detect outliers. We now compare the performance based on these two approximations. Let us first discuss the results at $\delta = 5\%$ contamination rate. For Rao's test statistic, the average $F1$ and MCC scores for the proposed Satterthwaite-approximated SPE scores ranged between 0.981 and 0.982, indicating excellent performance by the classifier. Both $F1$ and MCC scores showed similar values. In contrast, the $F1$ and MCC scores based on the Gaussian approximation used in [22] SPE scores ranged between 0.833 and 0.839, implying a decline in classifier performance. The standard error for all 250 runs was also higher for the Gaussian approximation-based scheme of [22]. This decline was more prominent when evaluating the model at $\delta = 10\%$ contamination rate, with $F1$ and MCC scores based on Gaussian-approximated SPE scores ranging between 0.521 and 0.595, while the proposed Satterthwaite-approximated SPE scores ranged between 0.793 and 0.863, still indicating superior performance for binary classification. We observed that the Gaussian approximation of [22] consistently yielded lower average $F1$ and MCC scores compared to the proposed Satterthwaite-approximated SPE scores, with values significantly decreasing as the contamination rate rose from $\delta = 5\%$ to $\delta = 10\%$.

Table 2.1: Average *F1* and *MCC* scores for $p = 9$ based on Rao's and Hawkins' test statistics at $\delta = 5\%$ contamination rate.

Samples	**Eval**	**Proposed R-PCA framework**		**R-PCA framework of [22]**	
	Scheme	**Rao-Satter**	**Hawkins-Satter**	**Rao-Gaussian**	**Hawkins-Gaussian**
1000	*F1*	0.982 [se:0.013]	0.957 [se:0.023]	0.834 [se:0.034]	0.686 [se:0.053]
	MCC	0.981 [se:0.013]	0.944 [se:0.023]	0.839 [se:0.034]	0.719 [se:0.053]
2000	*F1*	0.981 [se:0.009]	0.964 [se:0.017]	0.833 [se:0.021]	0.732 [se:0.029]
	MCC	0.980 [se:0.009]	0.931 [se:0.017]	0.840 [se:0.022]	0.749 [se:0.029]
10000	*F1*	0.982 [se:0.004]	0.891 [se:0.009]	0.833 [se:0.010]	0.689 [se:0.016]
	MCC	0.981 [se:0.004]	0.891 [se:0.009]	0.839 [se:0.010]	0.689 [se:0.016]

Table 2.2: Average *F1* and *MCC* scores for $p = 9$ based on Rao's and Hawkins' test statistics at $\delta = 10\%$ contamination rate.

Samples	**Eval**	**Proposed R-PCA framework**		**R-PCA framework of [22]**	
	Scheme	**Rao-Satter**	**Hawkins-Satter**	**Rao-Gaussian**	**Hawkins-Gaussian**
1000	*F1*	0.793 [se:0.025]	0.592 [se:0.041]	0.521 [se:0.027]	0.284 [se:0.037]
	MCC	0.797 [se:0.024]	0.558 [se:0.041]	0.571 [se:0.026]	0.371 [se:0.038]
2000	*F1*	0.793 [se:0.017]	0.594 [se:0.026]	0.518 [se:0.018]	0.316 [se:0.023]
	MCC	0.796 [se:0.017]	0.592 [se:0.026]	0.568 [se:0.018]	0.409 [se:0.023]
10000	*F1*	0.868 [se:0.007]	0.567 [se:0.012]	0.550 [se:0.008]	0.334 [se:0.011]
	MCC	0.863 [se:0.007]	0.595 [se:0.012]	0.590 [se:0.008]	0.428 [se:0.011]

Next, we consider the case of Hawkins' test statistic. We simulated $n = 1000, 2000$ and 10000 samples from the outlier model mentioned earlier at $\delta = 5\%$ and 10% rates of contamination. For $\delta = 5\%$, average $F1$ and MCC scores for the proposed Satterthwaite-approximated SPE scores ranged between 0.891 and 0.957, implying an excellent performance by the classifier. The lowest score was achieved when we used $n = 10000$ sample sizes. In contrast, $F1$ and MCC scores based on the Gaussian-approximated SPE scores of [22] ranged between 0.659 and 0.749. This implies a significant decline in the performance of the classifiers. We note that the decline in performance is quite visible even for $\delta = 5\%$. We observe that the Gaussian approximation used in [22] always yields less average $F1$ and MCC scores compared to the proposed Satterthwaite-approximated SPE scores. The scores significantly decrease when the contamination rate increases from $\delta = 5\%$ to $\delta = 10\%$.

In addition, we observed that the proposed Satterthwaite-approximated SPE scores always yield better $F1$ and MCC (average) scores compared to Gaussian SPE scores of [22] in all 250 Monte Carlo runs. The decline at $\delta = 10\%$ contamination level is due to the non-robust nature of both chi-square and Gaussian distributions. But, the decline in performance is severe for Gaussian SPE scores.

Table 2.3 and Table 2.4 present simulation-based results from 250 Monte Carlo runs when $p = 12$. We generated samples of sizes $n = 1000, 2000$ and 10000 from the previously described outlier model at contamination rates of $\delta = 5\%$ and 10%. Here, we used both Rao's and Hawkins' test statistics for calculating the SPE scores. First, we discuss the results at $\delta = 5\%$ contamination rate. For Rao's test statistic, the average $F1$ and MCC scores for the proposed Satterthwaite-approximated SPE scores ranged between 0.7712 and 0.809, and 0.682 and 0.704, respectively, indicating quite good performance by the classifier. $F1$ scores suggest better performance by the classifier compared to MCC scores. In contrast, the $F1$ and MCC scores based on the Gaussian-approximated SPE scores ranged between 0.593 and 0.601, and 0.607 and 0.613, implying a decline in classifier performance. The standard error for all 250 Monte Carlo runs was also higher for the Gaussian approximation.

This decline was more prominent when evaluating the model at $\delta = 10\%$, with $F1$ and MCC scores based on the Gaussian-approximated SPE scores of [22] ranging between 0.406 and 0.417 and 0.474 and 0.682, respectively, while the proposed Satterthwaite-approximated

Table 2.3: Average *F1* and *MCC* scores for $p = 12$ based on Rao's and Hawkins' test statistics at $\delta = 5\%$ contamination rate.

Samples	**Eval**	**Proposed R-PCA framework**		**R-PCA framework of [22]**	
	Scheme	**Rao-Satter**	**Hawkins-Satter**	**Rao-Gaussian**	**Hawkins-Gaussian**
1000	*F1*	0.809 [se:0.049]	0.800 [se:0.052]	0.601 [se:0.050]	0.550 [se:0.050]
	MCC	0.704 [se:0.049]	0.809 [se:0.052]	0.613 [se:0.050]	0.604 [se:0.050]
2000	*F1*	0.766 [se:0.036]	0.798 [se:0.035]	0.590 [se:0.035]	0.540 [se:0.037]
	MCC	0.684 [se:0.036]	0.803 [se:0.035]	0.604 [se:0.035]	0.601 [se:0.037]
10000	*F1*	0.771 [sd:0.017]	0.796 [se:0.014]	0.593 [sd:0.016]	0.542 [se:0.017]
	MCC	0.682 [se:0.017]	0.802 [se:0.014]	0.607 [se:0.016]	0.598 [sd:0.017]

SPE scores, respectively, ranged between 0.624 and 0.640 and 0.617 and 0.632, still indicating superior performance for binary classification. We noted a consistent trend where the Gaussian approximation used in [22] consistently produced lower average $F1$ and MCC scores when compared to the proposed Satterthwaite-approximated SPE scores. These values exhibited a significant decrease as the contamination rate increased from $\delta = 5\%$ to 10%.

Table 2.4: Average *F1* and *MCC* scores for $p = 12$ based on Rao's and Hawkins' test statistics at $\delta = 10\%$ contamination rate.

Samples	**Eval**	**Proposed R-PCA framework**		**R-PCA framework of [22]**	
	Scheme	**Rao-Satter**	**Hawkins-Satter**	**Rao-Gaussian**	**Hawkins-Gaussian**
1000	*F1*	0.624 [se:0.036]	0.541 [se:0.038]	0.406 [se:0.030]	0.313 [se:0.0302]
	MCC	0.617 [se:0.036]	0.576 [se:0.038]	0.474 [se:0.030]	0.412 [se:0.030]
2000	F1	0.657 [se:0.024]	0.535 [se:0.027]	0.407 [se:0.023]	0.313 [se:0.022]
	MCC	0.636 [se:0.024]	0.574 [se:0.027]	0.473 [se:0.023]	0.415 [se:0.022]
10000	*F1*	0.640 [se:0.010]	0.538 [se:0.011]	0.417 [se:0.009]	0.314 [se:0.010]
	MC	0.632 [se:0.010]	0.570 [se:0.011]	0.483 [se:0.009]	0.412 [se:0.010]

Now, we discuss the results based on Hawkins' test statistic. We present simulation-based results from 250 Monte Carlo runs. As before, we generated samples of sizes $n = 1000, 2000$ and 10000 from the previously described outlier model at contamination rates of $\delta = 5\%$ and 10%, for $p = 12$. For $\delta = 5\%$, the average $F1$ and MCC scores for the proposed Satterthwaite-approximated SPE scores ranged between 0.796 and 0.800, and 0.802 and 0.809, respectively, indicating quite good performance by the classifier. $F1$ scores suggest better performance by the classifier compared to MCC scores. In contrast, the $F1$ and MCC scores based on the Gaussian-approximated SPE scores of [22] ranged between 0.542 and 0.550, and 0.542 and 0.604, implying a decline in classifier performance. The standard error for all 250 Monte Carlo runs was also higher for the Gaussian approximation used in [22]. This decline was more prominent when evaluating the model for $\delta = 10\%$, with $F1$ and MCC scores based on Gaussian-approximated SPE scores of [22] ranging between 0.313 and 0.314 and 0.411 and 0.412, respectively, while the proposed Satterthwaite-approximated SPE scores ranged between 0.538 and 0.541 and 0.570 and 0.576, respectively, still indicating better performance for binary classification.

We also ran the simulation at $\delta = 15\%$ contamination rate for $p = 9$ and 12. For $p = 9$, $F1$ and MCC scores based on the Gaussian-approximated SPE scores of [22] ranged between 0.33 and 0.419, indicating poor classifier performance. However, $F1$ and MCC scores for the proposed Satterthwaite-approximated SPE scores still ranged between 0.646 and 0.656, indicating better binary classifier performance. Similarly, for $p = 12$, $F1$ and MCC ranged between 0.205 and 0.206 and 0.312 and 0.313, indicating poor classifier performance. However, $F1$ and MCC scores for the proposed Satterthwaite-approximated SPE scores still ranged between 0.409 and 0.416, and 0.444 and 0.448, indicating better binary classifier performance. It was consistently observed that the Gaussian approximation used in [22] yielded lower average $F1$ and MCC scores compared to the proposed Satterthwaite-approximated SPE scores. These scores demonstrated a significant decrease as the contamination rate increased from $\delta = 5\%$ to $\delta = 15\%$. However, the proposed Satterthwaite-approximated SPE scores always yielded better $F1$ and MCC (average) scores compared to the Gaussian SPE scores of [22] for all 250 Monte Carlo Runs. The decline at $\delta = 15\%$ contamination level is due to the non-robust nature of both chi-square and Gaussian distributions. But, the decline in performance is more severe for the Gaussian SPE scores.

We also observed that the performance of the algorithm decreases as we increase the

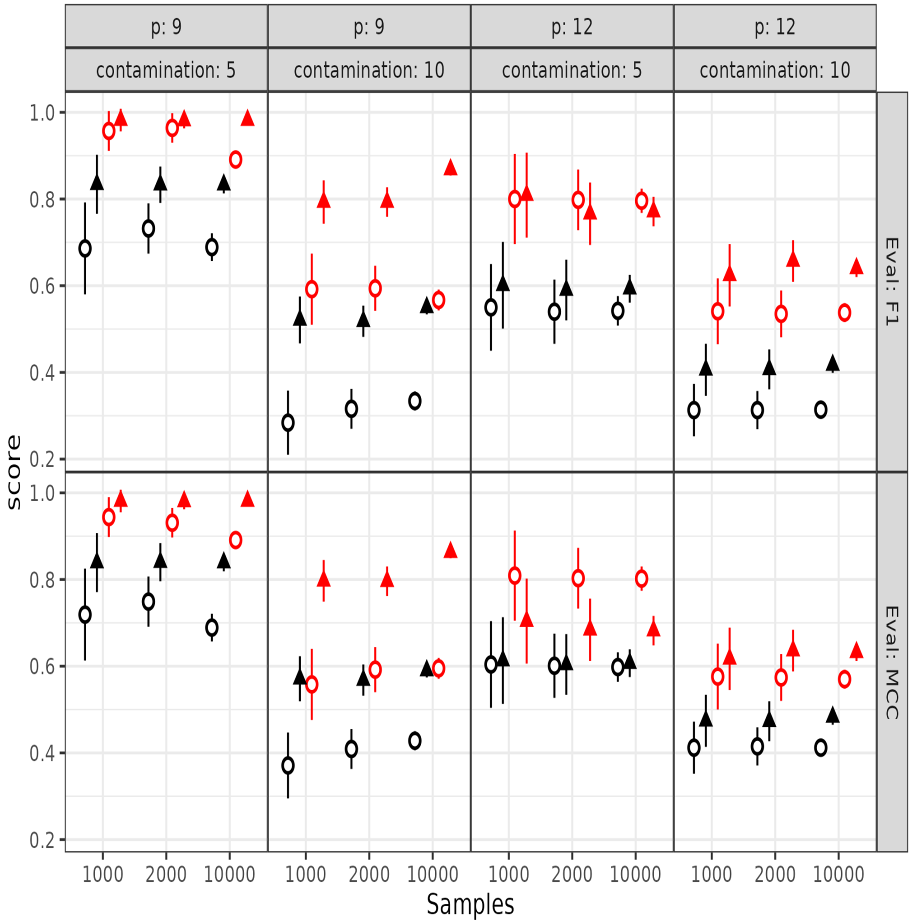


Figure 2.5: Comparisons on detection accuracy between the proposed method and the method from [22] across various sample sizes, dimensions, and contamination rates. The triangles and circles are Rao's and Hawkins' test statistics, respectively. Red triangles and red circles refer to the Satterthwaite-based improvement, and black triangles and black circles refer to the Gaussian method [22].

sample sizes of the data. We have used panel plots in Fig 2.5 for visually summarizing the results. Each panel of Fig 2.5 provides a visual representation of the performance based on the selected performance evaluation score. Fig 2.5 illustrates a decline in performance for both methods as the sample size and dimension increase, as evident from the trend observed in both $F1$ and MCC scores. This is because of the non-robustness of the distribution. If the contamination rate significantly increases, then the $F1$ scores might fall below the $0.45-0.5$ range. Also, higher $F1$ scores indicate a higher number of True positives i.e., higher True Positive Rate (TPR) and lower False Positive Rate (FPR). So, a decline in average $F1$ score reflects the non-robustness of the algorithm for both cases. This is due to the higher number of False Positives that the scheme is incorrectly classifying as outliers. High MCC scores also correspond to the fact that the classifier is able to detect the majority of the positive and negative classes correctly. That leads to the same conclusions as those based on a high $F1$ score. Additionally, Fig 2.5 illustrates the superior performance of Rao's statistic-based approximation compared to implementations based on the Hawkins' framework. So, a robustification based on Rao's test statistic should yield better performance.

Table 2.5: Comparison of the Average MCC and FPR scores among the proposed method and the benchmarks at $\delta = 5\%$ contamination rate for various sample sizes.

Samples	**Eval**	**Proposed Satter**	**Hotelling's T-Square**	**Mahalanobis Distance**
1000	MCC	0.981	0.918	0.909
	FPR	0.00054	0.0075	0.0089
2000	MCC	0.980	0.823	0.912
	FPR	0.00066	0.0230	0.0093
10000	MCC	0.981	0.908	0.909
	FPR	0.00046	0.009	0.0087

Additionally, we compared the proposed method with some classical benchmarks in Table 2.5 to better understand the performance. We have considered three sample sizes of $n = 1000, 2000$ and 10000 and $\delta = 5\%$ contamination rate for our Monte Carlo simulation. We can observe that the proposed method is performing well with the Benchmark methods.

All of the methods yield high MCC scores which correspond to the fact that the classifier is able to detect the majority of the positive and negative classes correctly. Mahanalobis distance-based approaches, i.e., those using D_i^2, are more robust and thus perform better than the Hotelling's T-squared approach. However, the proposed method is also performing well compared to the mentioned benchmark. Additionally, we have compared the False Positive Rate (FPR) of the proposed method with that of the benchmark for understanding the differences in MCC scores. The proposed method has generated fewer false positives compared to the benchmarks for all sample sizes, supporting the assertion in Section 2.4.

2.6.1 Complexity Analysis

In this section, we provide a summary of the computational complexity analysis for the three outlier detection algorithms applied to IoT-based systems using a data matrix $X_{[p\times n]}$. The algorithms under consideration are the original R-PCA-based Outlier Detection Algorithm (**Algorithm 1**), the Outlier Diagnosis Algorithm (**Algorithm 2**), and the proposed Satterthwaite-based R-PCA Outlier Detection Algorithm (**Algorithm 3**).

The original R-PCA algorithm (**Algorithm 1**) exhibits a computational complexity of approximately $O(p^2t)$, where t represents the number of cases and p denotes the number of features (columns) in the data matrix $X_{[p\times n]}$. The complexity arises from matrix operations, including eigenvalue decomposition. The algorithm's performance is directly influenced by the number of cases, making it more computationally demanding as t increases.

The Outlier Diagnosis Algorithm (**Algorithm 2**), designed for IoT-based systems, demonstrates a lower computational complexity compared to the original R-PCA algorithm. Its complexity is mainly $O(pt)$, involving iterations through cases and calculating anomalies for each case. This algorithm focuses on the number of cases as a primary driver of computational load.

The proposed Satterthwaite-based R-PCA algorithm (**Algorithm 3**) has a computational complexity of approximately $O(p^2t)$, similar to the original R-PCA algorithm.

2.6.2 Potential Limitations of the Satterthwaite-based R-PCA

While the points mentioned above highlight the advantages and benefits of the proposed method, it is also essential to point out some potential issues. The proposed method exhibits sensitivity to model assumptions. Its performance can be limited when the underlying data distribution significantly deviates from the algorithm's assumptions, particularly in the case of complex, non-Gaussian distributions. Additionally, the computational complexity of the algorithm increases with the dimensionality of the data. High-dimensional data can render the algorithm computationally expensive and potentially less accurate [25]. To alleviate this issue, one might look into outlier-robust [2] [12] tensor principal component analysis (OR-TPCA) method for simultaneous low-rank tensor recovery and outlier detection [26].

Moreover, the algorithm may require parameter tuning, such as setting the significance level (α) and the degrees of freedom. Choosing appropriate values for these parameters can be a challenging task and may influence the algorithm's ability to detect outliers. This method, based on the Satterthwaite approach, is more suitable for detecting univariate or multivariate outliers and may not perform as effectively in identifying contextual anomalies or more complex patterns often encountered in the IoT data, making outlier-type a potential concern.

Furthermore, IoT-based systems often handle data streams that evolve over time [28]. The algorithm might struggle to adapt to changes in data distribution and could necessitate continuous monitoring and re-calibration, leading to a limitation in its adaptability [27] [29]. These considerations underscore the need for a thorough assessment of the method's suitability for specific applications.

Chapter 3

Inferential Model for Understanding the Effects of Demographic and Gait Factors and Their Interactions on the Human Gait Index

3.1 Introduction

The Gait Index, developed through a systematic process outlined in [49], stands as an unique contribution to the field of gait analysis. Unlike all other gait indices developed in the last three decades [50], [51], [52], [53], [54], [55], [56], which either rely on complex statistical model, focus on specific pathology conditions or subject groups, or incorporate numerous parameters that add complexity, this index offers a simple yet comprehensive approach without

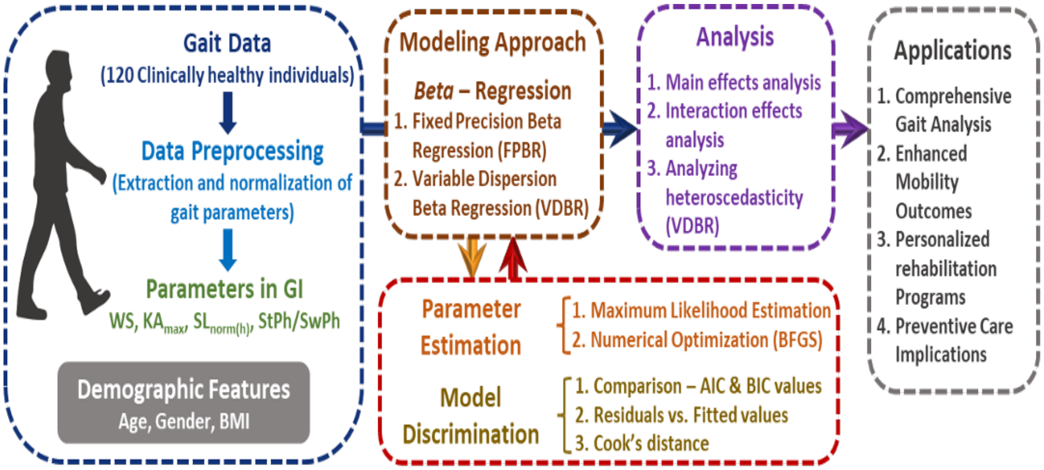

Gait Data
(120 Clinically healthy individuals)
Data Preprocessing
(Extraction and normalization of gait parameters)
Parameters in GI
WS, KA$_{max}$, SL$_{norm(h)}$, StPh/SwPh
Demographic Features
Age, Gender, BMI
Modeling Approach
Beta – Regression
1. Fixed Precision Beta Regression (FPBR)
2. Variable Dispersion Beta Regression (VDBR)
Analysis
1. Main effects analysis
2. Interaction effects analysis
3. Analyzing heteroscedasticity (VDBR)
Applications
1. Comprehensive Gait Analysis
2. Enhanced Mobility Outcomes
3. Personalized rehabilitation Programs
4. Preventive Care Implications
Parameter Estimation
1. Maximum Likelihood Estimation
2. Numerical Optimization (BFGS)
Model Discrimination
1. Comparison – AIC & BIC values
2. Residuals vs. Fitted values
3. Cook's distance

Figure 3.1: The overall flow diagram of the development of our inferential model to assess the effects of demographic and gait factors and their interactions on the human gait index.

compromising clinical relevance. The four key parameters—knee angle (KA), stride length (SL), walking speed (WS), and stance-to-swing phase ratio (StPh/SwPh)—were carefully selected based on a comprehensive literature review to ensure their clinical significance, ease of measurement, and ability to provide meaningful insights into gait dynamics [57], [58], [59], [60], [61], [62], [63], [64]. Many previous indices not only lack generalizability due to their complexity, but also face challenges in accurately estimating multiple parameters using wearable systems, making them less practical for routine and long-term use. In contrast, the GI was specifically developed for the preliminary assessment of gait health, ensuring applicability across diverse populations by normalizing parameters, such as stride length (SL) for height, to account for demographic variations and offering an intuitive framework for interpreting gait quality. The index's development involved rigorous validation using machine learning techniques, demonstrating its reliability and effectiveness. By requiring only a few essential and easily measurable parameters, this innovative GI is highly accessible for routine use in both clinical and research settings. It serves as a universal indicator of gait health, effectively addressing the practical challenges of previously developed indices, while bridging the intricate nature of gait with the practical needs of clinicians to provide a streamlined approach to gait assessment.

To comprehensively explore the GI as a marker of gait stability and its determinants, and to address its bounded nature and inherent variability within the study cohort from a modeling perspective, we evaluated various inferential models and ultimately employed Beta distribution-based regression models. Unlike the common approach in gait analysis, which involves using multiple regression models based on various spatio-temporal gait parameters [65], [66], [67], our study focused on developing a single model to infer overall gait stability while capturing the intricate relationships between key spatio-temporal gait markers included in the GI and demographic factors. The GI, serving as a comprehensive measure of both gait stability and overall gait health, provided a novel framework for this analysis, enabling a singular model-based investigation of the determinants of gait health. This approach not only enhances the understanding of gait dynamics, but also offers potential clinical applications by identifying key predictors that can be targeted for interventions aimed at improving mobility outcomes, even in clinical populations with gait abnormalities.

In this chapter, we have developed an inferential model to provide an in-depth examination of the GI and its key influencing factors (Fig 3.1). In the methodologies section, we outline the model frameworks, parameter estimation methods, and the rationale behind model selection. This framework analyzes the effects of gait and demographic parameters, their interactions on gait stability via the Gait Index, and identifies parameters driving variability in the Gait Index through the dispersion sub-model, reflecting gait stability variability within the cohort. The results section presents these findings, along with model diagnostics, emphasizing the insights gained into the diverse determinants of gait profiles. By integrating the GI as a central component, our study advances beyond traditional approaches, offering a singular model-based perspective on gait assessment. The discussion section synthesizes these results, underscoring their practical implications for clinicians and researchers. Finally, a succinct conclusion is presented summarizing the key contributions of our study and their potential to enhance clinical practice and patient care.

3.2 Literature Review

Nowadays, many models designed to comprehend gait patterns fall under the category of machine learning (ML) [68], [69], [70]. These models prioritize prediction by employing versatile learning algorithms such as Neural Networks (NN), Random Forest (RF), and Support Vector Machines (SVM) to uncover gait patterns within complex and extensive datasets [71]. Prediction serves the purpose of identifying optimal courses of action without necessitating an in-depth understanding of the underlying mechanisms. Consequently, ML models operate with minimal assumptions about the systems generating the data [67], [72]. However, despite yielding compelling prediction results, the absence of an explicit model can render ML solutions challenging to correlate directly with established biological knowledge. On the contrary, statistical inferential models such as linear regression, generalized linear models (GLM), and mixed models have traditionally emphasized inference [73], [74]. This is accomplished by developing and fitting a probability model specific to the project, which can be either Gaussian or non-Gaussian [75], [76]. These models enable us to quantify the confidence level that a detected relationship represents a "significant" effect unlikely to be due to random variation. Additionally, with sufficient data, we can explicitly test assumptions (such as equal variance) and adjust the model as necessary. With this motivation, our interest lies in developing an inferential model capable of understanding the intricate relationships among gait parameters and demographic factors to comprehend gait patterns. Our exploration of modeling the gait index (GI) delves deeper as we examine demographic effects and their influence on the GI. Various demographic factors significantly influence gait patterns [36], [38], [77], [78], [79]. For instance, aging contributes to changes in muscle mass, bone density, and motor control, impacting walking ability [38], [80], [81], [82], [83]. Male and female anatomical variations in the pelvis and thigh, as well as gender-specific muscle activity during walking, lead to distinct gait characteristics [84], [85]. Additionally, BMI significantly influences gait mechanics [86]. Abnormal BMI (underweight, overweight, or obese) can impact gait due to joint stress [87].

Beyond demographic effects, our inquiry extends to the intricate connections between certain aspects of how we walk such as specific gait features and demographic factors including age, gender, or BMI. Our focus is not only on evaluating these factors independently, but also on investigating their collective impact, and how they converge to shape the GI and its variations. Understanding these relationships is essential for unraveling the complexity of gait patterns and their underlying determinants. By recognizing the multifaceted nature of gait and its variation, clinicians can tailor personalized interventions to address specific contributing factors, optimizing gait outcomes in diverse patient populations. The current trend involves using multiple linear regression models and regression-based normalization methods in gait research. For instance, multiple linear regression models are applied to various gait parameters in [65], while 32 linear regression models are utilized without considering interaction effects in [66]. In our study, however, the challenge inherent in modeling the GI lies in its bounded scores, constrained within the [0, 1] range. This constraint poses a significant statistical challenge, urging us to adopt approaches that can effectively navigate the complexities within this confined space. While standard linear regression models are appreciated for their simplicity, interpretability, and ease of use, they fall short when dealing with bounded outcomes like the GI, often predicting values beyond the permissible range [88], [89]. In response to this challenge, recent advancements in statistical modeling have introduced specialized techniques designed for bounded outcomes [90], [91], [92], [93]. Among these, the logit transformation of the bounded response [90], [91] has become a common practice, also known as additive log-ratio transformation [94]. However, this method has its constraints, particularly in handling heteroscedasticity [95]. Barclay et al. [92] applied censored normal or Tobit models to handle proportional data, addressing some of the limitations associated with the logit transformation. Despite these efforts, the critique of using Tobit models for proportional data [96] indicates that not all aspects of the problem have been adequately addressed. In this landscape, the Beta regression model emerges as a methodologically robust solution. Its inherent flexibility and capacity to handle bounded scores make it well-suited for modeling the GI and capturing the intricate dynamics of gait [93], [97], [98], [99], [100], [101], [102], [103]. Johnson et al. (1995) [93] have described over a dozen examples from various physical sciences, showcasing the beta distribution as a better fitting model for proportional data as compared to alternative models. Hviid and Villadsen [97] extended this applicability to economics, while Mittelhammer [98] illustrated the beta distribution's efficacy using the proportion of heating oil in a tank measured at different times. Beta regression is closely linked to an expanded framework of generalized linear models (GLMs) outlined in Chapter 10 of [99], encompassing the joint modeling of means and dispersions. The initial instances of beta regression originated from studies within organizational economics and public management [97], [100]. Brehm and Gates [100] examined police compliance with supervision, employing the standard parameterization of the beta distribution with two shape parameters. However, this standard parameterization introduced complexities in formulating regression models and posed challenges in interpretation. Paolino [101] adopted the mean-dispersion parameterization, offering a more straightforward interpretation. Buckley [102] developed a Bayesian approach based on Paolino's model, employing Markov-chain Monte-Carlo (MCMC) techniques for estimation, while Ferrari and Cribari-Neto [103] derived an independent framework for beta regression model using Fisher scoring, which garnered significant popularity in this literature. However, they did not explicitly tackle model dispersion, but treated it as a nuisance parameter.

The issue of modeling variances has received significant attention in statistical literature, particularly within the field of econometrics [104]. As demonstrated by Cook and Weisberg [105] and Atkinson [91], traditional methodologies predominantly relied on graphical techniques to discern heteroscedasticity under normal errors. In the context of gait research, studies such as [65] acknowledged the presence of dispersion in the data, but did not integrate it into their modeling frameworks. Similarly, in [66], researchers identified heteroscedasticity in residuals for cadence in children, recognizing the failure to account for it as a limitation of their models. While their proposed regression framework aimed to address variations in spatiotemporal gait variables, these models lacked parameters to evaluate the effects of dispersion or to establish causal relationships between independent variables and the observed variability. This omission highlights a critical gap in capturing the complexities of heteroscedasticity within gait data analysis. Smyth [106] introduced a method for modeling the dispersion parameter within certain generalized linear models, expanding beyond normality. To address heteroscedasticity, Cysneiros et al. [107] explored linear models with symmetric errors, delving into diagnostic considerations within this framework. A notable contribution by Smithson and Verkuilen [108] introduced a flexible maximum-likelihood-based regression model, considering both location and dispersion using distinct sets of predictors. In this innovative approach, logit serves as the link function for the location sub-model, similar to logistic regression, while the dispersion sub-model follows a log-linear approach. Originally designed for capturing heteroscedasticity within psychological data, this method proved versatility in diverse applications. In addition, Simas et al. [109] proposed a general class of predictors that captured both linearity and non-linearity within the data. Employing linear predictors for both parameters, this approach offers a streamlined solution, simplifying computational complexities when applied in practical applications. Hence, in our study, we leveraged advanced Beta regression models to proficiently model and interpret the GI within its constrained range as well as perform a comparative analysis of their results to determine the best fit for our gait analysis.

3.3 Methodologies

The development of GI using clinically relevant parameters aims to simplify human gait assessment [49]. In this study, we employ a Beta regression model to explore the relationship between demographic and gait-related predictors and GI, offering valuable insights for clinicians. The methodologies detail the application of Beta distribution and re-parametrization techniques, the construction of regression models, choices of link functions, likelihood estimation methods, and model discrimination techniques using criteria such as Akaike's Information Criterion (AIC) [110] and Bayesian Information Criterion (BIC). This approach provides a comprehensive framework for GI-based gait assessment efficiently and accurately.

3.3.1 Beta Distribution and Re-Parametrization

The Beta distribution is renowned for its versatility in modeling data confined within specific intervals on the real line [93] [111]. Its application is particularly intriguing when dealing with data within the standard unit interval $(0, 1)$, where it can represent rates or proportions.

A random variable Y follows a Beta distribution with parameters p and q, both greater than 0, and is denoted by $\mathcal{B}(p, q)$. This distribution is characterized by its probability density function, with respect to the Lebesgue measure, given as follows:

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1, \tag{3.3.1}$$

where $\Gamma(\cdot)$ is the complete gamma function. The mean and variance of Y are, respectively, given by:

$$\mathbb{E}(Y) = \frac{p}{p+q}, \tag{3.3.2}$$

$$\mathrm{Var}(Y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{3.3.3}$$

To obtain a regression structure for the mean of the response and the precision (or dispersion) parameter [112], Ferrari and Cribari-Neto proposed a new parameterization [103]. Under this parameterization, the expectation and variance of the response can be written as

$$\mathbb{E}(Y) = \mu, \tag{3.3.4}$$

$$\mathrm{Var}(Y) = \frac{\mathrm{Var}(\mu)}{1+\phi}, \tag{3.3.5}$$

where $\mu = \frac{p}{p+q}$, $\phi = p+q$, $p = \mu\phi$, and $q = \phi(1-\mu)$. In this case, $\mathrm{Var}(\mu) = \mu(1-\mu)$ denotes a "variance function". Under this parameterization, we can rewrite the distribution of Y as $\mathcal{B}(\mu, \phi)$. Here, μ is the mean of the response variable, while ϕ can be interpreted as a precision parameter, indicating that for a fixed μ, higher values of ϕ correspond to smaller variances of Y. The density of Y then can be rewritten as

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))} y^{\mu\phi-1}(1-y)^{\phi(1-\mu)-1}, \quad 0 < y < 1 \tag{3.3.6}$$

where $0 < \mu < 1$ and $\phi > 0$, since $p, q > 0$. The log-density of the newly parameterized Y

is given by:

$$\log f(y;\mu,\phi)=\log\Gamma(\phi)-\log\Gamma(\mu\phi)-\log\Gamma(\phi(1-\mu))+(\mu\phi-1)\log y+(\phi(1-\mu)-1)\log(1-y). \tag{3.3.7}$$

3.3.2 Construction of the Regression Model

Considering a random sample, $\mathbf{y}=(y_1,\ldots,y_n)^T$, where $y_i\sim B(\mu_i,\phi_i)$ for $i=1,\ldots,n$, we can define the following functional relations for the mean and precision parameters of the response variable y_i

$$g_1(\mu_i)=\eta_{1i}=f_1(\mathbf{x}^T,\boldsymbol{\beta}) \quad \text{and} \quad g_2(\phi_i)=\eta_{2i}=f_2(\mathbf{z}^T,\boldsymbol{\theta}) \tag{3.3.8}$$

$$g_1(\mu_i)=\eta_{1i}=\mathbf{x}_i^T\boldsymbol{\beta} \quad \text{and} \quad g_2(\phi_i)=\eta_{2i}=\mathbf{z}_i^T\boldsymbol{\theta} \tag{3.3.9}$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_k)^T$ and $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_h)^T$ are vectors of unknown regression parameters that are assumed to be functionally independent ($\boldsymbol{\beta}\in\mathbb{R}^k$ and $\boldsymbol{\theta}\in\mathbb{R}^h$, $k+h<n$) and η_{1i} and η_{2i} are the predictors. Simas et al. [109] proposed a general class of predictors capturing both linearity and non-linearity in the data [109]. We adopted linear predictors for both parameters, to avoid computational complexity when implementing them in our application. Also, when $\phi_i=\phi$, the fixed-precision-based beta regression model proposed by Ferrari and Cribari-Neto [103] can be retrieved. The observations $x_{i1},\ldots,x_{iq_1}$ and $z_{i1},\ldots,z_{iq_2}$ correspond to q_1 and q_2 known predictors, which need not be exclusive.

3.3.3 Choices of Link Functions

The choice of link functions is critical in beta regression modeling. These functions, $g_1 : (0, 1) \to \mathbb{R}$ and $g_2 : (0, \infty) \to \mathbb{R}$, are strictly monotonic and twice differentiable. Monotonicity ensures that as the linear predictor (linear combination of predictor variables) increases, the mean of the response variable also increases (or decreases, depending on the relationship direction). This property is essential for maintaining the interpretability of the model. Twice differentiability ensures the smoothness and stability of the link function, facilitating model fitting and enhancing the reliability of statistical inference.

Various link functions are available [99], including the logit link, $g_1(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, and the probit link, $g_1(\mu) = \Phi^{-1}(\mu)$, where $\Phi(\cdot)$ denotes the standard normal distribution function. Similarly, for $g_2(\phi)$, the logarithmic link function, $g_2(\phi) = \log(\phi)$, the square root link function, $g_2(\phi) = \sqrt{\phi}$, and the identity link function, $g_2(\phi) = \phi$, are well-known, among some others.

3.3.4 Likelihood and Method of Estimation

As defined in Eq.(3.3.8), $\mu_i = g_1^{-1}(\eta_{1i})$ and $\phi_i = g_2^{-1}(\eta_{2i})$ are functions of β and θ, respectively. This beta regression family adheres to all regularity conditions outlined by Cox and Hinkley [113], ensuring its statistical robustness. Furthermore, one can demonstrate the uniqueness of the maximum likelihood estimators (MLE) within this framework. The components of the score vector, derived from differentiating the log-likelihood function with respect to the parameters, are presented in Appendix (Eqs. (A.0.1)–(A.0.10)). The maximum likelihood estimates (MLEs) of β and θ are derived by solving the nonlinear system $U(\zeta) = 0$. In practical applications, these MLEs are typically obtained through numerical maximization of the log-likelihood function using a nonlinear optimization algorithm, such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [114].

We choose linear models for our analysis to simplify computational complexity. Linear models are simpler to interpret, provide clear insights into variable relationships, yield useful statistics for assessing performance, and are easier to implement with better reproducibility. Their reduced complexity and lower optimization costs make them an efficient and practical choice.

Linear Beta Regression Model with Fixed Precision

In our approach, we adopt a linear regression framework to simplify computational complexity. For the fixed precision beta regression (FPBR) model, we use the logit mean for location modeling, where we only consider the fixed log link for the precision parameter:

$$\begin{aligned}\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1\mathrm{Age} + \beta_2\mathrm{BMI} + \beta_3\mathrm{Gender} + \beta_4\mathrm{KA} + \beta_5\mathrm{SL} + \beta_6\mathrm{WS} \\ + \beta_7\frac{\mathrm{StPh}}{\mathrm{SwPh}} + \beta_8\mathrm{KA}\times\mathrm{SL} + \beta_9\mathrm{BMI}\times\mathrm{WS} \\ + \beta_{10}\mathrm{Age}\times\mathrm{KA} + \beta_{11}\mathrm{Age}\times\mathrm{WS}\end{aligned} \quad (3.3.10)$$

Linear Beta Regression Model with Variable Dispersion

In this class of beta regression models, we employ a linear sub-model for the dispersion (or precision). For the variable dispersion beta regression (VDBR) model, we consider the following structure for the location and dispersion sub-models:

$$
\begin{aligned}
\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1\text{Age} + \beta_2\text{BMI} + \beta_3\text{Gender} + \beta_4\text{KA} + \beta_5\text{SL} + \beta_6\text{WS} \\
+ \beta_7\frac{\text{StPh}}{\text{SwPh}} + \beta_8\text{KA} \times \text{SL} + \beta_9\text{BMI} \times \text{WS} \\
+ \beta_{10}\text{Age} \times \text{KA} + \beta_{11}\text{Age} \times \text{WS}, \quad (3.3.11)
\end{aligned}
$$

$$
\ln(\phi) = \theta_0 + \theta_1\text{Age} + \theta_2\text{BMI} + \theta_3\text{Gender} + \theta_4\text{KA} + \theta_5\text{SL} + \theta_6\text{WS} + \theta_7\frac{\text{StPh}}{\text{SwPh}}. \quad (3.3.12)
$$

This entails using the same expressions as described in Eq. (3.3.9): $g_1(\mu_i) = \eta_{1i} = x_i^T\beta$, $g_2(\phi_i) = \eta_{2i} = z_i^T\theta$, where $\beta \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^h$.

3.3.5 Model Discrimination

As noted earlier, since the estimation is conducted through maximum likelihood, the standard inferential tools including Wald statistics, likelihood ratio tests, and Lagrange multiplier (score) tests become readily accessible. Model comparison is accomplished using the likelihood ratio test, which involves comparing twice the difference between the log-likelihoods of a full model and a restricted model where the covariates are a subset of the full model. Additionally, information criteria such as AIC [110] is employed for model evaluation. In model selection, metrics such as the AIC [110] and the BIC [115] are instrumental in assessing the goodness of fit of statistical models. Both metrics are represented as penalized chi-square values, with the penalties defined by the number of model parameters (k), and in the case of the BIC, the number of observations (N). The AIC is commonly defined as

$$
AIC = -2\ln L_{\text{fitt}} + 2k. \quad (3.3.13)
$$

while the BIC is defined as

$$BIC = -2 \ln L_{\text{fitt}} + 2k \ln N. \tag{3.3.14}$$

A recognized limitation of AIC is its susceptibility to small sample size, often leading to a tendency to favor more complex models, particularly with larger datasets. Hence, many researchers lean towards using BIC, which is derived from a Bayesian perspective and imposes stricter penalties on model complexity compared to AIC. Both AIC and BIC provide quantitative measures for model comparison, where lower values signify better fit while accounting for the number of parameters involved. To evaluate the overall goodness of fit of the model, an equivalent measure to the multiple R^2 in normal-theory ordinary least squares (OLS) regression would be beneficial. Ferrari and Cribari-Neto investigated the correlation between observed and predicted values as a potential measure of goodness of fit [103]. However, this approach overlooked the influence of dispersion covariates, thereby restricting its applicability [108]. So, we restrict our model comparisons to the AIC and BIC criteria, supported by various diagnostic plots. Among these tools, we use a half-normal plot with simulated envelopes, which also serves as a goodness-of-fit assessment for the model. Additional justification for selecting the model is provided in the 'Model Diagnostic' section, where we present further diagnostic plots.

3.4 Results

The development of GI using clinically relevant parameters is an effort to streamline human gait assessment [49]. Initially, a systematic literature review identified key gait parameters. A dataset comprising 120 healthy subjects was then analyzed to compute these parameters ([84], [116]) as well as their demographic factors such as age, gender, and BMI. Statistical analyses were conducted to identify the most significant gait parameters: walking speed –

WS, maximum knee flexion angle – $\mathrm{KA}_{\mathrm{max}}$, height normalized stride length – $\mathrm{SL}_{\mathrm{norm(h)}}$, and stance-to-swing phase ratio – StPh/SwPh, forming the basis for the GI as

$$\mathrm{GI} = \frac{WS \times KA_{\mathrm{max}} \times SL_{\mathrm{norm(h)}}}{StPh/SwPh}. \tag{3.3.15}$$

This composite GI would provide insights into an individual's gait pattern, enabling the detection of subtle abnormalities and facilitating continuous monitoring of changes over time. In order to make it robust, our aim in this study is to explore the influence of gait parameters, demographic factors, and gait-demographic interactions on GI.

We employ two beta regressions (fixed precision and variable dispersion) to model the GI, incorporating three demographic parameters (age, gender, and BMI) and the four gait parameters that are used in the GI formula as predictors. After exploring various combinations based on insights from previous studies [57], [58], [59], [60], [61], [62], [63], [64], [80], [81], [84], [85], [86], we carefully consider specific gait-gait and gait-demographic interaction terms for examination. For instance, we investigate the potential interaction effect between $\mathrm{KA}_{\mathrm{max}}$ and $\mathrm{SL}_{\mathrm{norm(h)}}$ on the GI, as suggested by literature findings [62], [117]. Additionally, considering the literature's indications, we examine the interaction between BMI and WS to evaluate their combined impact on the GI [118], [119]. Moreover, we explore the influence of interaction effects between age and $\mathrm{KA}_{\mathrm{max}}$, as well as age and WS [118], [119], on the GI of the selected participants.

Table 3.1: Model discrimination

Measure for model discrimination	FPBR	VDBR
ln(L)	338.1000	407.2000
AIC	-650.2007	-774.3000
BIC	-613.9633	-718.5000

We consider different link functions for the dispersion sub-model and select the log link based on model discrimination criteria (Table 3.2). The final model selection is based

Table 3.2: Comparison of the link functions for the Dispersion Sub-Model (VDBR)

Link functions for the precision parameter (ϕ)	AIC	BIC
$\log(\phi)$	-774.3199	-718.5701
ϕ	-741.9465	-686.1967
$\sqrt{\phi}$	-755.5501	-699.8002

on the criteria outlined in the "Model Discrimination" section. Initially, we applied a logit-transformed linear regression model, utilizing the additive logistic normal distribution approach [90], [91], [92], as a potential alternative to simple linear regression, which is unsuitable for bounded outcomes. However, the AIC and BIC scores for this model (-252.55 and -216.31, respectively – Appendix, Table A.1) are significantly higher compared to the VDBR and FPBR models, indicating its relative inferiority. Subsequently, we compare the VDBR and FPBR models based on AIC and BIC values (Table 3.1), where both metrics for the VDBR model are lower, demonstrating a better balance between model fit and complexity. Further justification for selecting the VDBR model is elaborated in the "Model Diagnostic" section, supported by diagnostic plots and an assessment of various precision parameter link functions (Table 3.2). Among these, the $log(\phi)$-based model within the VDBR framework exhibits the lowest AIC/BIC values, confirming its optimal fit.

3.4.1 Interpretations of the Fixed Main Effects

In this section, we present the interpretations of the model within the context of the gait data, as indicated in Table 3.3.

1. **Age:** In our study, we did not find age to be a statistically significant predictor of the GI. However, it is common knowledge that as we grow older, our walking tends to become less steady. The trend was captured by the positive slope of the regression coefficient in our study's findings. However, this change in GI due to the Age was negligible when interpreting quantitatively. For each passing year, the average GI

score decreased by roughly 0.0093 units on a standardized scale. In simpler terms, as we age, the likelihood of having a GI score above the average decreases slightly each year. For instance, if the coefficient for age was around 1.009 ($e^{0.0093}$) when exponentiated, it suggests that with every additional year, the odds of having a higher-than-average GI score decreased by 0.9% (1.009 – 1 = 0.009), provided all other factors remain constant;

2. **BMI:** Here, BMI was not a statistically significant predictor of GI at any given level of significance. This implied that changes in BMI might not have a significant impact on GI within the data we examined. However, the observed trend showed that as BMI increases, there was a slight decrease in the likelihood of having a GI score above the average. Specifically, for every increase in BMI, the odds of having a higher-than-average GI score decreased by 0.4% (1.004 – 1 = 0.004), provided all other factors remained constant;

3. **Gender:** Gender differences in gait are often explored. In our study, although there was a difference in the estimated mean GI between genders, it was not statistically significant at any listed significance level (Table 3.3) within our sample;

4. **Knee Angle ($\mathbf{KA_{max}}$):** An intriguing relationship regarding KA_{max}'s impact on the GI was observed in our analysis. The predictor was statistically significant for all the given levels of significance. As the KA_{max} increased by one unit (radians), the odds of the GI surpassing the grand mean increased by 28% ($e^{0.2437}$ – 1 = 1.28 – 1 = 0.28), when all other factors remained constant;

5. **Walking Speed (WS):** The influence of WS on the GI was statistically significant for all the given levels of significance in our study. With every one-unit increase in WS (in ms^{-1}), the odds of the GI surpassing the grand mean spiked by 23% ($e^{0.2037}$ – 1 = 1.23 – 1 = 0.23), when all other factors remained constant;

6. **Stride Length ($\mathbf{SL_{norm(h)}}$):** Another noteworthy finding involved the influence of SLnorm(h) on the GI. In our study design, the effect of $SL_{norm(h)}$ was statistically significant at the 0.1% significance level (thus, for all other levels of significance). $SL_{norm(h)}$ increased by one unit, the odds of the GI being above the grand mean increased significantly by 22% ($e^{0.1958} - 1 = 1.22 - 1 = 0.22$), hinting at the importance of longer strides for a higher GI;

7. **Stance-to-Swing Phase Ratio (StPh/SwPh):** The StPh/SwPh also emerged as a significant predictor at a 0.1% significance level (thus, for all other levels of significance). For every one-unit increase in this ratio, the odds of the GI exceeding the grand mean decreased by 16% ($e^{0.1498} - 1 = 1.16 - 1 = 0.16$), highlighting the role of this metric in understanding gait patterns.

3.4.2 Marginal Effects of Joint Interaction Factors on Gait Index (GI)

In this section, we carefully analyzed the interaction terms to understand their implications in the context of the location sub-model. These terms illuminated how the log odds of GI change due to the combined influences of specific predictor variables within the location sub-model framework. Our aim was to provide a clearer insight into the interplay between various gait and demographic factors (gait-gait and gait-demographic). Fig. 3.2 illustrates these marginal interaction effects in a 2×2 grid. Prior to model fitting, the data underwent mean-centering, ensuring each term's mean was set at 0. This normalization was crucial given the varying units of the parameters. We examined the marginal effects on the response while holding other factors constant at their respective reference levels.

1. **Knee Angle ($\mathbf{KA_{max}}$) and Stride Length ($\mathbf{SL_{norm(h)}}$):** The interaction effect between KA_{max} and $SL_{norm(h)}$ was statistically significant at all levels of significance.

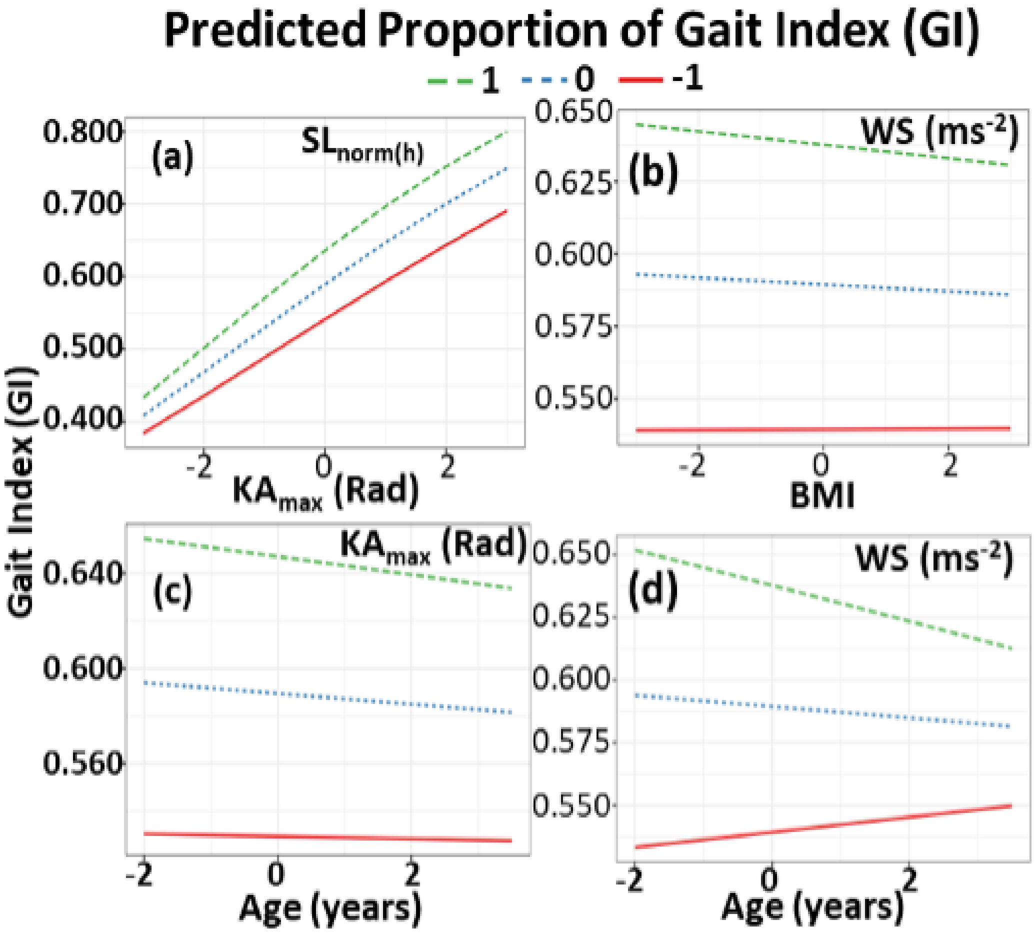


Figure 3.2: Marginal effects of joint interaction factors on the predicted Gait Index (GI): (a) KA_{max}*$SL_{norm(h)}$, (b) BMI*WS, (c) Age*KA_{max}, and (d) Age*WS. The effects are shown for mean + 2SD (dashed line, green), mean (dotted line, blue), and mean – 2SD (solid line, red).

Table 3.3: Regression coefficients and summary statistics for Fixed Precision Beta Regression Model (FPBR) and Variable Dispersion Beta Regression Model (VDBR)

FPBR				**VDBR**			
Parameters	**Coefficients**	**SE**	**p-values**	**Parameters**	**Coefficients**	**SE**	**p-values**
Location Sub-model				Location Sub-model			
β_0	0.3785	0.0121	$< 2e-16$ ***	β_0	0.3607	0.0064	$< 2e-16$ ***
β_1 (Age)	-0.0008	0.0098	0.9306	β_1 (Age)	-0.0093	0.0048	0.0526 ·
β_2 (BMI)	0.0073	0.0069	0.2855	β_2 (BMI)	-0.0048	0.0025	0.0596 ·
β_3 (Gender)	-0.0292	0.0138	0.0348 *	β_3 (Gender)	0.0009	0.0063	0.8843
β_4 (KA)	0.2590	0.0073	$< 2e-16$ ***	β_4 (KA)	0.2437	0.0043	$< 2e-16$ ***
β_5 (SL)	0.1970	0.0082	$< 2e-16$ ***	β_5 (SL)	0.1958	0.0041	$< 2e-16$ ***
β_6 (WS)	0.2123	0.0077	$< 2e-16$ ***	β_6 (WS)	0.2037	0.0039	$< 2e-16$ ***
β_7 (StPh/SwPh)	-0.1695	0.0071	$< 2e-16$ ***	β_7 (StPh/SwPh)	-0.1498	0.0028	$< 2e-16$ ***
β_8 (KA*SL)	0.0470	0.0081	$7.73e-09$ ***	β_8 (KA*SL)	0.0315	0.0036	$< 2e-16$ ***
β_9 (BMI*WS)	-0.0263	0.0077	0.0006 ***	β_9 (BMI*WS)	-0.0052	0.0037	0.1667
β_{10} (Age*KA)	-0.0063	0.0068	0.3533	β_{10} (Age*KA)	-0.0072	0.0027	0.0093 **
β_{11} (Age*WS)	-0.0204	0.0073	0.0056 **	β_{11} (Age*WS)	-0.0213	0.0032	$3.09e-11$ ***
Precision parameter (ϕ)				Dispersion sub-model			
Statistical significance levels (p-values): *** 0.001, ** 0.01, * 0.05, · 0.1	log(ϕ)	6.973	0.129	$< 2e-16$ ***			
	θ_0	7.6736	0.2371	$< 2e-16$ ***			
	θ_1 (Age)	0.0816	0.1730	0.6369			
	θ_2 (BMI)	0.6092	0.1513	$5.70e-05$ ***			
	θ_3 (Gender)(Male)	0.6580	0.2949	0.0257 *			
	θ_4 (KA)	-0.8841	0.1517	$5.67e-09$ ***			
	θ_5 (SL)	-0.4710	0.1565	0.0026 **			
	θ_6 (WS)	-0.2536	0.1658	0.1262			
	θ_7 (StPh/SwPh)	-0.4481	0.1634	0.00612 **			

We plotted the marginal effects of $\mathrm{KA}_{\mathrm{max}}$ on the predicted GI across three groups of $\mathrm{SL}_{\mathrm{norm(h)}}$, based on two standard deviations (SDs) above or below the mean, as shown in Fig. 3.2.(a). For individuals with SLnorm(h) two SDs above the mean, an increase in $\mathrm{KA}_{\mathrm{max}}$ corresponded to an increase in GI. Similar positive relationships were observed for those with average $\mathrm{SL}_{\mathrm{norm(h)}}$ and 2 SDs below the mean. This indicates that with each simultaneous increase in $\mathrm{KA}_{\mathrm{max}}$ and $\mathrm{SL}_{\mathrm{norm(h)}}$, the odds of the GI exceeding the grand mean rose significantly. This interaction term is statistically significant at all levels of significance;

2. **BMI and Walking Speed (WS)**: The interaction effect between BMI and WS was not statistically significant at any level. However, we observed the marginal effects of BMI on GI for individuals with WS two SDs above the mean, average WS, and 2 SDs below the mean (Fig. 3.2.(b)). While the pattern was not entirely conclusive, it appears that people with higher-than-average WS experienced a decrease in GI as BMI increased, suggesting a potential negative relationship;

3. **Age and Knee Angle ($\mathbf{KA}_{\mathbf{max}}$)**: The interaction effect between age and KAmax on GI was statistically significant at 1% level. The marginal effects of age on GI were

plotted for individuals with $\mathrm{KA}_{\max}$ change 2 SDs above the mean, average $\mathrm{KA}_{\max}$ change, and 2 SDs below the mean in Fig. 3.2.(c). The results indicate that higher $\mathrm{KA}_{\max}$ changes are associated with a decrease in GI as age increased, indicating a negative relationship;

4. **Age and Walking Speed (WS)**: The interaction effect between age and WS on GI was statistically significant at all levels in our study. We examined the marginal effects of age on GI for individuals with WS two SDs above the mean, average WS, and two SDs below the mean (Fig. 3.2.(d)). While individuals with average or higher than average WS show a decrease in GI as age increased, those lower-than-average WS exhibit the opposite patterns, suggesting a nuanced relationship between age, WS, and GI.

3.4.3 Uncertainty Analysis of Fixed Effects (FPBR vs VDBR)

Understanding the uncertainty inherent in estimating fixed effects is crucial for assessing the reliability of our findings. In this section, we analyze uncertainty surrounding the estimation of fixed effects on the GI and present the estimates for the location sub-model along with 95% confidence intervals (CI). Fig. 3.3 presents the uncertainty surrounding the estimation of each fixed effect on the GI, accompanied by a 95% CI. These estimates offer a range within which we can reasonably expect the true population parameter to lie. The estimates, presented in odds ratio scales, involve exponentiating the log-odds estimates of the predictors while preserving their original signs.

Combined coefficient plots of FPBR and VDBR explain how incorporating a dispersion sub-model in VDBR mitigates the higher uncertainty observed in FPBR's coefficient estimates (Fig. 3.3). This variable dispersion model helpa to handle the data heteroscedasticity, resulting in more precise estimates. The width of CIs reflects the uncertainty associated

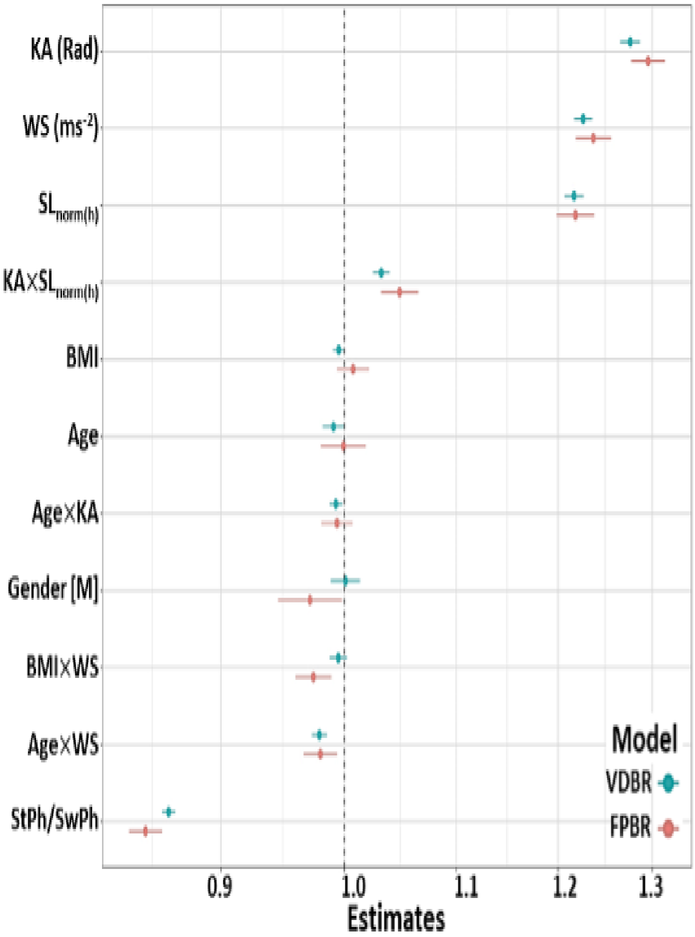


Figure 3.3: Coefficient plot comparing the estimates of VDBR and FPBR models for the location sub-model, shown in the odds-ratio scale, with 95% confidence intervals (CIs). Blue indicates VDBR effects, while red denotes FPBR effects of predictors on the response variable.

with each coefficient estimate, with wider intervals indicating greater uncertainty and narrower intervals suggesting more precise estimates. For instance, the gender coefficient in the FPBR model is statistically significant at a 5% level of significance for our study. However, the coefficient plot in Fig. 3.3 reveals that it has the widest CI, indicating high uncertainty in estimation. This is attributed to a relatively large standard error (SE: 0.0138) of the estimates, as shown in Table 3.3. Consequently, Gender cannot be interpreted as a predictor of GI in our study due to the variability of the GI within the male and female populations. VDBR addresses these issues by handling data heteroscedasticity through a dispersion sub-model, providing more stable CIs for the estimates. The indication of variability in the GI for male and female population is studied in more detail in the precision sub-model in sub-section 3.3.4.

In Fig. 3.3, coefficients with negative slopes are positioned left of the dashed reference line, indicating a negative relationship, while those with positive slopes are on the right, indicating a positive relationship between the predictor and the response variable. The color scheme distinguished between VDBR (blue) and FPBR (red) estimates and their corresponding CIs.

3.4.4 Interpretations of the Precision Sub-Model

In this section, we examine the precision sub-model's estimates (precision parameter (ϕ)), which assesses the dispersion of the beta distribution around its mean, providing crucial insights into the variability of the GI (Table 3.3). The intercept in the precision sub-model signifies the baseline precision of the GI when all predictors are at their reference levels (zero), which is approximately 7.67 (log estimate). Positive coefficients for the predictors indicate increased variability in the GI, while negative coefficients suggest reduced variability. As well, positive coefficients contribute to over-dispersion, while negative coefficients contribute to under-dispersion. Below, we interpret each predictor and its impact on the variability in the GI:

1. **Age:** The positive coefficient indicates that older age is associated with increased variability in the GI, although this effect is not statistically significant in the precision sub-model of VDBR, and its magnitude is relatively lower compared to other predictors;

2. **BMI:** Higher BMI (positive coefficient) is associated with increased variability in the GI. This coefficient is strongly significant in the dispersion sub-model (at all the given levels of significance);

3. **Gender:** Being male (as indicated by the gender coefficient) is associated with higher variability in the GI compared to females, a significant finding in the precision sub-model. This effect on the variability in the GI is statistically significant at a 5% level of significance for the dispersion sub-model;

4. **Stride Length ($\mathbf{SL_{norm(h)}}$):** This predictor is statistically significant at a 1% level of significance in our dispersion sub-model. An increase in $\mathrm{SL_{norm(h)}}$ is associated with a decrease in the variability of the GI. This indicates that individuals with higher $\mathrm{SL_{norm(h)}}$s exhibit less variability in their GI. This effect is significant as well;

5. **Walking Speed (WS):** An increase in WS was associated with a decrease in the variability of the GI, which implied a reduction in variability.

6. **Knee Angle ($\mathbf{KA_{max}}$):** This predictor is statistically significant for all the levels of significance given in our dispersion sub-model. Like SLnorm(h), greater $\mathrm{KA_{max}}$ values indicated lower variability in the GI. This effect is statistically significant, suggesting that $\mathrm{KA_{max}}$ plays a significant role in determining the consistency of GI measurements;

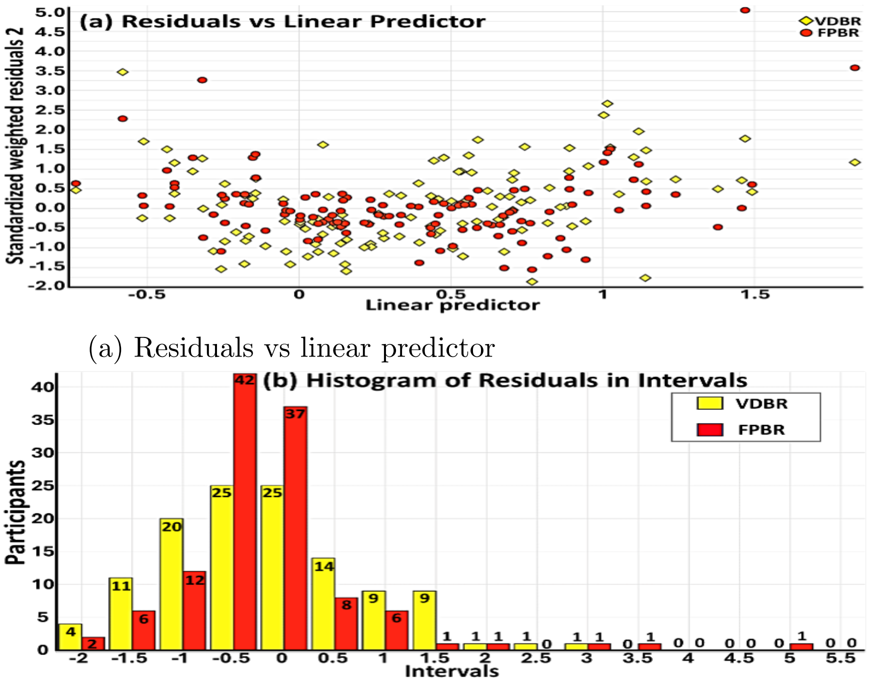
7. **Stance-to-Swing Phase Ratio (StPh/SwPh):** Finally, for StPh/SwPh, higher values of this ratio suggests lower variability in the GI. This effect is statistically significant ($p = 0.00263$) at a 1% level of significance, highlighting the importance of this ratio in gait index variability.

In summary, certain predictors like $\mathrm{SL}_{\mathrm{norm(h)}}$, StPh/SwPh, WS, and $\mathrm{KA}_{\mathrm{max}}$ had negative coefficients, suggesting that an increase in these predictors leads to a decrease in the dispersion parameter, potentially resulting in under-dispersion. In contrast, predictors such as age, BMI, and male gender had positive coefficients, indicating that an increase in these predictors is associated with an increase in the dispersion parameter, possibly contributing to over-dispersion. To understand the overall impact of predictors on dispersion, we need to consider the collective effect of all predictors in the model. This involves comparing the effects of predictors associated with under-dispersion to those associated with over-dispersion. If predictors linked to under-dispersion outweigh those associated with over-dispersion, the data are more likely to exhibit under-dispersion overall or vice-versa. Understanding these effects will help to comprehend the factors influencing variability in the GI and provide insights into the data's heterogeneity.

3.4.5 Model Diagnosis

In this section, we generate different diagnostic plots for two beta regression models: the VDBR model and the FPBR model. These plots enable us to assess model reliability by evaluating how well the model fits the data, identifying and alleviating the effects of outliers, and checking model assumptions. We focus on two key diagnostic measures: residuals versus fitted values and Cook's distance, along with half-normal plots with simulated envelopes as diagnostic tools for assessing the goodness of fit and the reliability of the model.

In diagnostics, it is crucial to measure influence and assess residuals. Initially, Ferrari



(a) Residuals vs linear predictor

(b) Histogram of residuals in intervals.

Figure 3.4: Model diagnosis for VDBR and FPBR.

and Cribari-Neto [103] derived deviance residuals based on log-likelihood contributions. However, their model lacked dispersion covariates, creating uncertainty when interpreting deviance residuals in the presence of dispersion covariates. This issue was later addressed in their 'betareg' package, which we use in our analysis to extract residuals [120]. To compare and evaluate the effectiveness of the VDBR and FPBR models in capturing the underlying data characteristics, we overlay the residuals versus linear predictor plots for both models. This comparison guided our model selection process. Additionally, we generated histograms of residuals in 0.5 intervals to compare the concentration of residuals around zero for both models and assess the impact of extreme observations.

The residuals versus fitted values plot is used to assess model performance, identify patterns, and detect influential outliers. Residuals represent the difference between observed and predicted values, while fitted values are model predictions. This plot is essential for detecting systematic deviations from model assumptions. In residual analysis, the raw response residuals $(y_i - \hat{\mu}_i)$ are often avoided due to the inherent heteroscedasticity. Instead, Pearson residuals, also known as standardized ordinary residuals, are used as a more reliable alternative (Equation 3.4.1) [103]:

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mathrm{Var}}(y_i)}}, \tag{3.4.1}$$

where

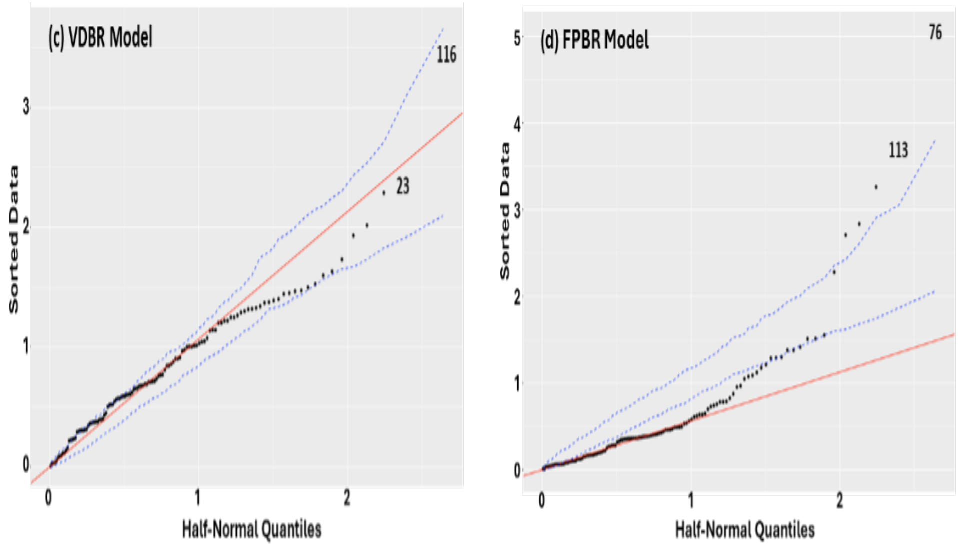
$$\hat{\mathrm{Var}}(y_i) = \frac{\hat{\mu}_i(1-\hat{\mu}_i)}{1+\hat{\phi}_i}, \quad \hat{\mu}_i = g_1^{-1}(x_i^T\hat{\beta}), \quad \hat{\phi}_i = g_2^{-1}(x_i^T\hat{\theta}). \tag{3.4.2}$$

This calculation incorporates the varying dispersion parameters into the residuals for further analysis. Additionally, Espinheira, Ferrari, and Cribari-Neto [121] introduced additional residuals including one with enhanced properties termed standardized weighted residual 2:

$$r_{SW2,i} = \frac{y_i^* - \mu_i^*}{\sqrt{\hat{\upsilon}_i(1-h_{ii})}}, \tag{3.4.3}$$

where $y_i^* = \log\left(\frac{y_i}{1-y_i}\right)$, $\upsilon_i = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)$ and $\psi(\cdot)$ is the digamma function. Standardization is then performed by $\mu_i^* = \psi'(\mu_i\phi_i) - \psi'((1-\mu_i)\phi_i)$ and h_{ii} refers to the i-th diagonal element of the hat matrix. Here, hats denote the evaluation at the maximum likelihood (ML) estimates.

In the FPBR model (constant dispersion), the residuals versus fitted values plot reveals heteroscedasticity and systematic trends, indicating violations of the constant dispersion assumption. Fig. 3.4.(a) shows a clustering of nearly 65% of residuals around the central region, with notable influences at the plot's extremes. The histogram in Fig. 3.4.(b) confirms this concentration near zero, reflecting standardized residuals but also showing the influence of extreme observations. In contrast, the VDBR model demonstrates residuals consistently spread across fitted values, effectively capturing data variability. In Fig. 3.4.(a), the VDBR residuals scatter randomly around zero, indicating robust relationships between predictors and responses.



(a) Half-normal plots of absolute residuals with simulated envelopes for VDBR model.

(b) Half-normal plots of absolute residuals with simulated envelopes for FPBR model.

Figure 3.5: Model diagnosis for VDBR and FPBR.

To further evaluate goodness-of-fit, we utilize half-normal plots of model diagnostics, including residuals, Cook's distance, and leverage. These plots are constructed by plotting the ordered absolute values of diagnostics against the expected order statistics of a half-normal distribution. Simulated envelopes, as proposed by Atkinson [91], are added to assess model consistency, with observed points expected to fall within the envelopes for a correctly specified model. The steps for creating Half-normal plots with a simulated envelope can be summarized as follows:

1. Fit the model and generate a simulated sample of n independent observations using the fitted model as if it were the true model;

2. Fit the model to the generated sample, and compute the ordered absolute values of the residuals;

3. Repeat Steps (1) ad (2) k times;

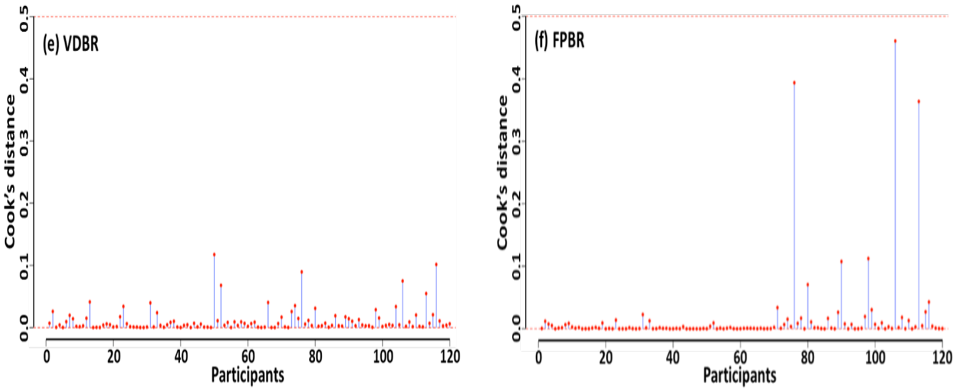
4. Consider the n sets of the k order statistics; for each set compute its average, minimum and maximum values;

5. Plot these values and the ordered residuals of the original sample against the half-normal scores:
$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right)$$

The minimum and maximum values of the k order statistics yield the envelope. In this study, we use $n = 100$ simulations to compute percentiles at each expected order statistic. Following Atkinson [91], we set $k = 19$, resulting in an approximate probability of 0.05 for an absolute residual to fall outside the envelope. Fig. 3.5.(b) (FPBR model) reveals that a significant proportion of points fall outside the simulated envelope, indicating poor model fit and inconsistency in residuals. Conversely, Fig. 3.5.(a) (VDBR model) shows more residuals within the simulated envelope, demonstrating consistency between the observed residuals and the fitted VDBR model. Also, Cook's distance is used to evaluate the influence of individual observations on model coefficients. High Cook's distance values indicate observations that could unduly affect model outcomes [122]. The 'betareg' package incorporates Cook's distance for variable dispersion models, extending the original formulation by Ferrari and Cribari-Neto to account for a variable precision parameter:

$$C_i = \frac{h_{ii} r^2_{SW2,i}}{1-h_{ii}}. \tag{3.4.4}$$

By comparing diagnostic plots, we assess the ability of the VDBR and FPBR models to capture underlying data characteristics. Fig. 3.6.(b) (FPBR model) shows abrupt spikes in



(a) Cook's distance for VDBR.

(b) Cook's distance for FPBR.

Figure 3.6: Model diagnosis for VDBR and FPBR.

Cook's distance, reflecting susceptibility to influential points and a lack of robustness. In contrast, Fig. 3.6.(a) (VDBR model) reveals better handling of data heteroscedasticity.

In conclusion, the comparison of diagnostic plots between the VDBR and FPBR models provides evidence of the VDBR model's superior ability to capture the underlying heteroscedasticity inherent in the dataset. Overall, the VDBR model demonstrates a better fit by effectively accounting for data variability and reducing the impact of extreme observations, providing more reliable and accurate results.

3.5 Discussions

In this study, we have employed Beta regression models to investigate the relationship between various demographic and gait-related predictors and GI. This provides insights for clinicians to understand and assess gait patterns from an inferential perspective. Our findings shed light on the main effects of age, BMI, gender, $\mathrm{KA}_{\mathrm{max}}$, $\mathrm{SL}_{\mathrm{norm(h)}}$, WS, and StPh/SwPh on GI variability and the interaction effects among these predictors.

Interestingly, none of the demographic factors were statistically significant predictors of GI in our analysis, though trends were observed. Although age did not show a statistically significant (not at 5%, but was significant at 10%) effect on the gait index (GI) in our study, there was a trend indicating a potential decline in GI with increasing age. This observation aligns with existing literature ([49], [71], [80]) suggesting that age-related changes in gait, such as reduced gait speed, altered stride length, and decreased joint mobility, can contribute to increased gait variability. BMI and gender also lacked statistical significance, indicating minimal direct influence on GI individually. However, a difference in the estimated mean GI between males and females hinted at a complex gender-related role. While FPBR analysis suggested an association between gender and GI, wide confidence intervals introduced uncertainty, whereas the VDBR model revealed a significant gender-related influence on GI variability. Hence, further research is needed to better understand these complex relationships. Among gait-related predictors, increases in $\mathrm{KA}_{\mathrm{max}}$, $\mathrm{SL}_{\mathrm{norm(h)}}$, and WS were associated with higher odds of GI being above the grand mean, while higher StPh/SwPh was linked to decreased odds. Overall, all the selected gait-related predictors demonstrated a significant effect on GI.

Additionally, we identified significant interaction effects between $\mathrm{KA}_{\mathrm{max}}$ and $\mathrm{SL}_{\mathrm{norm(h)}}$, age and $\mathrm{KA}_{\mathrm{max}}$, and age and WS. These interaction effects underscored the complex interplay between demographic and gait-related factors in shaping gait patterns. In our analysis, we observed a significant interaction effect between $\mathrm{KA}_{\mathrm{max}}$ and $\mathrm{SL}_{\mathrm{norm(h)}}$, exhibiting a positive slope. This indicated the importance of evaluating the individual impacts of $\mathrm{KA}_{\mathrm{max}}$ and $\mathrm{SL}_{\mathrm{norm(h)}}$ and their combined effect for a comprehensive gait assessment. Examining interactions between age and $\mathrm{KA}_{\mathrm{max}}$, as well as age and WS, suggested that aging influenced $\mathrm{KA}_{\mathrm{max}}$ and WS, with their combined effect subtly impacting gait patterns, particularly in older individuals. This may explain the near significance of age's effect on GI, as WS and $\mathrm{KA}_{\mathrm{max}}$ partially masked their individual impact. These findings underscored the importance of interaction effects in age-related gait changes and the need for further research.

Moreover, our analysis identified variability in gait performance (GI) associated with various predictors. Higher BMI was linked to increased variability in gait patterns (GI), while being male was associated with higher variability than females. On the other hand, shorter $\mathrm{SL}_{\mathrm{norm(h)}}$s, slower WSs, lower $\mathrm{KA}_{\mathrm{max}}$s, and higher StPh/SwPhs were all associated with increased variability in gait performance. Understanding these sources of variability is crucial for clinicians in accurately assessing gait patterns and for developing targeted interventions to improve mobility and function in patients with unstable gait patterns (lower GI value).

In addition to these findings, the VDBR model outperformed the FPBR model in capturing data heteroscedasticity by effectively identifying covariates, such as gender and BMI, that contribute to variability. Unlike traditional approaches that emphasize gender differences solely through means ([82], [85]), the VDBR model highlighted gender as a significant marker of gait variability, providing a more nuanced perspective—a similar observation applies to BMI. Although we did not find a strong association, BMI's impact on gait stability may be influenced by factors like joint pain, muscle weakness, and mobility limitations. Our previous work confirmed GI deterioration with abnormal BMI [49], [71]. In this study, individuals with higher WS showed a decline in GI as BMI increased, and BMI significantly contributed to data heteroscedasticity, further emphasizing its complex role in explaining gait stability. By incorporating a dispersion sub-model, the VDBR model delivered more stable CIs and effectively addressed outliers, as confirmed by Cook's distance analysis. Furthermore, half-normal plots demonstrated a better fit, highlighting the VDBR model's superiority in capturing GI score variability and its suitability for gait analysis.

Unlike prior studies that typically employed multiple regression models for analyzing individual variations in gait parameters [65], [66], [67], our work took a different approach by using the GI as a unified representation of gait stability. This model-based framework allowed us to explore how key gait and demographic parameters, and their interactions, influence gait stability more holistically. Furthermore, the dispersion sub-model was instrumental in identifying the specific gait and demographic parameters responsible for variability in GI, providing a clearer understanding of the factors contributing to fluctuations in gait stability within this cohort. Overall, our findings will provide clinicians with valuable insights into how both demographic and gait-related factors contribute to GI variability. By addressing specific factors contributing to variability and considering their interactions, clinicians can tailor interventions to optimize gait outcomes, ultimately improving the quality of life for individuals with unstable gait patterns (lower GI).

Chapter 4

Understanding Patient-level Gait Predictions using an Interpretable Machine Learning (ML) Model: A BART Framework

4.1 Introduction

The significance of gait analysis in understanding and addressing various health conditions has gained prominence, transcending from a biomechanical curiosity to a pivotal aspect of clinical evaluation [125], [126], [127], [128], [129], [130]. Gait analysis holds a unique position as a non-invasive, dynamic assessment tool that offers insights into musculoskeletal function, cardiovascular health, neurological integrity, and overall mobility [131], [132], [133], [134], [135], [136], [137], [138], [139]. Wearable sensor-based data collection is easy to implement in routine clinical practice, enabling researchers to gather data from large populations. Despite the high volume of published studies, only a small portion of this data is freely accessible and well-documented for reuse. Open and curated gait datasets aim to address this gap by serving two main purposes: first, they allow clinicians to test and compare clinical hypotheses—such as the ability of walking patterns to distinguish fallers from non-fallers [140], [141]; second, they enable bioengineers to design and evaluate the accuracy of algorithmic methods [142]. The current trend of patient-level gait analysis focuses on developing prediction models based on wearable devices. A wearable inertial measurement unit (IMU) attached to the lower limb region was found to be the most common approach. Wearable sensors' most extracted quantitative gait features were temporal, spatial, and spatiotemporal characteristics [143]. Based on such extracted parameters, different prediction models have been built to understand patient-level nuances in their gait characteristics between pathological groups and healthy cohorts. Early identification of disorder-specific gait deficits, coupled with monitoring disease progression, allows for implementing targeted, preventive, and personalized treatments for patients based on their pathology. Previous studies have investigated patient-level gait characteristics in various neurological disorder groups, such as Parkinson's disease (PD)[144] and its subtypes, as well as orthopedic disorder groups, including osteoarthritis [145] and its subtypes. Neurodegenerative diseases, such as PD have a significant effect on motor functions, especially movement impairments [146], [147], [148]. Multiple simple threshold-based analytic approaches have been used for Freezing of gait (FOG) detection in PD patients using wearable sensors [149], [150]. Traditional machine-learning (ML) models, such as support vector machines (SVM) [151], [152], and decision-tree-based models like random forests (RF) [153], [154], [155], have been used along with deep-learning models, including convolutional neural networks (CNN) [154], [156] and, recurrent neural networks (RNN) [157] for FOG detection using wearable sensors. More cutting-edge black box deep-learning tools such as transformers, autoencoders and Long Short Term Memory (LSTM) network have been employed for quantitative gait analysis

[158], [159], [160]. These studies focused on prediction or enhancing prediction accuracy rather than on the interpretability of the results. Furthermore, while advanced machine learning approaches offer high prediction accuracy, their practical use in clinical settings remains limited because their predictions are often difficult to interpret and, therefore, not actionable. Interpretable methods clarify why a specific prediction was made for a patient by identifying the patient-level gait characteristics that contributed to the result. To date, this lack of interpretability has constrained the adoption of powerful techniques like deep learning and ensemble models in medical decision support. Previous review studies [143] have highlighted similar challenges when summarizing the contributions of machine learning algorithms in wearable sensor gait analysis. Only a few studies, such as the one in [161], have considered deep learning approaches for estimating spatiotemporal gait features. However, these models still lack local interpretability, such as defining the decision path, ensuring model convergence, explaining uncertainty in predictions, or attributing importance to features used in predictions. To the best of our knowledge, this is the first study on wearable sensor gait analysis to propose a Bayesian Additive Regression Tree (BART)-based [162] framework that not only predicts patient-level classifications between healthy and pathological cohorts but also explains the predictions by offering a comprehensive Bayesian inference for the decision path, model diagnostics, and feature importance. Understanding what drives a prediction is important for determining targeted interventions in a clinical setting. For this reason, machine-learning methods employed in clinical gait analysis applications avoid using complex yet more accurate models and retreat to simpler interpretable (for example, linear) models at the expense of accuracy. Studies like [163] often use correlation analysis-based approaches or linear models to achieve interpretability at the expense of the prediction accuracy. These approaches cannot account for non-linearity and higher-order interactions within the data. Additionally, simple linear models are highly

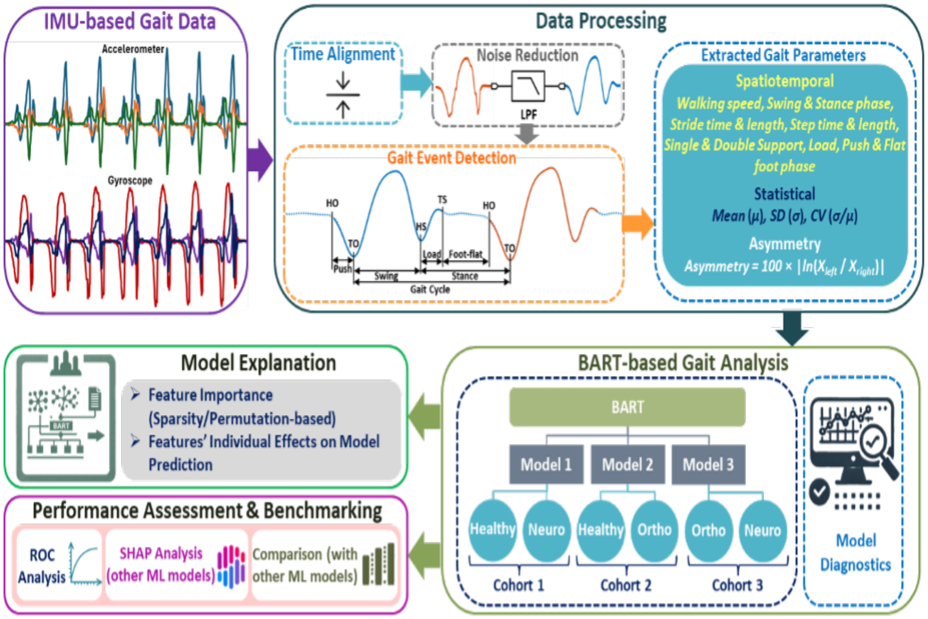

IMU-based Gait Data
Accelerometer
Gyroscope
Data Processing
Time Alignment
Noise Reduction
LPF
Gait Event Detection
HO
TO
HS
TS
HO
TO
Push
Load
Foot-flat
Swing
Stance
Gait Cycle
Extracted Gait Parameters
Spatiotemporal
Walking speed, Swing & Stance phase, Stride time & length, Step time & length, Single & Double Support, Load, Push & Flat foot phase
Statistical
Mean (μ), SD (σ), CV (σ/μ)
Asymmetry
Asymmetry = 100 × |ln(X_left / X_right)|
Model Explanation
BART
Feature Importance (Sparsity/Permutation-based)
Features' Individual Effects on Model Prediction
BART-based Gait Analysis
BART
Model 1
Model 2
Model 3
Healthy
Neuro
Healthy
Ortho
Ortho
Neuro
Cohort 1
Cohort 2
Cohort 3
Model Diagnostics
Performance Assessment & Benchmarking
ROC Analysis
SHAP Analysis (other ML models)
Comparison (with other ML models)

Figure 4.1: Preprocessing of the datasets and development of the interpretable BART-based framework.

sensitive to correlated feature spaces, which can lead to biased results—such as inaccurate p-values and inflated confidence intervals—if left unaddressed. In our study, we have demonstrated how to retain interpretability with complex models, such as non-parametric Bayesian methods, by developing a framework that provides theoretically justified explanations of model predictions, model diagnostics, and feature importance. This framework aligns with recent advances in model-agnostic prediction explanation methods [164], [165], [166].

The Bayesian Additive Regression Tree (BART) is a "sum-of-trees" model in which each tree is constrained by a regularization before acting as a weak learner. Fitting and inference are performed using an iterative Bayesian back-fitting MCMC algorithm that generates samples from the posterior distribution [162]. BART can accommodate collinearity, non-linearity present in the data subspace and higher-order interactions, along with estimating random effects present in the model [167]. BART has gained significant attention in various applications like ecology [168], modelling uncertainty in the economy [169], and hourly streamflow forecasting [170]. BART is highly interpretable due to its built-in feature importance criteria. Moreover, it facilitates permutation-based tests that provide a statistical justification for the significance of each feature [171]. BART's performance has been evaluated in both lower-dimensional cases and in high-dimensional, settings with sparsity considerations. The sparsity-induced prior-based modifications result in reasonable feature importance and improved performance compared to other decision-tree ensemble methods [172]. Furthermore, recent theoretical studies indicate that these models achieve a near-minimax posterior concentration rate across a wide range of prediction functions, supporting the empirical success of BART and its variants from a theoretical standpoint [173].

In this research, we utilize the IOPL dataset [142], which comprises 1,020 time series, each accompanied by various contextual metadata. The dataset captured approximately 8.5 hours of gait signals from 230 subjects performing a sequence of simple movements, including standing, walking, and turning, in a fixed order. More than 40,000 footsteps were manually annotated with specified start and end timestamps. We preprocessed the data (see subsection on Preprocessing) and converted it into a standard tabular format with extracted temporal, spatial, and spatiotemporal gait features. A correlation analysis of the predictor space indicated a high correlation among the predictors. We performed binarization based on healthy versus pathological cohorts and further differentiated between two pathological groups: orthopedic and neurological. This binarization enabled us to better understand the patient-level gait characteristics of the disorder groups in comparison to the healthy cohort, which served as the reference level.

We evaluated BART's out-of-sample performance across cohorts using various performance metrics. Model diagnostics were also assessed to verify convergence and ensure prediction reliability for each cohort. We benchmarked BART's performance against traditional machine-learning models, including support vector machines (SVM), decision-tree-based models, and logistic regression. Each cohort was evaluated, and a comparison of model performance was conducted. Unlike previous studies [151], [152], [153], [154], [155], [156], BART provides an explanation of the importance of each feature for each cohort. Individual effect plots, such as Accumulated Local Effects [174], which are more robust towards collinearity, were evaluated for selected features to better understand their relationship with the model's predictions. In our study, BART produced biologically meaningful results aligned with previous studies [175], [176], [177], [178], [179], [180], [181], [182], [183], [184], [185], [186], [187], [188], [189], [190], [191], [192]. To further assess the consistency of important features indicated by BART, we conducted a SHAP analysis [193] with benchmarked machine learning (ML) algorithms across each cohort. Feature importance does not establish causation and so does not offer a complete diagnosis of a patient's disorder. However, they provide clinicians with insights into the gait characteristics and procedural factors that influenced the model's predicted pathology, supporting more informed diagnostic decisions.

4.2 Methods

We deployed BART on this transformed dataset and evaluated the performance. To illustrate the value of the BART-explained predictions and provide insight into factors influencing the classification between healthy and pathological groups, we present the following details.

4.2.1 Procedures

Preprocessing: This study focuses on analyzing IMU-based gait data collected from 230 participants across three distinct pathology groups: Healthy, Orthopedic, and Neurological

[142]. Time-series data from IMU sensors was processed to extract critical gait features, which were essential for distinguishing walking patterns across the groups. IMU sensors were placed on each participant's feet, capturing accelerometer and gyroscope data at 100 Hz. These sensors captured three-dimensional data, including accelerometer readings for acceleration and gyroscope readings for angular velocity along the X, Y, and Z axes. Key gait events, such as Heel-Off (HO) and Toe-Strike (TS), were manually annotated in the dataset by specialists using a software tool that displayed the relevant sensor signals, allowing precise marking of these events. In addition to utilizing the manually annotated events, Toe-Off (TO) and Heel-Strike (HS) events were estimated by us using a peak detection method applied to the accelerometer and gyroscope signals.

After detecting key gait events, we extracted 35 features capturing temporal, spatial, and phase-specific characteristics for each trial (see Table 4.1). Temporal features included stride time (time between successive heel strikes of the same foot), swing time (duration the foot is off the ground), and stance time (duration the foot is in contact with the ground), along with single and double support times. Spatial features included walking speed (from stride length and cadence), stride length (distance between successive heel strikes of the same foot), and step length (distance between consecutive heel strikes of alternating feet). Phase-specific characteristics captured load (heel-strike to flat foot), push (heel-off to toe-off), and flat foot phases. To assess gait variability, standard deviation (SD) and coefficient of variation (CV $= \sigma/\mu$) were calculated for key parameters, and asymmetry between left and right feet was computed for several parameters. Asymmetry between the left and right feet was also computed for multiple parameters using the following equation:

$$\text{Asymmetry} = 100 \times \left| \ln \left(\frac{X_{\text{left}}}{X_{\text{right}}} \right) \right|, \tag{4.2.1}$$

where X_{left} and X_{right} represent the values of a specific gait parameter for the left and right feet.

Finally, several demographic features, including age, gender, BMI (calculated from height and weight), dominant foot (laterality), and pathology group (Healthy, Orthopedic, or Neurological), were included as part of the feature set. These features collectively enabled a detailed examination of gait mechanics and the identification of patterns unique to each pathology group.

4.2.2 Model Architecture

Bayesian Additive Regression Tree or BART is a Bayesian approach to non-parametric function estimation by growing multiple decision trees. Suppose we have a continuous response variable Y and p covariates X for n subjects. The objective is to develop a model capable of capturing intricate relationships between X and Y, with a focus on predictive performance. BART aims to estimate $f(X)$ using models of the form: $Y = f(X) + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$. To approximate $f(X)$, a summation of regression trees is formulated as

$$f(X) = \sum_{j=1}^{m} g(X; T_j, M_j). \tag{4.2.2}$$

Predictions are based on the sum-of-these-tree models, where each tree is composed of leaves and terminal nodes. So, predictions are taken from each tree and added together to get the total estimates. In Eq. (4.2.2), T_j is the j^{th} binary tree structure and $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ is the vector of terminal node parameters associated with T_j. For each binary regression tree T_j and its associated set of terminal node parameters M_j, the function $g(x; T_j, M_j)$ assigns a value $\mu_{kj} \in M_j$ to x. Under this formulation, the conditional expectation $E(Y \mid X)$ is given by the sum of all terminal node parameters μ_{kj} assigned by the functions $g(x; T_j, M_j)$. Here, μ_{kj} is the mean parameter of the k^{th} node for the j^{th}

regression tree. Additionally, each μ_{kj} represents a main effect when $g(x; T_j, M_j)$ depends on a single component of x (i.e., a single variable), whereas it represents an interaction effect when $g(x; T_j, M_j)$ depends on multiple components of x (i.e., multiple variables).

Thus, the defined tree-based ensemble model is capable of capturing both main effects and interaction effects. Moreover, since the trees in this formulation can vary in size, the interaction effects can be of different orders. In the special case where every terminal node assignment depends only on a single component of x, the sum-of-trees model simplifies to an additive function (a sum-of-step functions) based on the individual components of x. However, there are some computational challenges on detecting higher order (more than 2) interaction terms [162]. There are many descriptions and derivations of the architecture of BART models [162] [167]. In this work, we follow the original formulation by Chipman et al. [162] and the general BART framework outlined in [167].

Priors of BART

Now that we have a conceptual understanding of how the BART algorithm operates, we proceed with a more rigorous explanation. We begin by specifying the prior distributions for BART.

The prior distribution for Eq. (4.2.2) is given by $P(T_1, M_1, \ldots, T_m, M_m, \sigma)$. A common prior specification assumes that $\{(T_1, M_1), \ldots, (T_m, M_m)\}$ and σ are independent, and that each pair (T_j, M_j) is independent of the others. This allows us to simplify the prior specification problem to the specification of forms expressed as

$$P((T_1, M_1), \ldots, (T_m, M_m), \sigma) = P((T_1, M_1), \ldots, (T_m, M_m))P(\sigma)$$

$$= \left[\prod_{j=1}^{m} P(T_j, M_j)\right] P(\sigma)$$

$$= \left[\prod_{j=1}^{m} P(M_j|T_j)P(T_j)\right] P(\sigma)$$

$$= \prod_{j=1}^{m}\prod_{k=1}^{b_j} P(\mu_{kj}|T_j)P(T_j)P(\sigma). \tag{4.2.3}$$

From the third to the fourth line in Eq. (4.2.3), recall that $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ represents the vector of terminal node parameters associated with T_j, and each node parameter μ_{kj} is typically assumed to be independent.

Thus, Eq. (4.2.3) indicates that we need to specify prior distributions only for $\mu_{kj}|T_j$, σ, and T_j.

Regularization prior or $P(T_j)$

A prior on T_j governs the depth of each tree using a negative power distribution, which can be specified by three aspects:

1. The probability that a node at depth $d = 0, 1, \ldots$ will split is given by $\frac{\alpha}{(1+d)^{\beta}}$.

   where the parameter $\alpha \in (0, 1)$ determines the likelihood of a node splitting—higher values of α increase the probability of a split. The parameter $\beta > 0$ controls the number of terminal nodes, with larger values of β leading to fewer terminal nodes.

   This property plays a crucial role in BART as it acts as a regularization mechanism, preventing overfitting and ensuring the convergence of BART to the target function $f(X)$ (Rokčová and Saha, 2018)[194]. As discussed in the previous subsection, this feature also allows many shallow (weak) regression trees to be fit and eventually summed together to obtain a stronger model;

2. The distribution used to determine which covariate an internal node splits on is typically chosen as the uniform distribution. However, if we have correlated predictors

or high-dimensional predictor spaces, the assumption of a uniform distribution could lead to biased results in variable selection. Recent studies [194][172] have suggested that a uniform distribution does not inherently encourage variable selection;

3. The distribution used to determine the cutoff point within an internal node, after selecting the covariate, is typically chosen as the uniform distribution by default.

Prior on $\mu_{kj}|T_J$

Here, we use conjugate normal distribution $N(\mu_\mu, \sigma_\mu^2)$ as the prior for $P(\mu_{kj}|T_J)$. For the hyperparameters μ_μ and σ_μ, they are chosen so that the conditional expectation $E[Y|X]$ follows a normal distribution, $N(m\mu_\mu, m\sigma_\mu^2)$, which is essentially the sum of $m\mu_{kj}$'s under the sum-of-trees model. Further, μ_{kj}s are considered apriori and are independently and identically distributed. Here, it is highly probable that $E[Y|X]$ is between the interval $(\min(Y), \max(Y))$. This is ensured by setting v such that $\min(Y) = m\mu_\mu - v\sqrt{m\sigma_\mu}$, and $\max(Y) = m\mu_\mu + v\sqrt{m\sigma_\mu}$. To facilitate posterior distribution calculations, Y is transformed as follows: $\tilde{Y} = \frac{Y - \min(Y) + \max(Y)}{2}$. This transformation results in $\tilde{Y} \in (-0.5, 0.5)$, where $\min(Y) = -0.5$ and $\max(Y) = 0.5$. This transformation allows the hyperparameter μ_μ to be set as 0 and σ_μ to be determined as $\sigma_\mu = 0.5$, where v is chosen accordingly.

For $v = 2$, the normal distribution $N(m\mu_\mu, m\sigma_\mu^2)$ assigns 95% prior probability to the interval $(\min(Y), \max(Y))$, which is the default setting. Finally, for ν and λ, the default value of ν is set to 3, while λ is chosen such that $P(\sigma^2 < s^2; \nu, \lambda) = 0.9$, where s^2 represents the estimated residual variance from a multiple linear regression model with Y as the response and X as the set of covariates.

Prior on σ

The prior for σ or $P(\sigma)$ is taken as $\sigma^2 \sim \mathrm{IG}(\nu/2, \nu\lambda/2)$, where $\mathrm{IG}(\alpha, \beta)$ represents the inverse gamma distribution with shape parameter α and rate parameter β.

Additionally, three priors are used to control how a decision tree evolves in BART. One prior governs the depth of each tree using a negative power distribution, which has three associated hyperparameters. The other prior distribution controls the selection of covariates for splitting at an internal node. By default, this is set to a uniform distribution, meaning that each covariate has an equal probability of being selected for a split at an internal node.

The first hyperparameter controls the likelihood of a node splitting, where larger values increase the chance of a split. The second regulates the number of terminal nodes (the endpoints of each tree where predictions were made), preventing BART from overfitting. The third hyperparameter manages the tree depth. Together, these priors allowed for fitting many shallow trees, or "weak learners," which are ultimately summed to form a stronger model.

If we have correlated predictors or high-dimensional predictor spaces, the assumption of a uniform distribution could lead to biased results in variable selection. To address this, we employ a more general distribution like the Dirichlet distribution as a sparsity-inducing prior, which helps guide covariate selection in the tree more effectively [172]. This is an important aspect in understanding the feature importance for our model. Finally, after the covariate is selected, another prior distribution was used to select the cut-off point in an internal node. The default suggested distribution was the uniform distribution [162]. All the hyperparameters are tuned using a 5-fold cross-validation. In this way, a fully explainable decision path could be provided to fit the model to the given prediction problem.

Posterior inference for BART

BART utilizes a full Bayesian approach for decision-making purposes. Sampling from the posterior distribution is obtained via Gibbs sampling, in combination with a Metropolis-Hastings step. This whole sampling mechanism utilizes the idea of Bayesian back-fitting procedure [162] which is a general procedure for posterior sampling from additive and generalized additive models.

The general form of the posterior distribution can be described as follows:

Step 1: The joint posterior distribution can be written as follows:

$$P\left[(T_1,M_1),\ldots,(T_m,M_m),\sigma \mid Y\right] \propto P\left(Y \mid (T_1,M_1),\ldots,(T_m,M_m),\sigma\right) P\left((T_1,M_1),\ldots,(T_m,M_m),\sigma\right) \tag{4.2.4}$$

which can be decomposed into two primary posterior distributions using Gibbs sampling. First, we successively draw

$$P\left[(T_j,M_j) \mid T_{(-j)},M_{(-j)},Y,\sigma\right] \tag{4.2.5}$$

for $j=1,\ldots,m$, where $T_{(-j)}$ and $M_{(-j)}$ represent all tree structures and terminal nodes excluding the jth tree structure and its corresponding terminal nodes.

Step 2: Next, we sample

$$P\left[\sigma \mid (T_1,M_1),\ldots,(T_m,M_m),Y\right] \tag{4.2.6}$$

from the inverse gamma distribution: $\sigma^2 \sim IG\left(\frac{\nu+n}{2},\frac{\nu\lambda+\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{m}g(X_i,T_j,M_j)\right)^2}{2}\right)$.

To generate a draw from Eq. (4.2.5), note that this distribution depends on $(T_{(-j)},M_{(-j)},Y,\sigma)$ through the residuals:

$$R_j = Y - \sum_{w\neq j} g(X,T_w,M_w), \tag{4.2.7}$$

which represent the residuals obtained after removing the contribution of the j^{th} tree from the sum-of-trees model. Consequently, Eq. (4.2.5) is equivalent to obtaining a posterior draw from a single regression tree: $R_{ij}=g(X_i,T_j,M_j)+\epsilon_i$, or more formally,

$$P[(T_j, M_j) \mid R_j, \sigma]. \tag{4.2.8}$$

Step 3: To sample from Eq. (4.2.8), we first integrate out M_j to obtain $P(T_j \mid R_j, \sigma)$, which is feasible due to the conjugate normal prior on μ_{kj}. We then use a Metropolis-Hastings (MH) algorithm to sample from $P(T_j \mid R_j, \sigma)$, where we propose a candidate tree T_j^* using a probability distribution $q(T_j, T_j^*)$. The proposed tree T_j^* is accepted with probability

$$\alpha(T_j, T_j^*) = \min\left(1, \frac{q(T_j^*, T_j)P(R_j \mid X, T_j^*, M_j)P(T_j^*)}{q(T_j, T_j^*)P(R_j \mid X, T_j, M_j)P(T_j)}\right). \tag{4.2.9}$$

Here, $q(T_j^*, T_j)$ represents the probability of transitioning from the new tree back to the previous tree, while $q(T_j, T_j^*)$ is the probability of moving from the previous tree to the new tree. The term $P(R_j \mid X, T_j^*, M_j)/P(R_j \mid X, T_j, M_j)$ is the likelihood ratio comparing the new and previous trees, and $P(T_j^*)/P(T_j)$ represents the prior ratio of the new and previous trees.

Additionally, a new tree T_j^* can be proposed from the previous tree T_j using the following four localized steps:

1. **Grow**: A terminal node is split into two new child nodes;
2. **Prune**: Two terminal child nodes under the same non-terminal node are merged, turning their parent non-terminal node into a terminal node;
3. **Swap**: The splitting criteria of two non-terminal nodes are exchanged;
4. **Change**: The splitting criterion of a single non-terminal node is modified.

Once a draw from $P(T_j \mid R_j, \sigma)$ is obtained, we subsequently draw

$$P(\mu_{kj} \mid T_j, R_j, \sigma) \sim N\left(\frac{\sigma_\mu^2 \sum_k R_{kj} + \sigma^2 \mu_\mu}{n_k \sigma_\mu^2 + \sigma^2}, \frac{\sigma^2 \sigma_\mu^2}{n_k \sigma_\mu^2 + \sigma^2}\right),$$

where R_{kj} represents the subset of residuals R_j allocated to the terminal node parameter μ_{kj}, and n_k is the number of elements in $\bar{R}_{kj}$ assigned to μ_{kj}. Next, we will provide the derivations of the posteriors in the following sections.

Posterior distributions for μ_{kj}

Let $R_{kj} = (R_{kj1}, \ldots, R_{kjn_k})^T$ be a subset of residuals R_j, where n_k is the number of residuals R_{kjh} assigned to the terminal node associated with the parameter μ_{kj}. The index h denotes individual subjects allocated to this terminal node.

We assume that

$$R_{kjh} \mid g(X_{kjh}, T_j, M_j), \sigma \sim N(\mu_{kj}, \sigma^2)$$

and

$$\mu_{kj} \mid T_j \sim N(\mu_\mu, \sigma_\mu^2).$$

Then, the posterior distribution of μ_{kj} is given by

$$P(\mu_{kj} \mid T_j, \sigma, R_j) \propto P(R_{kj} \mid T_j, \mu_{kj}, \sigma) P(\mu_{kj} \mid T_j).$$

Expanding the likelihood and prior terms, we have

$$\propto \exp\left[-\frac{\sum_h (R_{kjh} - \mu_{kj})^2}{2\sigma^2}\right] \exp\left[-\frac{(\mu_{kj} - \mu_\mu)^2}{2\sigma_\mu^2}\right].$$

Rearranging the exponent terms, we obtain

$$\propto \exp\left[-\frac{(n_k\sigma_\mu^2+\sigma^2)\mu_{kj}^2-2(\sigma_\mu^2\sum_h R_{kjh}+\sigma^2\mu_\mu)\mu_{kj}}{2\sigma^2\sigma_\mu^2}\right].$$

Thus, completing the square, the posterior distribution of μ_{kj} follows as

$$\mu_{kj} \mid T_j, \sigma, R_j \sim N\left(\frac{\sigma_\mu^2\sum_h R_{kjh}+\sigma^2\mu_\mu}{n_k\sigma_\mu^2+\sigma^2}, \frac{\sigma^2\sigma_\mu^2}{n_k\sigma_\mu^2+\sigma^2}\right).$$

Here, $\sum_h(R_{kjh}-\mu_{kj})^2$ represents the sum of squared differences between μ_{kj} and the residuals R_{kjh} allocated to the corresponding terminal node.

Posterior distributions for σ^2 or $P\left[\sigma \mid (T_1, M_1), \ldots, (T_m, M_m), Y\right]$

Let $Y = (Y_1, \ldots, Y_n)^T$, where the index i represents subjects $i = 1, \ldots, n$. Given that $\sigma^2 \sim IG(\nu/2, \nu\lambda/2)$, the posterior draw of σ^2 is obtained as follows:

$$P\left(\sigma^2 \mid (T_1, M_1), \ldots, (T_m, M_m), Y\right) \propto P\left(Y \mid (T_1, M_1), \ldots, (T_m, M_m), \sigma\right) P(\sigma^2).$$

Since

$$P(Y \mid g(X, T_j, M_j), \sigma)P(\sigma^2) = \prod_{j=1}^{m} P(Y \mid Xg(X, T_j, M_j), \sigma)P(\sigma^2),$$

we expand the likelihood and prior as

$$P(Y \mid g(X, T_j, M_j), \sigma) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(Y_i-\sum_{j=1}^{m} g(X_i, T_j, M_j))^2}{2\sigma^2}\right]$$

$$P(\sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{\nu\lambda}{2\sigma^2}\right).$$

Thus, combining terms, we obtain

$$P(\sigma^2 \mid (T_1, M_1), \ldots, (T_m, M_m), Y) \propto (\sigma^2)^{-\left(\frac{\nu+n}{2}+1\right)} \exp\left[-\frac{\nu\lambda + \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{m} g(X_i, T_j, M_j))^2}{2\sigma^2}\right].$$

Since the resulting posterior follows an inverse gamma distribution, we conclude that

$$\sigma^2 \mid (T_1, M_1), \ldots, (T_m, M_m), Y \sim IG\left(\frac{\nu+n}{2}, \frac{\nu\lambda + \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{m} g(X_i, T_j, M_j))^2}{2}\right).$$

Here, $\sum_{j=1}^{m} g(X_i, T_j, M_j)$ represents the predicted value from BART assigned to the observed outcome Y_i.

Metropolis-Hasting ratio for the grow and prune step

This section is adapted from Appendix A of Kapelner and Bleich [195]. The acceptance probability for the Metropolis-Hastings step is given by

$$\alpha(T_j, T_j^*) = \min\left\{1, \frac{q(T_j^*, T_j)P(R_j \mid X, T_j^*, M_j)P(T_j^*)}{q(T_j, T_j^*)P(R_j \mid X, T_j, M_j)P(T_j)}\right\}.$$

Here, the terms in the acceptance ratio are defined as follows: $\frac{q(T_j^*,T_j)}{q(T_j,T_j^*)}$ represents the transition probability ratio, and $\frac{P(R_j|X,T_j^*,M_j)}{P(R_j|X,T_j,M_j)}$ denotes the likelihood ratio, $\frac{P(T_j^*)}{P(T_j)}$ is the prior probability ratio of tree structures.

We now derive explicit formulas for each of these ratios under the grow and prune proposal.

Grow proposal

Grow proposal steps can be summarized into the following steps.

Transition ratio

The probability of transitioning from T_j to T_j^*, denoted by $q(T_j^*, T_j)$, represents the probability of selecting a terminal node in T_j and growing two child nodes. This can be expressed as

$$\begin{aligned} P(T_j^* \mid T_j) = P(\text{grow}) \times P(&\text{selecting a terminal node to grow from}) \\ &\times\ P(\text{selecting a covariate to split on}) \times P(\text{selecting a value to split on}). \end{aligned}$$

Substituting the individual probabilities, we obtain:

$$P(T_j^* \mid T_j) = P(\text{grow}) \times \frac{1}{b_j} \times \frac{1}{p} \times \frac{1}{\eta}.$$

Here, $P(\text{grow})$ is a user-defined probability of choosing a grow step, typically set to 0.25, b_j is the number of available terminal nodes that can be split in T_j, p is the number of variables available for splitting, and η is the number of unique values left in the chosen variable after accounting for parent node splits.

On the other hand, the probability of transitioning from T_j^* back to T_j, denoted by $q(T_j, T_j^*)$, corresponds to a pruning move, which involves selecting an internal node with exactly two terminal children and collapsing them into a single terminal node. This probability is given by

$$P(T_j \mid T_j^*) = P(\text{prune}) \times P(\text{selecting the correct internal node to prune}).$$

Substituting the relevant probabilities, we obtain

$$P(T_j \mid T_j^*) = P(\text{prune}) \times \frac{1}{w_2^*},$$

where w_2^* represents the number of internal nodes with exactly two terminal children.

Thus, the transition ratio is

$$q(T_j^*, T_j) = \frac{P(T_j^* \mid T_j)}{P(T_j \mid T_j^*)} = \frac{P(\text{prune}) \cdot b_j \cdot p \cdot \eta}{P(\text{grow}) \cdot w_2^*}.$$

If no variables have two or more unique values left, this transition ratio is set to 0.

Likelihood ratio

Since the overall tree structure remains unchanged between T_j^* and T_j, except for the terminal node where the two child nodes are introduced, we only need to focus on this specific terminal node.

Let l denote the selected terminal node in T_j, which is split into two child nodes: l_L (left child) and l_R (right child) in the grow step. Then, the likelihood ratio is given by

$$\frac{P(R_j \mid X, T_j^*, M_j)}{P(R_j \mid X, T_j, M_j)} = \frac{P(R_{l(L,1),j}, \ldots, R_{l(L,n_L),j} \mid \sigma^2) \cdot P(R_{l(R,1),j}, \ldots, R_{l(R,n_R),j} \mid \sigma^2)}{P(R_{1,j}, \ldots, R_{n,j} \mid \sigma^2)}.$$

Expanding the probability terms, we obtain

$$\frac{P(R_j \mid X, T_j^*, M_j)}{P(R_j \mid X, T_j, M_j)} = \frac{\sigma^{-2}(\sigma^2 + n_L\sigma_\mu^2)^{-\frac{1}{2}}(\sigma^2 + n_R\sigma_\mu^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n_L} R_{l(L,k),j}^2 - \frac{\left(\sum_{k=1}^{n_L} R_{l(L,k),j}\right)^2}{\sigma^2+n_L\sigma_\mu^2}\right)\right]}{\sigma^{-2}(\sigma^2 + n\sigma_\mu^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n} R_{k,j}^2 - \frac{\left(\sum_{k=1}^{n} R_{k,j}\right)^2}{\sigma^2+n\sigma_\mu^2}\right)\right]},$$

where

- n_L and n_R represent the number of observations assigned to the left and right child nodes, respectively;
- $R_{l(L,k),j}$ and $R_{l(R,k),j}$ are the observed values assigned to the left and right child nodes;

- σ^2 is the variance parameter;
- σ_μ^2 is the variance of the prior distribution of terminal node parameters.

This formulation accounts for the changes due to the grow step, focusing only on the modified terminal node.

Tree structure ratio

Since T_j can be characterized by three aspects, we define

$P_{\text{SPLIT}}(\theta)$ as the probability that a selected node θ will split, and $P_{\text{RULE}}(\theta)$ as the probability of selecting a particular variable and value for the split.

Given that $P(\theta) \propto \alpha(1+d_\theta)^{-\beta}$ and that T_j and T_j^* differ only at the child nodes, we express the ratio as

$$\frac{P(T_j^*)}{P(T_j)} = \frac{\prod_{\theta \in H^*_{\text{terminals}}}(1 - P_{\text{SPLIT}}(\theta)) \prod_{\theta \in H^*_{\text{internals}}} P_{\text{SPLIT}}(\theta) \prod_{\theta \in H^*_{\text{internals}}} P_{\text{RULE}}(\theta)}{\prod_{\theta \in H_{\text{terminals}}}(1 - P_{\text{SPLIT}}(\theta)) \prod_{\theta \in H_{\text{internals}}} P_{\text{SPLIT}}(\theta) \prod_{\theta \in H_{\text{internals}}} P_{\text{RULE}}(\theta)}.$$

Since T_j^* introduces two new terminal nodes θ_L and θ_R, while modifying θ, the expression simplifies to

$$\frac{P(T_j^*)}{P(T_j)} = \frac{[1 - P_{\text{SPLIT}}(\theta_L)][1 - P_{\text{SPLIT}}(\theta_R)]P_{\text{SPLIT}}(\theta)P_{\text{RULE}}(\theta)}{1 - P_{\text{SPLIT}}(\theta)}.$$

Substituting $P_{\text{SPLIT}}(\theta) \propto \alpha(1+d_\theta)^{-\beta}$, we obtain

$$\frac{P(T_j^*)}{P(T_j)} = \frac{(1 - \frac{\alpha}{(1+d_{\theta_L})^\beta})(1 - \frac{\alpha}{(1+d_{\theta_R})^\beta})\alpha(1+d_\theta)^{-\beta}\frac{1}{p}\frac{1}{\eta}}{1 - \frac{\alpha}{(1+d_\theta)^\beta}}.$$

Since $d_{\theta_L} = d_{\theta_R} = d_\theta + 1$, we finally obtain

$$\frac{P(T_j^*)}{P(T_j)} = \frac{\alpha(1 - \frac{\alpha}{(2+d_\theta)^\beta})^2}{[(1+d_\theta)^\beta - \alpha]p\eta}$$

Prune proposal

A prune proposal serves as the reverse operation of a grow proposal. In this step, an internal node with two terminal children is selected, and both children are removed. Consequently, the corresponding ratios will be approximately the reciprocals of those derived for the grow proposal in the previous sub-section. Additionally, prune steps are not applicable to trees that contain only a single root node [195].

BART for binary prediction problem

The BART model defined in Eq. (4.2.2) can be easily extended to a classification problem. For classification problems, a Probit model is used over the sum-of-tree-based predictions. In this set-up, the sum-of-trees model serves as an estimate of the conditional Probit of the features which can be easily transformed into the conditional probability estimate of the target class.

For binary outcomes, BART can be extended using a probit model. Specifically, the probability of an outcome $Y_i = 1$, given the predictors and tree structures, is given by

$$P(Y_i = 1 \mid X_i, (T_1, M_1), \ldots, (T_m, M_m)) = \Phi\left(\sum_{j=1}^{m} g(X_i; T_j, M_j)\right), \tag{4.2.10}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, and i indexes the subjects ($i = 1, \ldots, n$). To estimate the posterior distribution, data augmentation (Albert and Chib, 1993) can be implemented. Specifically, we first draw a latent variable $Z = \{Z_1, \ldots, Z_n\}$ as follows:

$$Z_i \sim N(-\infty, 0]\left(\sum_{j=1}^{m} g(X_i; T_j, M_j), 1\right) \quad \text{if } Y_i = 0,$$

$$Z_i \sim N(0, \infty] \left(\sum_{j=1}^{m} g(X_i; T_j, M_j), 1 \right) \quad \text{if } Y_i = 1,$$

where $N(a, b)[\mu, \sigma^2]$ denotes a normal distribution with mean μ and variance σ^2, truncated to the interval (a, b).

Next, we treat Z as a continuous outcome in a BART model:

$$Z = \sum_{j=1}^{m} g(X; T_j, M_j) + \epsilon,$$

where $\epsilon \sim N(0, 1)$ due to the probit link function. Given this setup, posterior estimation for a continuous-outcome BART model with $\sigma \equiv 1$ can be employed for a single MCMC iteration. The updated function $\sum_{j=1}^{m} g(X; T_j, M_j)$ is then used to draw a new Z, and this updated Z is subsequently used to sample another iteration of $\sum_{j=1}^{m} g(X; T_j, M_j)$. This process is repeated until convergence.

Choices of priors for Probit-based BART

In this setup, only priors for $(T_1, M_1), \ldots, (T_m, M_m)$ are required. The same decomposition as in Eq. (4.2.3) can be used without σ, and similar prior specifications for $\mu_{kj} \mid T_j$ and T_j can be applied, i.e.,

$$P((T_1, M_1), \ldots, (T_m, M_m)) = P((T_1, M_1), \ldots, (T_m, M_m))$$

$$= \left[\prod_{j=1}^{m} P(T_j, M_j) \right]$$

$$= \left[\prod_{j=1}^{m} P(M_j | T_j) P(T_j) \right]$$

$$= \prod_{j=1}^{m} \prod_{k=1}^{b_j} P(\mu_{kj}|T_j)P(T_j). \tag{4.2.11}$$

The choice for $P(T_j)$ is the same as that of the continuous outcome model in (4.2.2). Here, the hyperparameters α and β remain the same as in the continuous-outcome BART model. However, he hyperparameter settings for $P(\mu_{kj} \mid T_j)$ differ slightly from those used for continuous outcomes model in (4.2.2). The hyperparameters μ_μ and σ_μ are specified differently.

To define these hyperparameters, we set

$$\mu_\mu = 0, \quad \sigma_\mu = \frac{3}{v\sqrt{m}},$$

where choosing $v = 2$ ensures that, with approximately 95% probability, the sum $\sum_{j=1}^{m} g(X; T_j, M_j)$ falls within the range $(-3, 3)$. Since this range aligns with the probit model's natural scale, no transformation of the latent variable Z is required. Additionally, Chipman et al. [162] also recomended to use $v = 2$ as a default choice for getting better shrinkage to the μ_{kj}'s. Alternatively, the value of v may be chosen by cross-validation.

4.2.3 Feature importance using BART and consistency check

Feature importance is another important aspect of the local interpretability of tree-based machine learning (ML) models, which can be seen as building blocks for global insights [196] [197]. Global feature importance values are computed across an entire dataset (i.e., for all samples) using three primary methods [197] as follows.

Gain: Introduced by Breiman et al. in 1984 [198], gain is a traditional measure of feature importance. It quantifies the total reduction in loss or impurity resulting from all splits involving a given feature. Although its foundation is largely heuristic, gain remains widely used as the basis for various feature selection methods [199], [200];

Split Count: This method determines feature importance by counting the number of times a feature is used for splitting. Since splits are selected based on their informativeness, a higher split count indicates a more influential feature;

Permutation: This approach evaluates feature importance by randomly permuting a feature's values in the test set and measuring the resulting change in model error. If a feature is crucial, its permutation should significantly increase the model's error. Different variations of this method exist, depending on how the feature values are permuted [201], [202], [203].

BART can be easily used for assessing feature importance across the entire dataset using the last two approaches mentioned above, i.e., split count and permutation-based tests. To select features or variables, BART uses variable inclusion proportions (VIP): the proportion of times each predictor is chosen as a splitting rule divided by the total number of splitting rules appearing in the model.

Permutation-based tests due to Bleich et al. [171]

Bleich et al. [171] developed a permutation-based framework for feature importance using the previously defined VIP scores. To determine how large a variable inclusion frequency p_k or VIP score is required for selecting a predictor variable x_k, they established appropriate selection thresholds. This was achieved through a permutation-based approach, which is used to generate a null distribution for the variable inclusion proportions $p = (p_1, p_2, \ldots, p_K)$.

In this method, we generate P permuted versions of the response vector: $y_1^*, y_2^*, \ldots, y_P^*$. For each permuted response vector y_p^*, we fit the BART model using y_p^* as the response while keeping the original predictor variables $x_1, \ldots, x_K$ unchanged. This permutation approach preserves potential dependencies among the predictor variables while eliminating any association between the predictors and the response variable.

From each BART run with a permuted response y_p^*, we retain the estimated variable inclusion proportions. Let $p^*_{k,p}$ denote the variable inclusion proportion for predictor x_k obtained from the pth permuted response. We further define p^*_p as the vector of all variable inclusion proportions from the pth permutation. The collection of variable inclusion proportions across all P permutations, $p^*_1, p^*_2, \ldots, p^*_P$, serves as the null distribution for the variable inclusion proportions derived from the actual (unpermuted) response y.

The remaining challenge is to establish an appropriate selection threshold for each predictor x_k based on the permutation-based null distribution $p^*_1, p^*_2, \ldots, p^*_P$. To address this, three different thresholding strategies were employed by [171], each varying in the strictness of the resulting variable selection procedure. Let us now explore the three thresholding schemes for determining important features:

1. **Local threshold:** To determine a selection threshold for each variable inclusion proportion p_k corresponding to predictor x_k, the calculation relies solely on the permutation-based null distribution of p_k. Specifically, the $(1 - \alpha)$ quantile of the distribution $p^*_{k,1}, p^*_{k,2}, \ldots, p^*_{k,P}$ is computed, and a predictor x_k is selected only if p_k exceeds this quantile;

2. **Global max threshold:** To establish a selection threshold for the variable inclusion proportion p_k of predictor x_k, the calculation is performed based on the maximum across the permutation distributions of the variable inclusion proportions for all predictor variables. For each permutation p, the maximum variable inclusion proportion across all predictor variables is first calculated as $p^*_{\max,p} = \max\{p^*_{1,p}, p^*_{2,p}, \ldots, p^*_{K,p}\}$. Next, the $(1 - \alpha)$ quantile of the distribution $p^*_{\max,1}, p^*_{\max,2}, \ldots, p^*_{\max,P}$ is determined and select predictor x_k only if p_k exceeds this $(1 - \alpha)$ quantile;

3. **Global SE threshold:** A predictor x_k is selected if its VIP score p_k exceeds a threshold based on the mean and standard deviation of its null distribution, using a global multiplier shared by all predictors. Let m_k and s_k be the mean and standard

deviation of variables inclusion proportion p^*_k for predictor x_k across all permutations. Then, the following is calculated:

$$C^* = \inf_{C \in \mathbb{R}^+} \left\{ \forall k : \frac{1}{P} \sum_{p=1}^{P} \mathbf{1}\{p^*_{k,p} \leq m_k + C \cdot s_k\} > 1 - \alpha \right\} \tag{4.2.12}$$

The value C^* is the smallest global multiplier that ensures simultaneous $1-\alpha$ coverage across the permutation distributions of p_k for all predictor variables. The predictor x_k is selected only if $p_k > m_k + C^* \cdot s_k$. This third strategy represents a compromise between the local permutation distribution for variable k, which incorporates the mean m_k and standard deviation s_k, and the global permutation distributions of the other predictor variables, through C^*.

Dealing with the correlated predictor space: Sparsity-induced prior-based scheme

One critique of these methods is that the optimal procedure depends on the underlying sparsity of the problem, which is often unknown. We also incorporate a sparsity-induced prior-based scheme to BART to assess feature importance. Given that the human gait data consists of highly correlated predictors, there is a significant risk of the Markov chain becoming trapped in a posterior mode [172]. We adopt the sparsity-inducing Dirichlet hyperprior on the splitting proportions as framed by Linearo et al. [172].

This approach had already been introduced in the existing BART literature to mitigate the effects of highly correlated nuisance predictors.

Sparsity-inducing Dirichlet prior and Variable selection [172]

As mentioned, the probability that a node at depth $d = 0, 1, ...$ will split is $\frac{\alpha}{(1+d)^\beta}$. Here, to each internal node, a splitting rule of the form $[X_i < C]$ is assigned. Each X associated with this internal node is then directed to one of its children depending on whether it satisfies the splitting rule. The predictor used for constructing a splitting rule is selected according to the probability vector $s = (s_1, \dots, s_p)$. There are multiple possible distributions for C given that predictor i is chosen for the splitting rule. Most BART implementations adopt a data-dependent prior as proposed by Chipman, George, and McCulloch [162]. A splitting rule is considered trivial if it contradicts a higher-level splitting rule in the tree.

Assumption 1. Given that predictor i is selected, the value C is drawn uniformly from the set of observed values $X_{1i}, \dots, X_{ni}$ that result in nontrivial splitting rules. If no such rule exists, a new predictor is selected based on s, and the process is repeated. The node is designated as terminal if constructing a nontrivial splitting rule is not feasible.

This is a standard and traditional assumption used for variable splitting and subsequent variable selection in BART. Additionally, this mechanism, like Shapley values, allocates credit uniformly among all features, thereby avoiding inconsistency problems. Linero et al. [172] slightly modified this assumption to greatly simplify the analytic properties of the prior for $P(T_j)$;

Assumption 2. Given that predictor i is selected, the value C is drawn uniformly from the set of observed values $X_{1i}, \dots, X_{ni}$ that result in nontrivial splitting rules. If no such rule exists, construct a split on predictor i by drawing C uniformly from $X_{1i}, \dots, X_{ni}$.

These two assumptions differ only in how they handle situations where no further splitting is possible on a selected predictor. Since trees constructed under $\frac{\alpha}{(1+d)^\beta}$ with typical values of (α, β) tend to be shallow, the distinction between these assumptions becomes significant primarily when certain predictors have only a limited number of unique sample values.

Now, let the variable selection probabilities be denoted by s_j for $j = 1, \dots, P$. Instead of using a uniform variable selection prior in BART, a Dirichlet prior is introduced. A beta prior is placed on the parameter θ, as follows:

$$[s_1, \ldots, s_P] \mid \theta \sim \text{Dirichlet}\left(\frac{\theta}{P}, \ldots, \frac{\theta}{P}\right),$$

$$\frac{\theta}{\theta + \rho} \sim \text{Beta}(a, b). \tag{4.2.13}$$

This allows for the data to determine an appropriate degree of sparsity. When $a = b = 1$, this corresponds to the prior density $\frac{\rho}{(\theta+\rho)^2}$ for θ, which exhibits Cauchy-like tails with a median of ρ. The heavy tails in this distribution allow for large values of θ, enabling the prior to revert to the standard BART prior when $f_0(x)$ is not sparse. The values considered are $b = 1$ and $a \in \{0.5, 1\}$, where $a = 0.5$ provides additional preference for sparsity in the prior. In our analysis, we have considered the default chosen values: $a = 0.5$, $b = 1$, and $\rho = P$. If additional sparsity is required, the argument ρ can be set to a value smaller than P. This may be more appropriate when there is strong prior knowledge suggesting that f_0 is sparse.

Another approach involves treating θ as a tuning parameter and selecting its value through cross-validation. This method is effective and circumvents the complexities of prior specification. However, its primary drawback is the increased computational cost associated with executing cross-validation [172].

We have employed the Dirichlet prior-based BART for getting stable variable inclusion proportions (VIP) for each classification cohorts. Additionally, a Monte Carlo study is conducted to examine the uncertainties in the VIP scores of the selected predictors (see Figure 4.8, 4.9, and 4.10). The Monte Carlo study reveals a visual summary of the predictors consistently showing higher VIP scores across multiple iterations. This allows us to compare the two approaches' feature importance based on their VIP scores.

4.2.4 Prediction explainability of BART through the individual effect plots

One of the most important aspects of interpretable machine learning (ML) models is to understand and visualize the individual effects of the predictors on the prediction function. Partial Dependence Plots (PDPs) [204] are the most common approach for visualizing the individual predictor's impact on the prediction function (i.e., on the predicted classes). However, PDPs are often impacted by collinearity among predictors, which can lead to biased interpretations [174]. Here, Accumulated Local Effects (ALE) plots [174] are used to understand the individual effects of the predictors on the predicted classes due to BART. We focus on the variables identified as important (as mentioned in the previous section) and examine their biological interpretations with the predicted classes, as inferred from the ALE plots.

Friedman's Partial Dependence Plots (PDPs) [204]

Viewing higher-dimensional functions presents a greater challenge. Therefore, it is beneficial to examine the partial dependence of the approximation of the prediction function on selected small subsets of input variables. To be more specific, consider a scenario where a supervised learning model has been trained to approximate the conditional expectation $E[Y \mid X = x] \approx f(x)$. Here, Y represents a scalar response variable, while $X = (X_1, X_2, \ldots, X_d)$ is a vector consisting of d predictor variables. The function $f(\cdot)$ denotes the fitted model, which is used to predict Y or, in the case of classification, the probability that Y belongs to a particular class as a function of X. The training data used to fit the model consists of n observations, each comprising $(d + 1)$ variables: $\{(y_i, x_i) \mid x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d}), \quad i = 1, 2, \ldots, n\}$.. Here, y_i represents the response variable, while x_i is a vector of d predictor variables corresponding to the i^{th} observation.

As noted earlier, our goal is to visualize and interpret the “individual” or “main effects” of the prediction function $f(x) = f(x_1, x_2, \ldots, x_d)$ for each predictor, along with the lower-order “interaction” effects between specific pairs of predictors. To study the influence of a single predictor or a small subset of predictors, say X_S (with at most two predictors considered in X_S), on the predicted response, the PD function is defined as:

$$\hat{f}_{S,\mathrm{PD}}(x_S) \equiv \mathbb{E}[f(x_S, X_C)] = \int p_{X_C}(x_C) f(x_S, x_C)\, dx_C, \tag{4.2.14}$$

where $p_{X_C}(\cdot)$ represents the marginal distribution of X_C. Here, X_C are the other features used in the machine learning model $f(.)$. The feature(s) in S are those for which we want to know the effect on the prediction. The feature vectors X_S and X_C combined make up the total feature space X.

A pointwise estimate of Eq. (4.2.14), computed for different values of X_S, is given by

$$\hat{f}_{S,\mathrm{PD}}(x_S) \equiv \frac{1}{n} \sum_{i=1}^{n} f(x_S, x_{i,C}). \tag{4.2.15}$$

Partial dependence functions provide valuable insights for interpreting models produced by black-box prediction methods such as neural networks, support vector machines, nearest neighbors, and radial basis functions. When the number of predictor variables is large, these functions offer an effective means to analyze and understand model behavior.

Remark 1: The closer the dependence of $f(x)$ on the predictors in X_S is to being additive or multiplicative, the more completely the partial dependence function $\hat{f}_{S,\mathrm{PD}}(x_S)$ (Eq. 4.2.14) captures the nature of the influence of the variables in X_S on the derived approximation $\hat{f}(x)$.

Remark 2: For regression trees that utilize single-variable splits, the partial dependence of $f(x)$ on a specified target variable subset X_S can be directly evaluated using only the tree structure, without requiring reference to the original training data. Given a specific set of values for the variables X_S, the evaluation proceeds through a weighted traversal of the tree. At the root node, an initial weight of 1 is assigned. For each non-terminal node encountered during traversal:

- If the split variable belongs to the target subset X_S, the traversal continues to the appropriate left or right daughter node without modifying the weight;
- If the split variable is part of the complement subset X_C, both daughter nodes are visited, and the current weight is multiplied by the fraction of training observations that proceeded left or right at that node.

Each terminal node visited during this process retains the current weight. Once the tree traversal is complete, the partial dependence function $\hat{f}_{S,\mathrm{PD}}(x_S)$ is obtained as the weighted average of the $\hat{f}(x)$ values across the terminal nodes that were visited. For classification tasks where the machine learning model produces probability outputs, the partial dependence plot illustrates the probability of a specific class as a function of different values of the feature(s) in S. For an ensemble of M regression trees, such as those generated through boosting, the final partial dependence estimate is obtained by averaging the results from individual trees. Same idea will be used while interpreting the individual effects of the predictors on the objective function, in BART.

Remark 3: The assumption of independence is the biggest issue for the PD plots. It is assumed that the feature(s) used to compute partial dependence are not correlated with other features. In the integral of Eq. (4.2.14), the weighted average of $f(x_S, X_C)$ as X_C varies over its marginal distribution. So, this requires severe extrapolation beyond the training data. If a simple parametric model of the correct form were fitted, then the extrapolation might be reliable. However, due to its inherent flexibility, a non-parametric supervised learning model, such as a regression tree, cannot be expected to extrapolate reliably. [174] demonstrated that this renders PD plot an unreliable indicator of the effect of variable of interest.

Accumulated Local Effect (ALE) Plots: An improvement over the PDPs [174]

To estimate local effects, we divide the feature into many intervals and compute the differences in the predictions. Accumulated Local Effect (ALE) plots are more reliable than PDPs for explaining the individual effects of features on model predictions, especially in a highly correlated feature space. Instead of looking at the whole picture (like the usual PDPs), ALE breaks it down step by step. In this way, ALEs isolate the effect of the feature of interest and block the effect of correlated features. ALE plots also enable us to look at the interaction effects second order) of two predictors on the predicted classes without accounting for the highly correlated nature of the predictors.

Accumulated Local Effects (ALE) plots average the changes in predictions and accumulate them over a grid. The ALE function for a subset of features S is defined as

$$\hat{f}_{S,\mathrm{ALE}}(x_S) = \int_{z_{0,S}}^{x_S} \mathbb{E}_{X_C|X_S=z_S}\left[\frac{\partial \hat{f}_S(X_S, X_C)}{\partial X_S}\Big| X_S = z_S\right] dz_S - \text{constant}. \qquad (4.2.16)$$

which can be rewritten as

$$\hat{f}_{S,\mathrm{ALE}}(x_S) = \int_{z_{0,S}}^{x_S} \left(\int_{X_C} \frac{\partial \hat{f}_S(z_S, X_C)}{\partial X_S} dP(X_C|X_S = z_S)\right) dz_S - \text{constant}. \qquad (4.2.17)$$

Here, $z_{0,j}$ is an approximate lower bound of X_S. The 'constant' is chosen such that $f_{S,\mathrm{ALE}}(X_S)$ has a mean of zero with respect to the marginal distribution of X_S. An Accumulated Local Effects (ALE) plot of the main effect of x_S is a plot of an estimate of $f_{S,\mathrm{ALE}}(x_S)$ versus x_S. Visually, the main effect depends on $f(\cdot)$ as a function of x_S.

As discussed earlier, ALE examines small changes in a feature and how they affect the prediction. Then, it pieces all these small effects together to get the overall impact of that feature. Thus, the estimate of the ALE main effect is obtained by replacing the integral in Eq. (4.2.16) with a summation and the derivative with a finite difference, i.e.,

$$\hat{f}_{S,ALE}(x_S) = \sum_{k_S(x)} \frac{1}{n_S(k)} \sum_{i:x_S^{(i)} \in N_j(k)} \left[\hat{f}(z_{k,S}, x_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, x_{-S}^{(i)})\right] - \hat{contsant}. \quad (4.2.18)$$

The $\hat{constant}$ is chosen so that $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{S,ALE}(x_S^{(i)}) = 0$.

4.3 Results

4.3.1 Demographic characteristics and cohort description

The IOPL dataset [142] contains three pathology groups: Healthy, Orthopedic, and Neurological. Fifty-two subjects are classified as the Healthy cohort, who underwent 242 predescribed gait trials. The average age is 36.4 (SD, 20.6), with 67.3% identified as male. This group's average BMI is 23.6 (SD, 3.9). 96.2% of the cohort has right laterality, while the remaining participants have left laterality. Healthy subjects have no known medical impairment.

53 subjects were allocated to the Orthopedic cohort, with an average age of 60.1 (SD, 19.3). These subjects underwent a total of 243 pre-defined gait trials. 49% of the population are identified as male. The average BMI is noted at 27.1 (SD, 5.6). Right laterality is observed in 94.3% of the cohort, with the remaining participants exhibiting left laterality. The orthopedic group is also composed of two cohorts of distinct pathologies: lower limb osteoarthrosis and cruciate ligament injury. None of the quantitative analysis is provided

for these cohorts in the dataset.

Finally, the dataset include a cohort of 125 subjects with neurological disorders. The average age for this pathology group is reported as 61.5 (SD, 13.2), with 64% of the population identified as male. The average BMI is reported at 25 (SD, 4.4). Right laterality is present in 85.6% of the cohort, left laterality in 3.2%, ambidexterity in 0.8%, and 10.4% is not reported. This group is composed of 4 cohorts: hemispheric stroke, Parkinson's disease, toxic peripheral neuropathy, and radiation-induced leukoencephalopathy. No demographic characteristics are available for these smaller cohorts in the dataset. Differences in demographic variables among the Healthy, Orthopedic and Neurological groups were compared using the analysis of variance (ANOVA) or Kruskal-Wallis test for parametric and non-parametric tests, respectively. All hypotheses were non-directional, and a p-value of < 0.05 indicated a statistically significant difference. The difference analysis reveals statistically significant differences in demographic characteristics (age, height, and BMI) between the healthy and pathology groups (See Table 4.2).

4.3.2 Model Diagnostics

We have primarily used 'bartMachine' package in R for all the BART-related computations. The package uses Java and is integrated into R via 'rJava. It supports multi-threading, which speeds up the computation. During model creation, parallelization was implemented by generating one independent Gibbs chain per core. With the default setting of 250 burn-in samples and 1,000 post-burn-in samples, utilizing four cores results in each core sampling 500 times—250 for burn-in and 250 for post-burn-in. The final model aggregates the four post-burn-in chains from all cores, yielding a total of 1,000 post-burn-in samples. While this approach effectively runs the burn-in phase sequentially, making it susceptible to Amdahl's Law, it also reduces the autocorrelation of the sum-of-trees samples in the posterior distribution, potentially enhancing predictive performance due to the independence of the chains [195].

We monitored the convergence diagnostic plots for our three models, providing insights into model stability across the MCMC iterations. For each model (for the three classification cohorts), the convergence diagnostic plots consist of three panels: (a) Percent Acceptance by MCMC Iteration (Top-Left): This plot shows the percent acceptance of Metropolis-Hastings (MH) proposals across the trees where each point plots one MCMC iteration. The black line represents the average acceptance rate, which stabilizes around 0.6 after the burn-in period (indicated by the shaded region). A stable acceptance rate after burn-in suggests that the sampler is moving consistently through the parameter space, indicating convergence. Each computing core is coloured differently; (b) Tree Number of Nodes and Leaves by MCMC After Burn-in (Top-Right): This panel shows the number of nodes and leaves for each tree across the MCMC iterations after burn-in. The blue line represents the mean number of nodes and leaves for all trees, while the black lines display the spread across iterations. The relatively stable blue line indicates that the average tree complexity remains consistent, suggesting that the model has reached equilibrium in terms of tree structure, which is an indicator of convergence. Computing cores are separated by vertical gray lines; (c) Tree Depth by MCMC Iteration after burn-in (Bottom): This panel tracks the depth of the trees across MCMC iterations after burn-in. The blue line again represents the average tree depth, while the black lines capture the variation in tree depths across iterations. The consistency in mean tree depth after the burn-in period implies that the tree growth process is stable, further supporting convergence. We can observe the stability of our models across all the classification cohorts, and computing cores are separated by vertical gray lines.

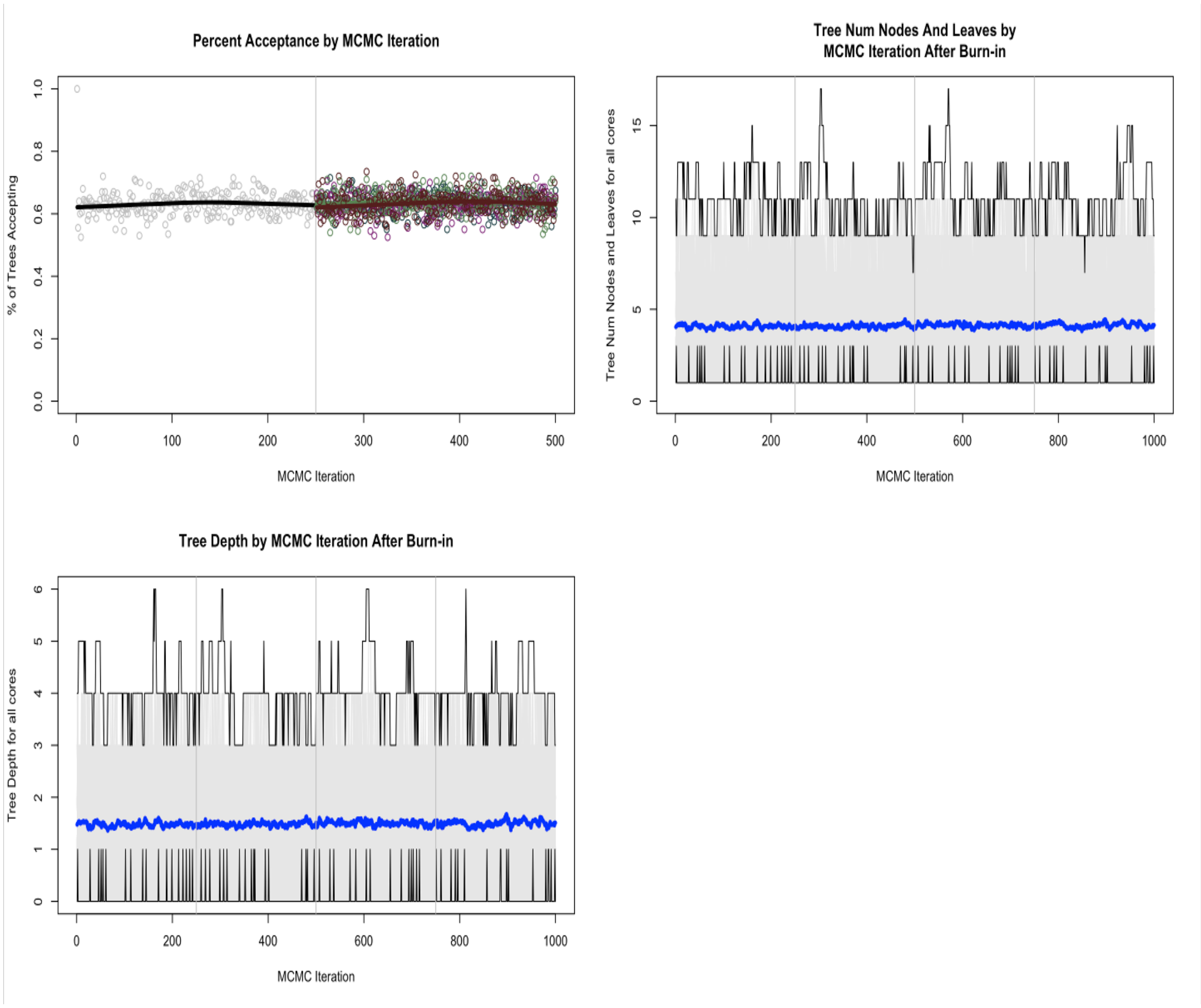


Figure 4.2: Diagnostic check for the cross-validated BART model deployed for the classification cohort based on the Healthy vs. Neurological comparison.

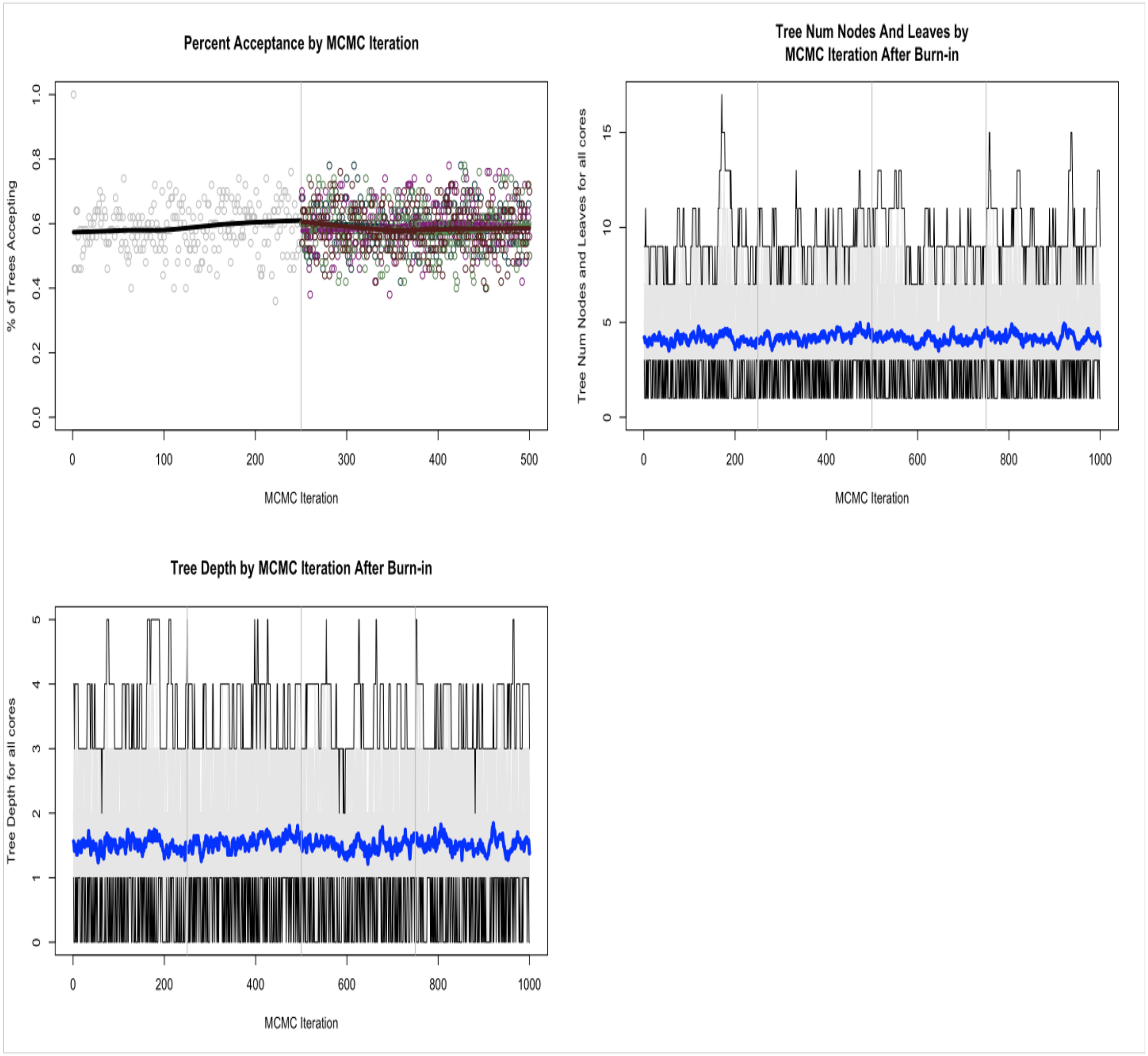
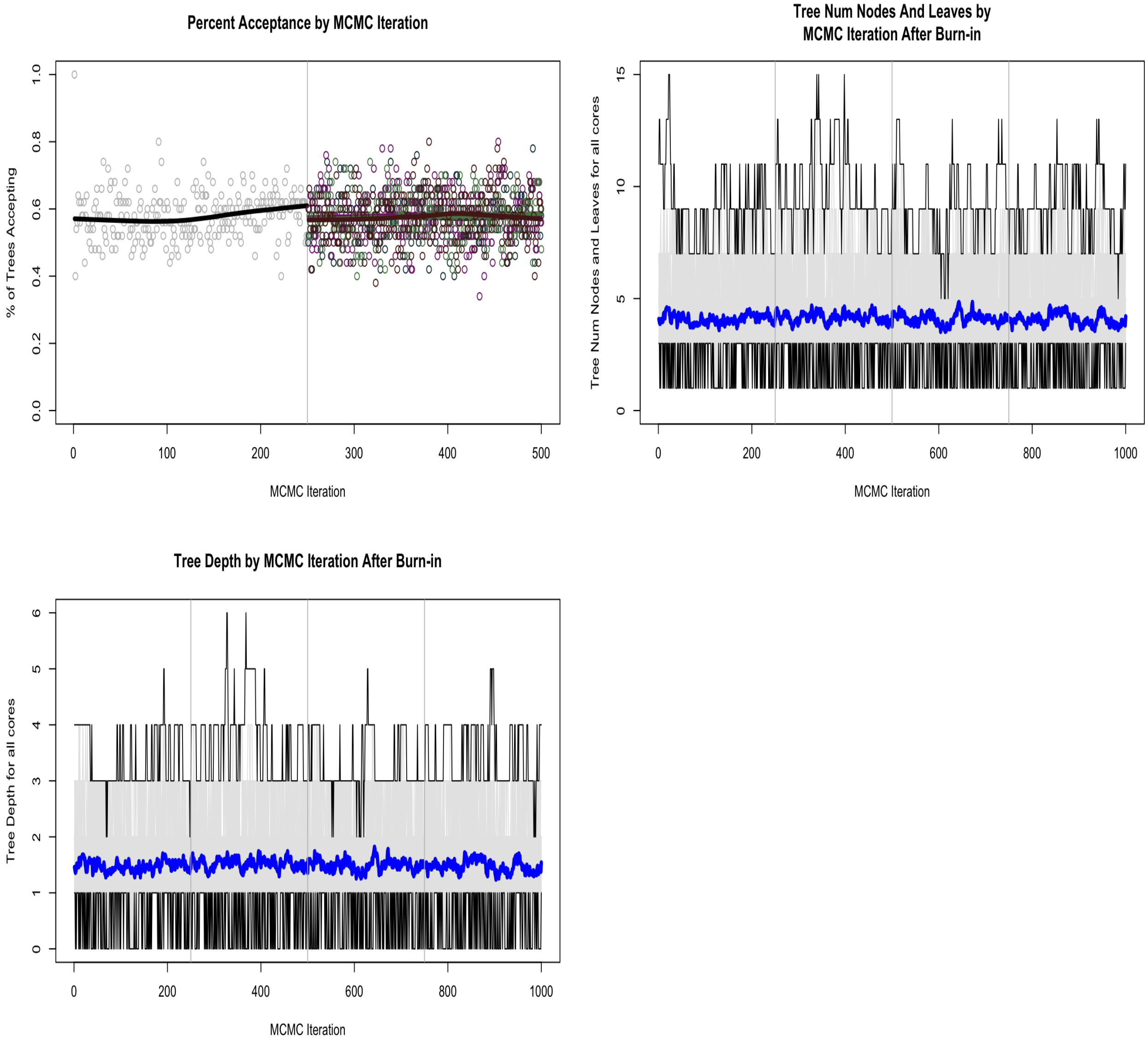


Figure 4.3: Diagnostic check for the cross-validated BART model deployed for the classification cohort based on the Healthy vs. Orthopedic comparison.



Percent Acceptance by MCMC Iteration

Tree Num Nodes And Leaves by MCMC Iteration After Burn-in

Tree Depth by MCMC Iteration After Burn-in

Figure 4.4: Diagnostic check for the cross-validated BART model deployed for the classification cohort based on the Orthopedic vs. Neurological comparison.

4.3.3 Explainability of BART

The current trend in gait analysis is mostly limited to the predictive performance of the model. However, the explanation of the model predictions is particularly important in medical applications, where the patterns a model uncovers can be more important than the model's predictive performance. We investigate BART's in-built feature importance based on permutation tests and sparsity-induced priors. Variable importance is investigated based on the variable inclusion proportions (VIP), i.e., how many times a particular feature is used in the splitting process (i.e., the split counts and permutation-based methods). Apart from investigating only the global interpretations of BART, we also investigate the local interpretations of the model by looking at the effects of important input features on individual prediction cohorts.

We investigate the permutation-based test based on three thresholds as mentioned in Subsection 4.2.3. The sparsity-induced prior-based method is also applied to get stable variable inclusion proportions as mentioned in Subsection 4.2.3. A Monte Carlo simulation is performed to examine the uncertainty of variable inclusion proportions for sparsity-induced BART. We examine multiple methods to evaluate the consistency and impact of individual features on the prediction function, along with their biological interpretation. In this way, we can understand the individual effects of the features responsible for distinguishing healthy subjects from pathological groups. To understand and validate the consistency among these in-built feature-importance procedures, a traditional SHAP analysis [193] is performed based on the bench-marked classes of ML algorithms in Chapter 5.

Feature importance and individual effects on the prediction function for the Healthy vs. Neurological disorder cohort

For the Healthy vs Neurological cohort, permutation-based procedures listed six features, among which the "Average mean load" received the highest ranking regarding its VIP score and came out as important for all three thresholds. “Right foot CV load,” “Average stride length,” “Average walking speed,” “Average step length,” and “Right foot CV stride length” came as important based on the “local” procedure. “Right foot CV load,” “Average stride length,” and “Right foot CV stride length” came as important features based on the “Global SE” threshold.

Sparsity-induced prior-based BART also ranks the “Average mean load Phase” as the most important feature among the six. “Average mean load,” “Left foot CV stance,” and “Left foot CV stride time” came as the important features. “Average mean load” ranked much higher in terms of its variation inclusion proportion score (VIP) when compared with the other features. The uncertainty analysis of the VIP values based on a 1000 Monte Carlo runs is given in the Fig 4.8.

After evaluating the individual average effects of the Average mean load Phase on the prediction function (with the healthy group taken as the target class), a non-linear relationship is observed, as shown in Fig 4.11 (a). Also, it is observed that there is a high positive effect for predicting the Healthy group (targeted class) for a smaller average mean load and a high negative impact for the larger average mean load. Subjects with higher loads in their gait cycle are more likely to be classified within the Neurological disorder cohort. Additionally, regarding biological importance, Average Walking Speed and Average Stride length are two important variables captured by the permutation-based procedure. Fig 4.11 (b) reveals that subjects with slower walking speed have a relatively high negative impact as the predictor of the healthy group. These subjects are more likely to be classified within the Neurological disorder cohort. Conversely, a strong positive effect is observed in subjects with higher walking speeds when predicting the healthy group. This indicates that BART assigns a higher probability to the Healthy class for subjects with higher average walking speeds. The same trend is observed for Average walking speed and the feature Average stride length. Some of the previous studies have biologically validated these trends in prediction (please see Discussion section).

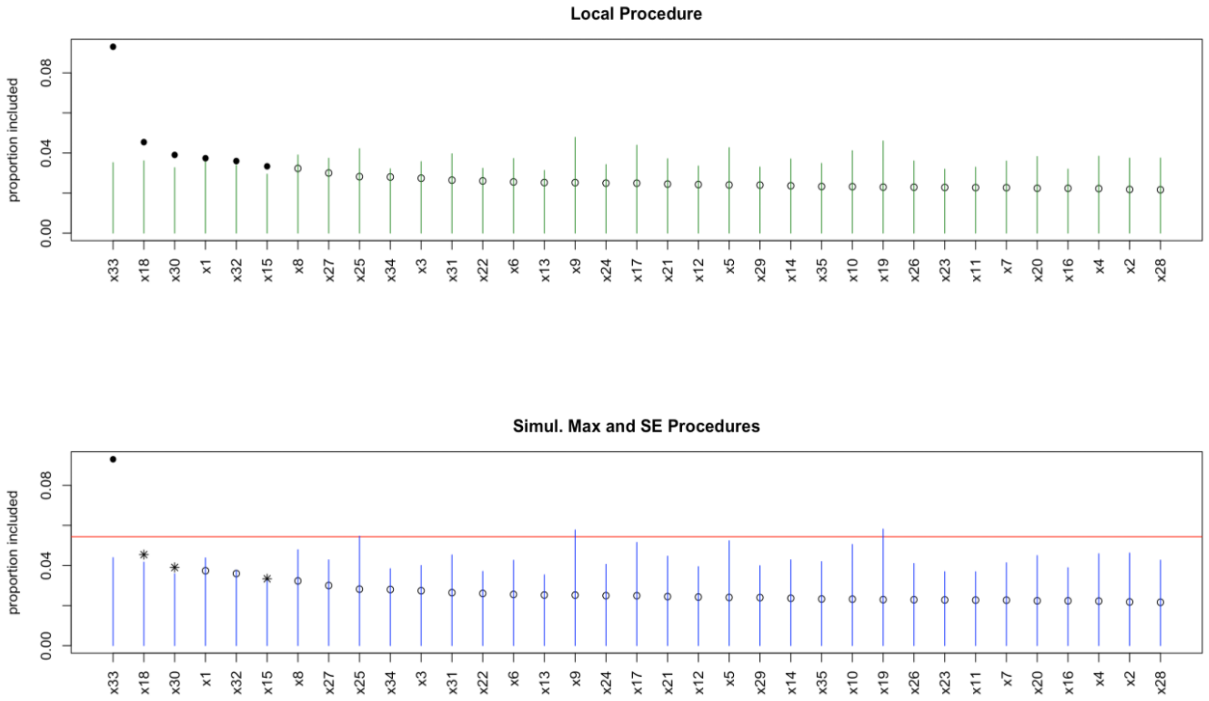


Figure 4.5: Visualization of permutation-based variable importance using three thresholds for the classification cohort based on the Healthy vs Neurological comparison. The top plot shows the "Local" procedure, where green lines mark threshold levels from permutation distributions. Solid dots indicate variables included (observed value exceeds the threshold), while open dots show variables not included. The bottom plot displays "Global SE" and "Global Max" thresholds: red lines mark the "Global Max" cut-off (solid dots for included variables), and blue lines represent the "Global SE" threshold (asterisks for variables that exceed only this threshold). Open dots exceed neither threshold.

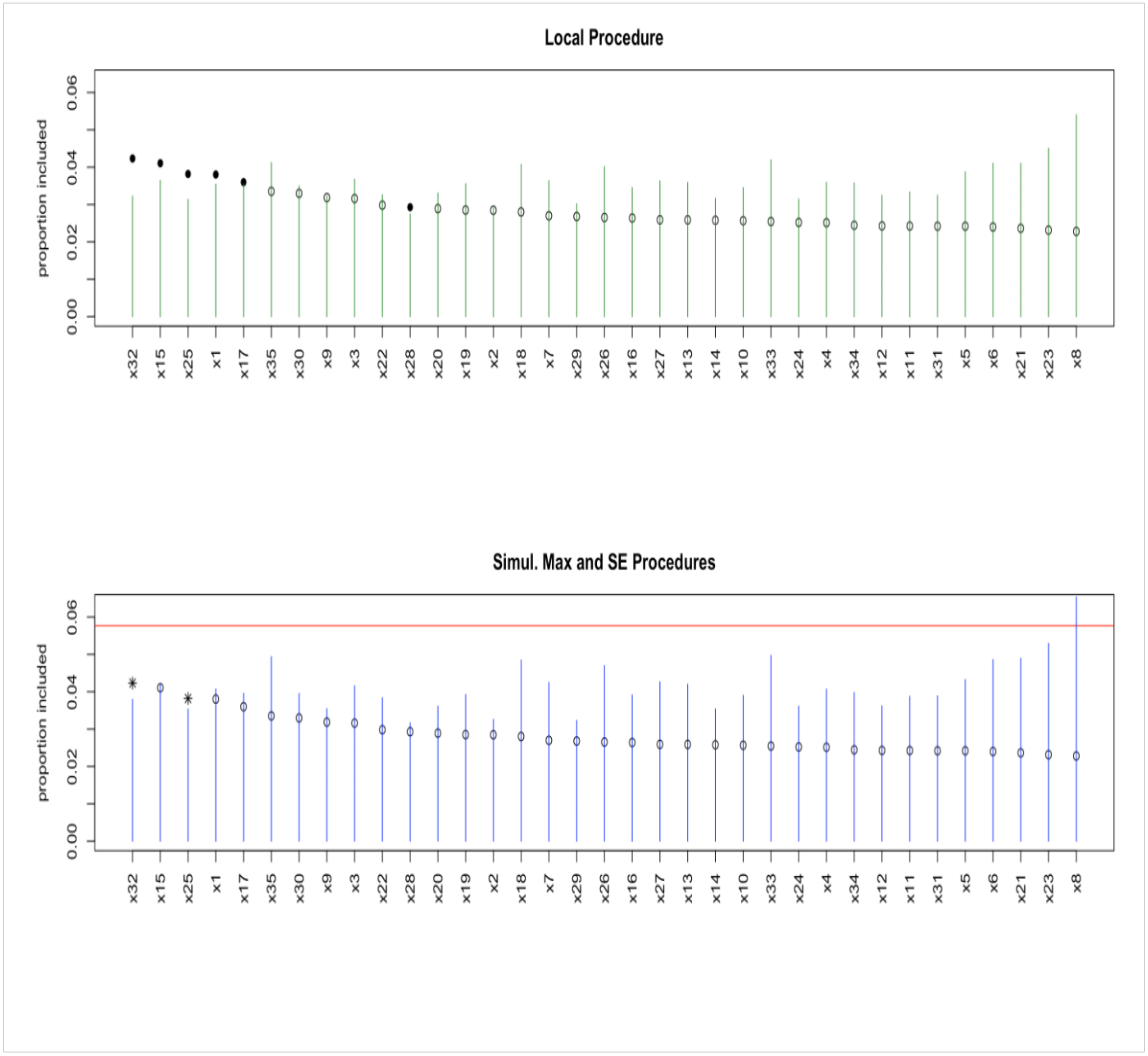


Figure 4.6: Visualization of permutation-based variable importance using three thresholds for the classification cohort based on the Healthy vs Orthopedic comparison. The top plot shows the “Local” procedure, where green lines mark threshold levels from permutation distributions. Solid dots indicate variables included (observed value exceeds the threshold), while open dots show variables not included. The bottom plot displays “Global SE” and “Global Max” thresholds: red lines mark the “Global Max” cut-off (solid dots for included variables), and blue lines represent the “Global SE” threshold (asterisks for variables that exceed only this threshold). Open dots exceed neither threshold.

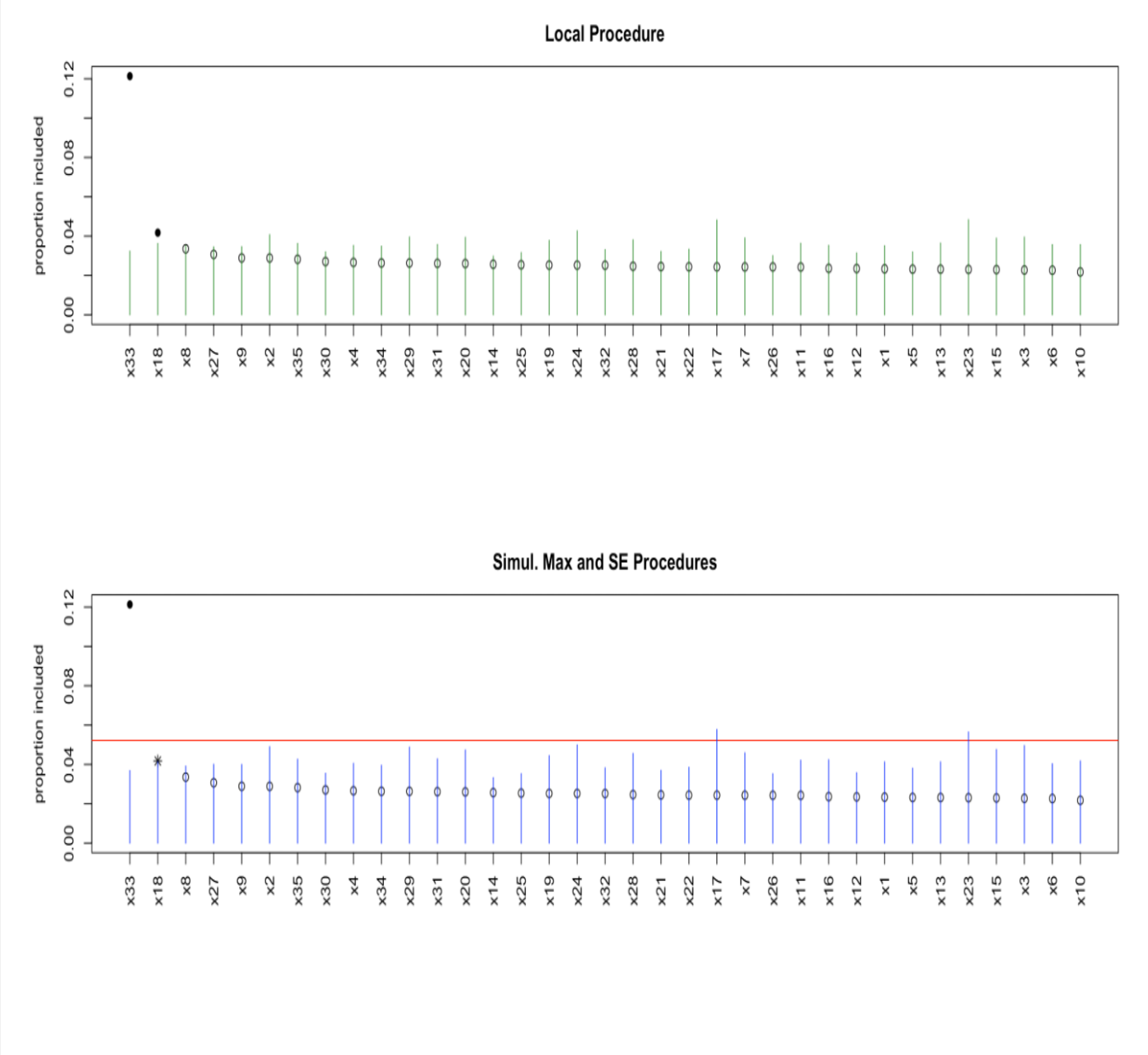


Figure 4.7: Visualization of permutation-based variable importance using three thresholds for the classification cohort based on the Orthopedic vs. Neurological comparison. The top plot shows the “Local” procedure, where green lines mark threshold levels from permutation distributions. Solid dots indicate variables included (observed value exceeds the threshold), while open dots show variables not included. The bottom plot displays “Global SE” and “Global Max” thresholds: red lines mark the “Global Max” cut-off (solid dots for included variables), and blue lines represent the ”Global SE” threshold (asterisks for variables that exceed only this threshold). Open dots exceed neither threshold.

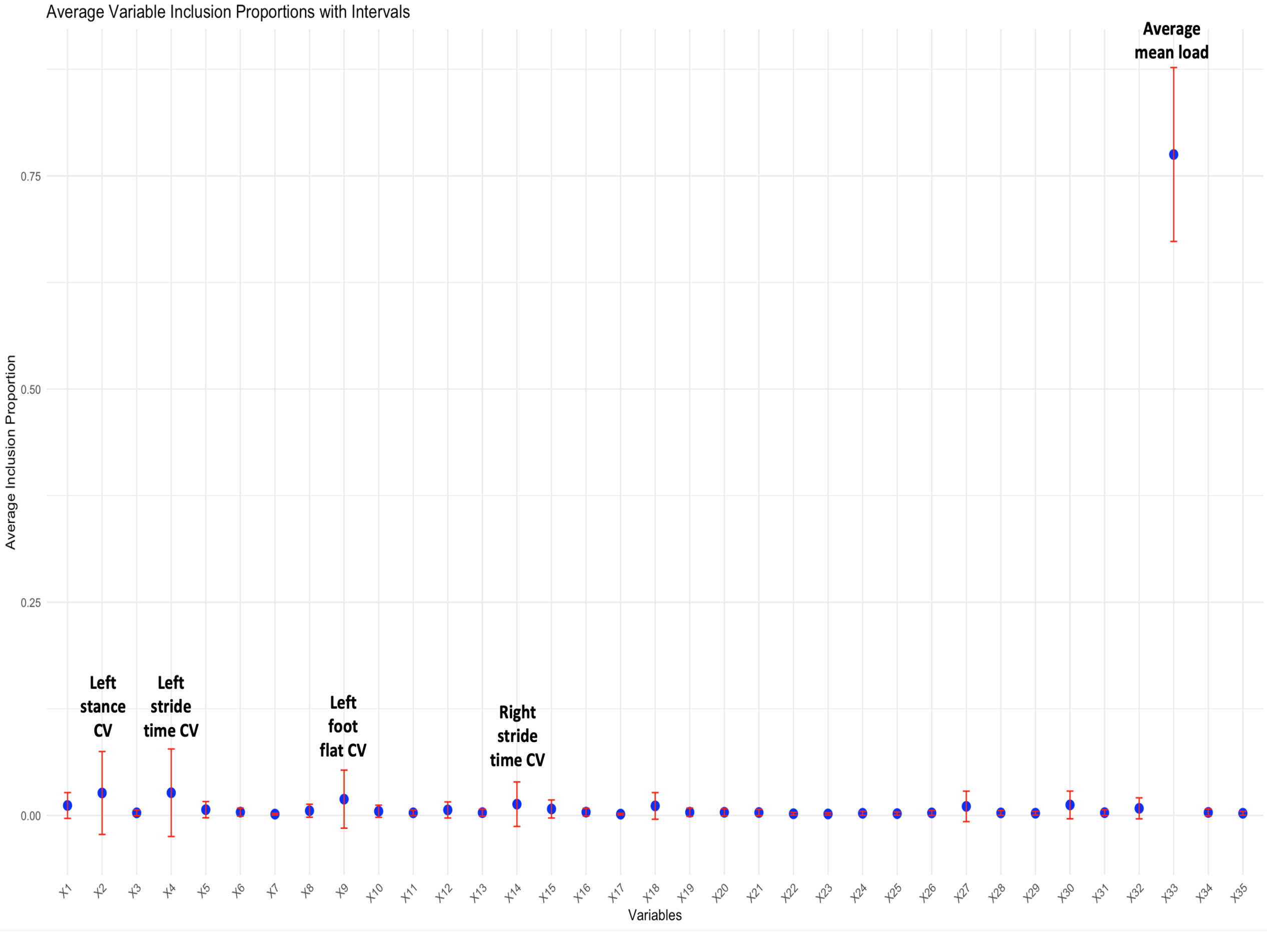


Figure 4.8: Visualization of uncertainty from an MCMC study on average variable inclusion proportions (VIP) for various gait parameters, using a sparsity-induced prior-based BART model for the classification cohort based on the Healthy vs Neurological comparison. We investigated each gait parameter's variable inclusion proportions (VIP) using a sparsity-induced prior-based BART model to understand feature importance across the three cohorts. We carried out an MCMC simulation study with 1000 samples to understand the uncertainty in the VIP scores for each run. This approach allowed us to assess the stability of feature importance resulting from the prior-modified BART model. The uncertainty in the bounds reflects variation from the Markov chains used in each iteration. We highlighted the important features due to the sparsity-induced prior-based BART for each cohort.

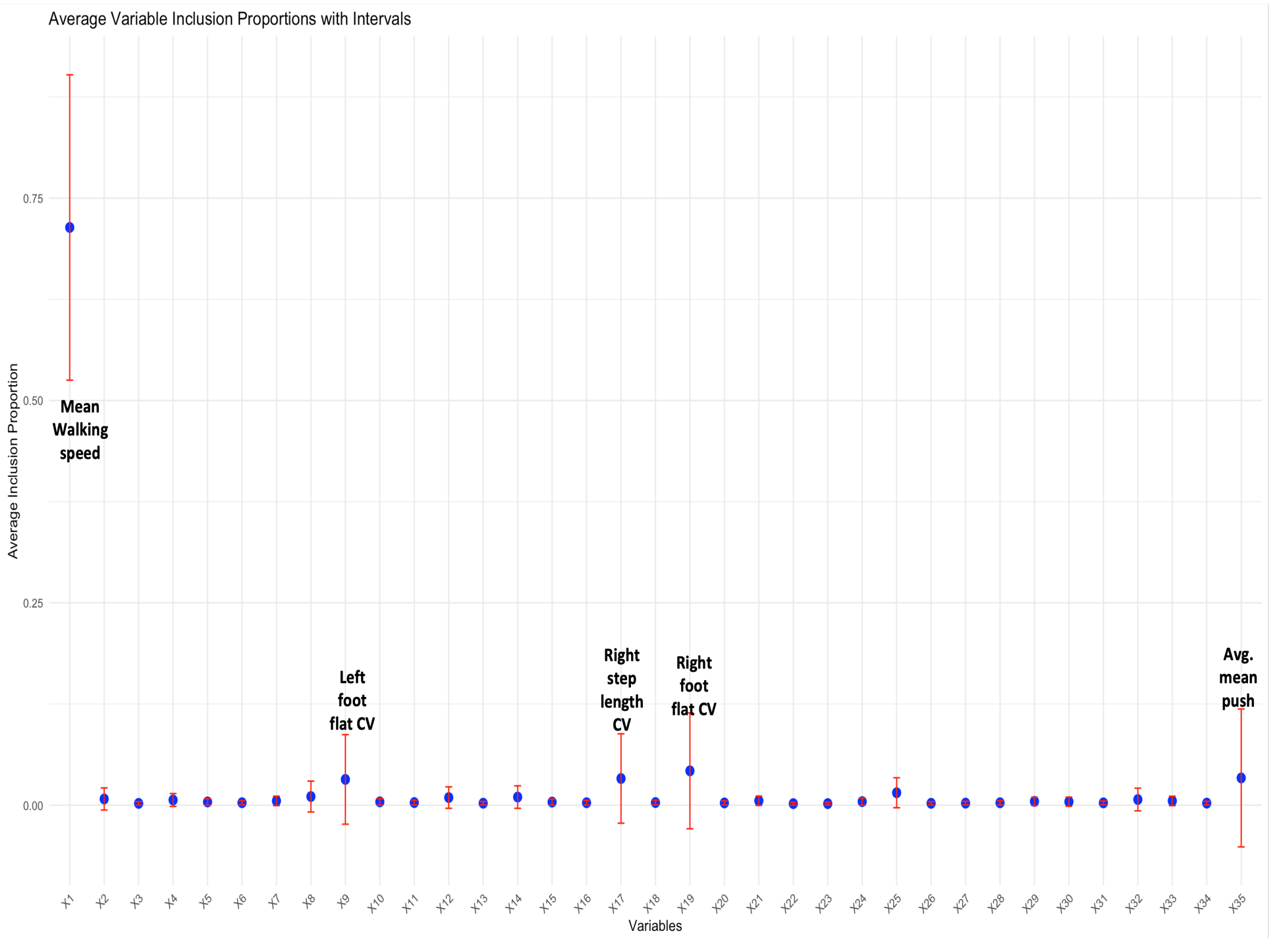


Figure 4.9: Visualization of uncertainty from an MCMC study on average variable inclusion proportions (VIP) for various gait parameters, using a sparsity-induced prior-based BART model for the classification cohort based on the Healthy vs Orthopedic comparison. We investigated each gait parameter's variable inclusion proportions (VIP) using a sparsity-induced prior-based BART model to understand feature importance across the three cohorts. We carried out an MCMC simulation study with 1000 samples to understand the uncertainty in the VIP scores for each run. This approach allowed us to assess the stability of feature importance resulting from the prior-modified BART model. The uncertainty in the bounds reflects variation from the Markov chains used in each iteration. We highlighted the important features due to the sparsity-induced prior-based BART for each cohort.

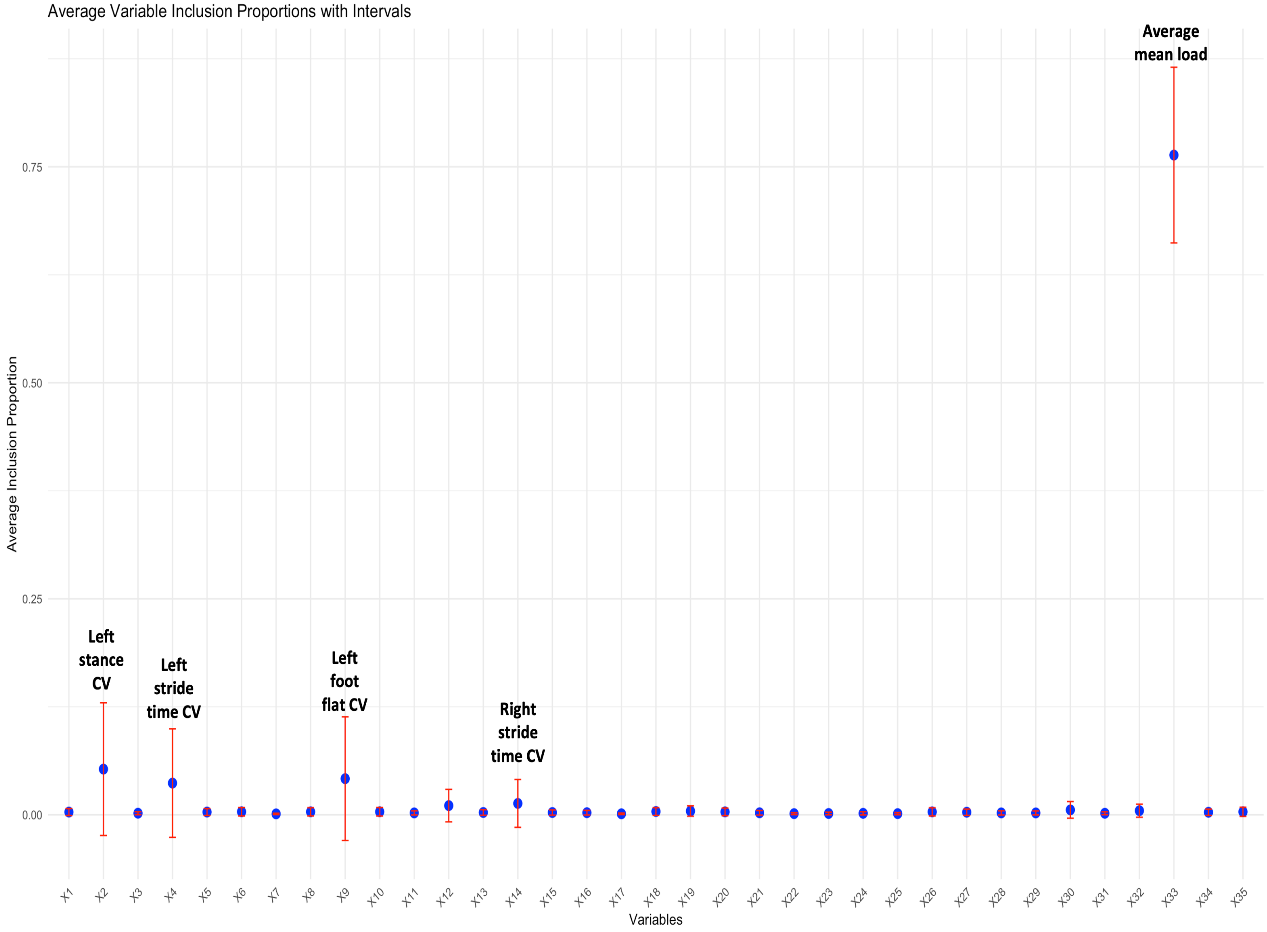


Figure 4.10: Visualization of uncertainty from an MCMC study on average variable inclusion proportions (VIP) for various gait parameters, using a sparsity-induced prior-based BART model for the classification cohort based on the Orthopedic vs. Neurological comparison. We investigated each gait parameter's variable inclusion proportions (VIP) using a sparsity-induced prior-based BART model to understand feature importance across the three cohorts. We carried out an MCMC simulation study with 1000 samples to understand the uncertainty in the VIP scores for each run. This approach allowed us to assess the stability of feature importance resulting from the prior-modified BART model. The uncertainty in the bounds reflects variation from the Markov chains used in each iteration. We highlighted the important features due to the sparsity-induced prior-based BART for each cohort.

Feature importance and individual effects on the prediction function for the Healthy vs. Orthopedic disorder cohort

For the Healthy vs Orthopedic cohort, the permutation-based procedure flags five features. “Average step length,” “Right foot CV stride length,” “Mean asymmetry single support,” “Average walking speed,” “Right foot CV step length,” and “Average stance” came as important based on the ‘local’ threshold. “Average step length” and “Mean asymmetry single support” came out as important features based on the ”Global SE” threshold. So, ”Average mean step length” and ”Mean asymmetry single support” appeared twice as important features based on two threshold-based schemes. The software-generated plot from the permutation test is given in Fig. 4.6.

Sparsity-induced prior-based BART ranks ”Average Walking Speed” as the most important feature concerning the VIP scores. Additionally, “Left foot CV foot flat,” “Right foot CV step length,” “Right CV foot flat,” and “Average push” came out as important features. Here, “Average walking speed” received the highest rank in terms of its VIP scores. A Monte Carlo study is performed to assess the uncertainty in VIP scores (see Fig. 4.9).

Fig. 4.12 (a) reveals a non-linear relationship between the Average mean step length and the model’s prediction for the healthy class. As the average step length increases, the model’s prediction for a person being classified as Healthy initially decreases (negative effects), then sharply increases (positive effects) before plateauing. This suggests that very short step lengths might not be a significant differentiating factor in predicting the Healthy class. Subjects with short step lengths in their gait cycle are more likely to be classified within the Orthopedic disorder cohort. A similar effect is observed for the Average Walking Speed on the model’s prediction. Here, BART also assigns a higher probability to the healthy class for subjects with higher average walking speeds.

After evaluating the individual average effects of the Average mean asymmetry single support, we notice an interesting non-linear relationship with the model's prediction. The curve shows a sharp downward trend followed by a flattening at lower effect levels. Fig. 4.12 (c) shows that at lower levels of asymmetry, the likelihood of being classified as healthy increases, indicating that low asymmetry in single support is a key indicator of healthy individuals. However, as asymmetry increases, the model predicts a sharp decline in the probability of a healthy classification, suggesting that greater asymmetry is more indicative of orthopedic-related issues. Beyond a certain threshold (around 10), further increases in asymmetry provide little new information, and the model likely attributes those instances to Orthopedic individuals.

Feature importance and individual effects on the prediction function for the Orthopedic vs. Neurological disorder cohort

Finally, we evaluate the feature-importance of the Orthopedic vs Neurological cohort. The permutation-based procedure flags two features. “Average mean load” became an important feature based on all three thresholds for the permutation-based approach. Additionally, it ranks higher than all the features based on its VIP score. “Right foot CV load” comes out as an important feature based on the “local” and “Global SE” thresholds.

Sparsity-induced prior-based BART captures four features as important in terms of the VIP scores. “Average mean load,” “Left foot CV stance,” “Left foot CV stride time,” and “Left foot CV foot flat” all come as the important features. Here, “Average mean load” ranks much higher in terms of its variation inclusion proportion score (VIP) when compared with the other features.

Thus, the ”average mean load phase” emerges as the most important feature based on its VIP score for both methods. The software-generated plot from the permutation test, along with the uncertainty analysis in VIP scores due to the sparsity-induced BART model, is shown in Fig. 4.7 and Fig. 4.10.

Fig. 4.13 reveals a non-linear relationship between the individual effects of the Average mean load Phase and the model's prediction for the healthy class. Subjects with higher average loads in their gait cycle are more likely to be classified within the Neurological disorder cohort. This trend for the "Average mean load Phase" feature is consistent with the Healthy vs. Neurological cohorts' classification.

Additionally, Sparsity-induced prior-based BART framework also indicates that alteration in gait parameters associated with the loading phase (i.e. the "Average mean load Phase" feature) can be considered as some of the important features for understanding patient-level classification among the two disorder groups. Let us evaluate these gait features from the ALE plots shown in Fig. 4.14 in the following manner: (a) left CV flat foot: The accumulated local effect (ALE) plot indicates that the model assigns higher probabilities to orthopedic patients with less variability in left foot flatness. In contrast, patients with greater variation in left foot flatness are more likely to be classified in the neurological cohort. (b) left foot CV stride time: Higher variation in left foot stride time is again associated with the neurological cohort when compared with the orthopedic group. (c) left foot CV stance: Stance phase is a parameter associated with the loading phase. The ALE plot reveals that higher variation in left foot stance is more prevalent in the neurological disorder group than in the orthopedic group; (d) right foot CV stride time: Stride time is also an important gait feature which gets affected by the loading phase. In the ALE plot, the model predictions show that higher variation in right foot stride time negatively impacts the orthopedic group. Consequently, patients with greater variation in right foot stride time are more likely to be classified in the neurological disorder cohort. A more detailed discussion of their biological interpretations is provided in the Concluding Chapter.

These key features underscore the effectiveness of our BART-based gait classification model in delivering both robust gait predictions and detailed insights into how specific gait-related characteristics influence the likelihood of classifying individuals as healthy or pathological. The model effectively captures the non-linear impact of gait-related features on classification outcomes, providing a clearer understanding of how variations in these features contribute to changes in the risk of pathology.

We further performed a benchmark comparison with the standard ML (kernel-based/tree-based) classification models. They include SVM (Support vector Machine), KNN (K-Nearest Neighbours), Decision Trees, Logistic Regression, NN (Neural Nets), Naïve Bayes, Ensemble-Bagged Trees and Discriminant. All classification tasks were carried out for each cohort. Prediction performance was evaluated using the same metrics as those applied to BART. The same split ratio and cross-validation folds (as mentioned in the Preprocessing section) were used for model training and to assess out-of-sample performance on the test set. A SHAP (SHapley Additive exPlanations) [193] analysis was conducted for all models across the classification cohorts to examine the consistency of feature importance between the bench-marked models and BART (see next chapter).

Extracted Gait Parameters	
Gait Parameters	**Description**
Spatiotemporal	• Walking speed (ms^{-1}) • Swing and stance phase (%) – left and right • Single and double support time (s) – left and right • Stride time (s) and length (m) • Step time (s) and length (m) – left and right • Load, push, and flat foot phase (%)
Statistical	• Mean (μ) • Standard Deviation (SD - σ) • Coefficient of Variations (CV $= \sigma/\mu$)
Asymmetry	Asymmetry $= 100 \times \left\lvert \ln\left(\frac{X_{\text{left}}}{X_{\text{right}}}\right) \right\rvert$, where X_{left} and X_{right} are the values of a specific gait parameter for the left and right sides, respectively.

Table 4.1: Extracted Gait parameters

Table 4.2: Participants' characteristics of three separate participant cohorts.

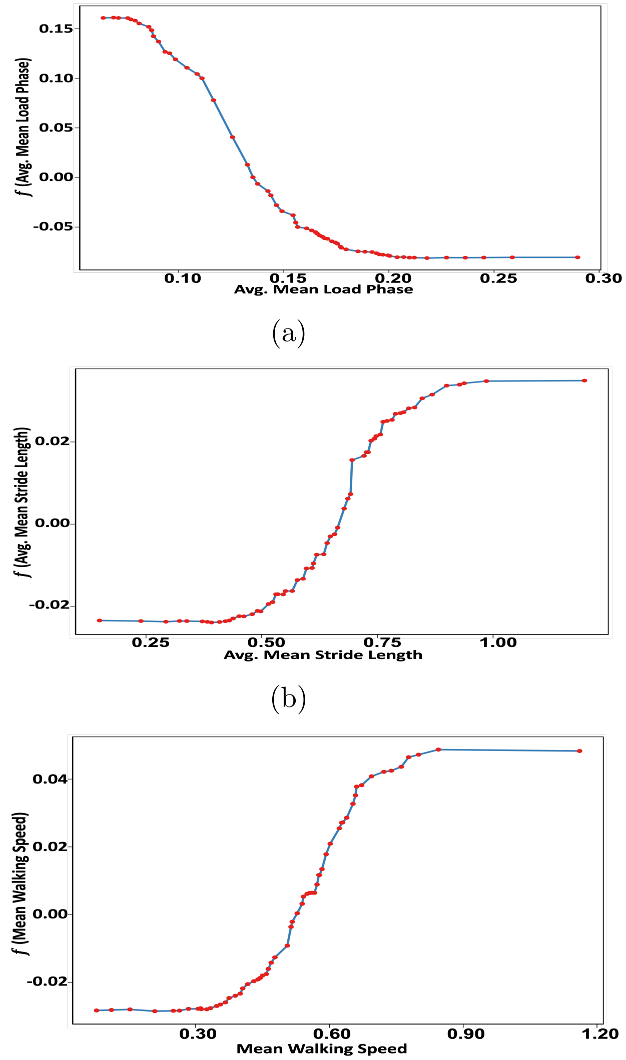
	Healthy	**Orthopedic**	**Neurological**	**Total**	**p-value**
Number of subjects	52	53	125	230	
Number of trials	242	243	535	1020	
Gender (%)					
Males	35 (67.3%)	26 (49.1%)	80 (64.0%)	141 (61.3%)	
Females	17 (32.7%)	27 (50.9%)	45 (36.0%)	89 (38.7%)	
Age (years ± SD)	36.4 ± 20.8	60.1 ± 19.4	61.6 ± 13.2	55.5 ± 19.6	$< 0.01 **$
Height (cm ± SD)	173.4 ± 10.8	169.2 ± 10.2	169.8 ± 8.7	170.5 ± 9.7	$< 0.05*$
Weight (kg ± SD)	70.7 ± 12.4	77.7 ± 17.0	72.7 ± 15.7	73.4 ± 15.4	0.053
BMI					
BMI ± SD	23.6 ± 3.9	27.1 ± 5.6	25.0 ± 4.4	25.2 ± 4.7	$< 0.01 **$
BMI Group (%)					
UW	5.8%	0.0%	5.6%	4.3%	
HW	63.5%	41.5%	42.4%	47.0%	
OW	23.1%	28.3%	39.2%	33.0%	
Ob	7.7%	28.3%	9.6%	13.5%	
NR	-	1.9%	3.2%	2.2%	
Laterality (%)					
R	96.2%	94.3%	85.6%	90.0%	
L	3.8%	5.7%	3.2%	3.9%	
A	-	-	0.8%	0.4%	
NR	-	-	10.4%	5.7%	

p-values for continuous variables are calculated via ANOVA, and for binary variables via the χ^2 test.

Statistical significance levels: ** $p < 0.01$, * $p < 0.05$.

M = Male, F = Female, UW = Underweight, HW = Healthy weight, OW = Overweight, Ob = Obese,

R = Right, L = Left, A = Ambidextrous, NR = Not reported.

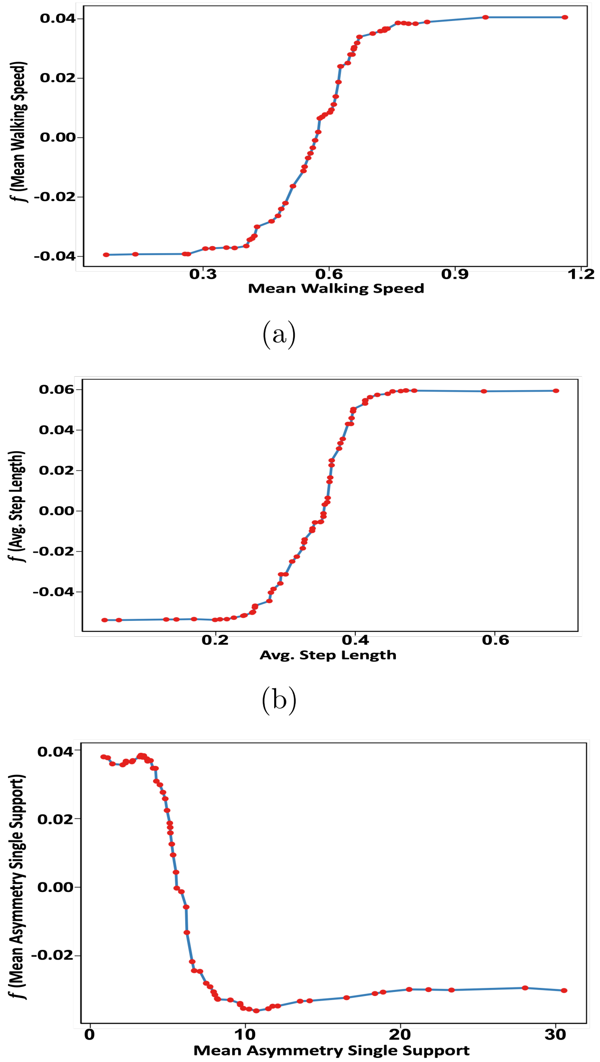


(a)

(b)

(c)

Figure 4.11: Effect of varying individual spatio temporal gait feature values on the model's prediction for the Healthy vs Neurological cohort. These accumulated local effect (ALE) (centered at 0) plots show the change in relative probability for each class based on a feature's value compared to its average effect. An increase in the likelihood for one class means a corresponding decrease in another class's probability.



(a)

(b)

(c)

Figure 4.12: Effect of varying individual spatio-temporal gait feature values on the model's prediction for the Healthy vs Orthopedic cohort. These accumulated local effect (ALE) (centered at 0) plots show the change in relative probability for each class based on a feature's value compared to its average effect. An increase in the likelihood for one class means a corresponding decrease in another class's probability.

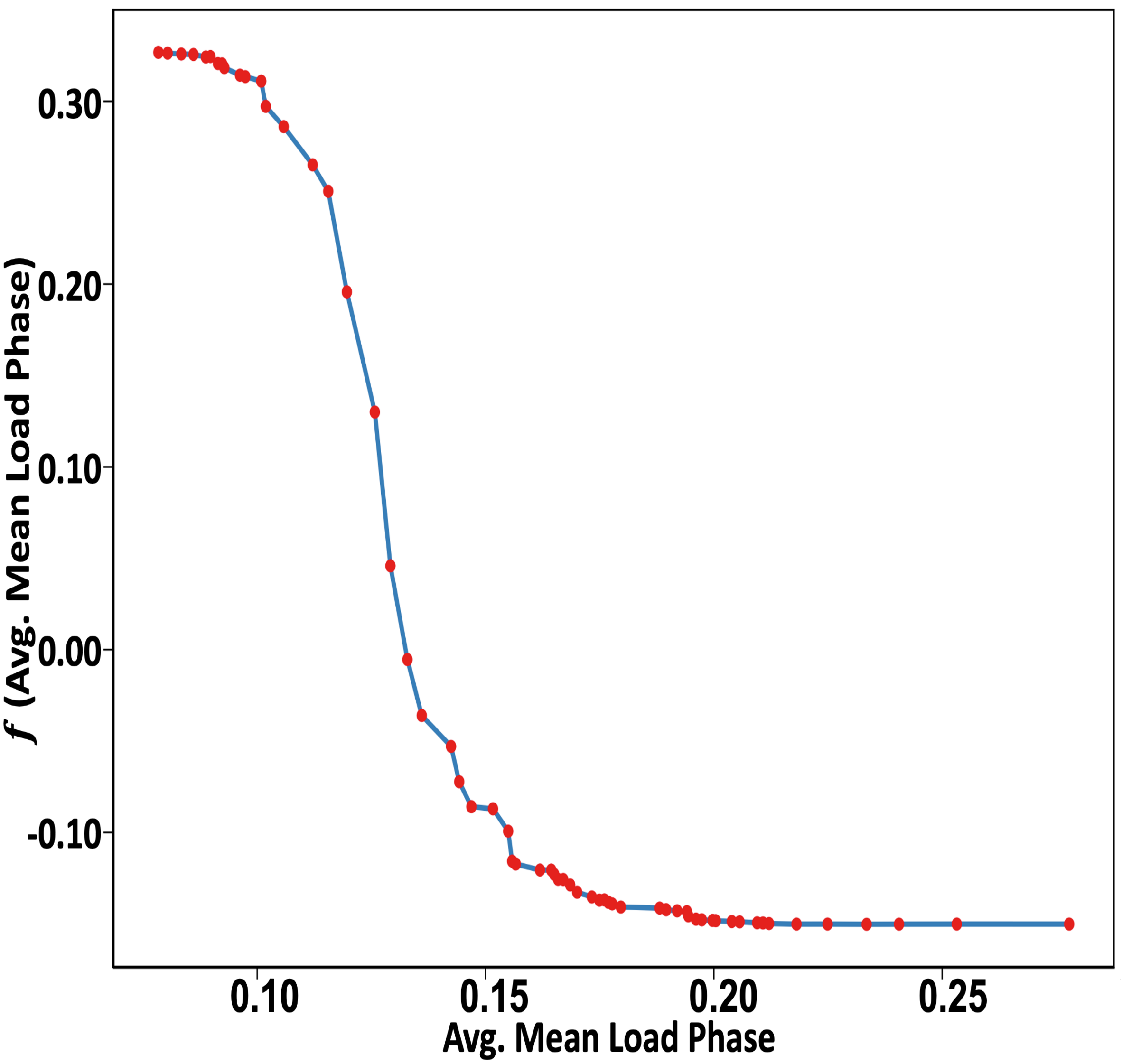
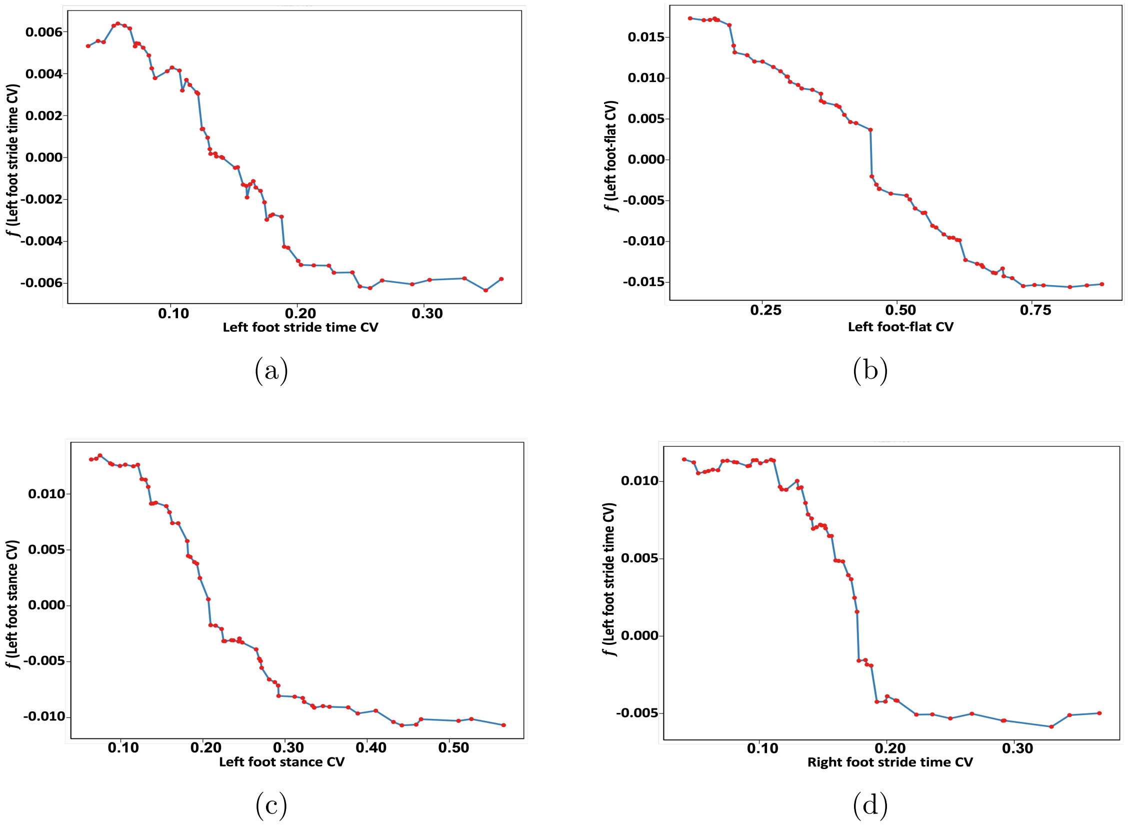


Figure 4.13: Effect of varying individual spatio temporal gait feature values on the model's prediction for the Orthopedic vs Neurological cohort. These accumulated local effect (ALE) (centered at 0) plots show the change in relative probability for each class based on a feature's value compared to its average effect. An increase in the likelihood for one class means a corresponding decrease in another class's probability.



(a)

(b)

(c)

(d)

Figure 4.14: Effect of varying individual spatiotemporal gait feature values on the model's prediction for the Orthopedic vs Neurological cohort. Our study revealed that average loading phase is a key differentiator between the orthopedic and neurological disorder cohort. The sparsity-induced prior-based BART framework also indicated that alterations in gait parameters associated with the loading phase can be considered important features for understanding patient-level classification among the two disorder groups.

Chapter 5

Comparison of BART's Predictive Performance Against Traditional Machine Learning Algorithms

5.1 Introduction

Bayesian Additive Regression Trees (BART) is a powerful non-parametric machine learning model that offers advantages in predictive performance and feature importance analysis. In this chapter, we compare BART's performance with several traditional machine learning (ML) algorithms, including the following:

- **Support Vector Machine (SVM)**: Support Vector Machines (SVM) [205] aim to optimally separate classes by maximizing the margin between them while minimizing classification errors. In most real-life problems, the data are not linearly separable, necessitating the use of a nonlinear kernel in the SVM process [205].

  The nonlinear radial basis function (RBF) kernel is selected to map the data into a higher-dimensional space, where a k-dimensional hyperplane separates the classes. Here, k represents the number of features used in the model. The kernel parameters are optimized to balance maximizing the margin between classes while minimizing misclassification costs.

A RBF kernel-based, cost-sensitive SVM classification model is developed for the binary classification task across the three cohorts by solving the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C_{PC} \sum_{i=1}^{n_{PC}} \xi_i + C_{HC} \sum_{j=1}^{n_{HC}} \xi_j, \tag{5.1.1}$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{5.1.2}$$

where w, b, and ξ are optimization parameters, and C_{PC} and C_{HC} represent the misclassification costs for Pathological cohort (PC) and healthy cohorts, respectively. The label vector $y \in \{-1, 1\}$ assigns 1 to PD subjects and -1 to control subjects. Here, x denotes the feature values of data points to be classified. Classification is determined based on the sign of $w^T x + b$, which indicates on which side of the hyperplane a given data point falls. Thus, classification gets done as follows:

$$\text{Class} = \begin{cases} \text{PC}, & \text{if } w^T x_i + b \geq 1 - \xi_i, \quad y_i = 1 \\ \text{HC}, & \text{if } w^T x_i + b \leq -1 + \xi_i, \quad y_i = -1 \end{cases} \tag{5.1.3}$$

When misclassification costs are equal, i.e., $C_{HC} = 1$, the model treats both classes symmetrically. However, in predicting PC, the consequences of misclassification may differ. A cost-sensitive learning approach is implemented to account for these differences. In a conservative classification strategy, a subject is more likely to be classified as healthy rather than PC when uncertainty exists. This is achieved by assigning a higher misclassification cost for incorrectly classifying a healthy subject as PC than for misclassifying a PC subject as healthy, resulting in $C_{PC} < C_{Healthy}$.

- **Naive Bayes**: For a feature vector $\mathbf{F} = \{F_1, \ldots, F_n\}$ extracted from sensor data, a Naïve Bayes classifier [206] is employed to infer the probability of a patient belonging to one of two possible states, given $\mathbf{F}$. The classifier is implemented using a Naïve Bayes model [206], which relies on the conditional probability $p(C|\mathbf{F})$, where C is a binary class variable representing the patient's state, i.e., Healthy vs Pathology. $\mathbf{F}$ is the feature vector.

  Applying Bayes' theorem and assuming the Naïve Bayes independence assumption—i.e., given the class label C, each feature F_i is conditionally independent of every other feature F_j $(j \neq i)$—the likelihood can be factorized as

  $$p(\mathbf{F}|C) = \prod_{i=1}^{n} p(F_i|C), \quad p(\mathbf{F}) = \prod_{i=1}^{n} p(F_i). \tag{5.1.4}$$

  Thus, the posterior probability of the class variable can be expressed as

  $$p(C|\mathbf{F}) = \frac{p(C) \prod_{i=1}^{n} p(F_i|C)}{\prod_{i=1}^{n} p(F_i)}. \tag{5.1.5}$$

  This formulation enables efficient inference of the patient's state based on the extracted sensor features. The simplicity of Eq. (5.1.5) makes Naïve Bayes classifiers highly suitable for resource-constrained applications, as the model parameters can be estimated efficiently.

- **K-Nearest Neighbors (KNN)**:

K-Nearest Neighbors (KNN) is a non-parametric, supervised machine learning algorithm [207] used for classification tasks. The core principle of KNN is based on computing the Euclidean distance between an unknown data point (test sample) and the training data samples.

Let $x \in \mathbb{R}^{n\times d} = (x_1, \ldots, x_n)$ be the feature matrix, where n represents the number of training samples, and d denotes the number of features. Given an arbitrary test sample x_o, the Euclidean distance in the feature space $\mathbb{R}^p$, with $p = 2$, is defined as

$$d_i = \|x_r - x_o\|_p = \left(\sum_{i=1}^{n} |x_i - x_o|^p\right)^{\frac{1}{p}}. \tag{5.1.6}$$

To classify a set of features into M distinct classes, the classified entities can be represented as

$$\Omega = \{\Omega_1, \ldots, \Omega_m\}, \quad 1 \leq m \leq d. \tag{5.1.7}$$

By selecting the k training samples closest to the unknown data point x_o, the KNN algorithm determines the number of nearest neighbors assigned to each class label $l \in \mathbb{R}^{1\times d} = (l_1, \ldots, l_d)$ in the training set:

$$S_r = \{(x_1, l_1), \ldots, (x_n, l_d)\}, \tag{5.1.8}$$

where $x_r \in \mathbb{R}^{n\times 1} = (x_1, \ldots, x_n)$ represents the training examples associated with S_r. Each entry in S_r corresponds to a class label in Ω. The classification process involves estimating the conditional probability of each class as an empirical fraction:

$$P_r = P(m(l) \in l \mid x = x_o) = \frac{1}{k} \sum_{i \in N(l, S_r)} I(x_r \in l), \tag{5.1.9}$$

where $N(l, S_r)$ represents the indices of the k nearest neighbors to class l in the training set S_r. The indicator function $I(.)$ is defined as:

$$I(w) = \begin{cases} 1, & \text{if } w \text{ is True} \\ 0, & \text{otherwise.} \end{cases} \tag{5.1.10}$$

All gait-related features were analyzed using the implemented KNN algorithm for the three cohorts. Here, we considered $M = 2$, i.e; $\Omega = \{\Omega_1, \Omega_2\}$. for the three cohorts. The k parameter was tuned by doing a 5-fold cross validation.

- **Neural Networks (NN)**: A deep learning approach that models complex relationships between features and target variables. A single-layer neural network model with weight w, bias b, and activation function f is given by

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right). \tag{5.1.11}$$

A three-layer Multi-Layer Perceptron (MLP) has an input layer, a hidden layer, and an output layer. Successive layers are fully connected by weights. An MLP updates the weights iteratively to map a set of input vectors $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ to a set of corresponding output vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p)$. An input $\mathbf{x}_i$ is presented to the input layer and multiplied by the weights. All the weighted inputs to each unit in the upper layer are then summed up and produce an output h_i as given by the following equations:

$$h_i = f(\mathbf{W}_h \mathbf{x}_i + \theta_h), \tag{5.1.12}$$

$$y_i = f(\mathbf{W}_o h_i + \theta_o), \tag{5.1.13}$$

where $\mathbf{W}_h$ and $\mathbf{W}_o$ are the hidden and output layer weight matrices, h_i is the vector denoting the response of the hidden layer to input $\mathbf{x}_i$, θ_o and θ_h are the output and hidden layer bias vectors, respectively, and $f(\cdot)$ is the sigmoid activation function. The cost function to be minimized is the sum of squared error E, defined as

$$E = \frac{1}{2}\sum_i (\mathbf{t}_i - \mathbf{y}_i)^\top (\mathbf{t}_i - \mathbf{y}_i), \tag{5.1.14}$$

where $\mathbf{t}_i$ is the target output vector for pattern i. The standard backpropagation training algorithm [208] uses gradient descent techniques to minimize E, but suffers from slow convergence and frequently stuck in local minima. There are a number of variations to backpropagation algorithm to achieve faster convergence and to avoid local minima. However, generalization ability of neural network is the most important factor. A desired neural network model should produce small error not only on sample data but also on out of sample data. A potential problem is the unsmoothening of the trained weights which may contribute to a network's poor performance in generalization. To produce a network with better generalization ability, [209] proposed a method to constrain the size of network parameters by regularization. Regularization technique forces the network to settle to a set of weights and biases having smaller values. This causes the network response to be smoother and less likely to overfit [210] and capture noise. In regularization technique, the cost function F is defined as

$$F = \gamma E + (1 - \gamma) E_w \tag{5.1.15}$$

where E is the same as defined in Eq. (5.1.14), $E_w = \|\mathbf{w}\|^2/2$ is the sum of squares of the network parameters, and γ (< 1.0) is the performance ratio parameter, the magnitude of which dictates the emphasis of the training on regularization. A large

γ will drive the error E to small value whereas a small γ will emphasize parameter size reduction at the expense of error and yield smoother network response. One approach of determining optimum regularization parameter automatically is the Bayesian framework. It considers a probability distribution over the weight space, representing the relative degrees of belief in different values for the weights. Using Gaussian probability distribution and Bayesian rule, the optimum value of γ at the minimum point of F can be determined. A detailed description of the method is available in [209].

- **Ensemble - Bagged Trees**: Bagging, short for *bootstrap aggregating* [211], is an ensemble learning technique that involves two key components: bootstrap sampling of the training data and aggregation of base learners. For classification problems, predictions are combined via majority voting, whereas for regression problems, outputs are averaged.

  This method significantly enhances generalization performance by combining multiple *unstable* base learners (e.g., decision trees). In contrast, *stable* learners such as k-nearest neighbors, radial basis function (RBF) networks, or support vector machines, tend to be insensitive to variations introduced by bootstrap sampling [211].

  Given a dataset of N observations (gait patterns), the bootstrap sampling procedure generates new training datasets $\{\mathbf{f}_n^{bd}\}$, each of size N, for training the individual base learners. This is achieved by randomly sampling (with replacement) from the original dataset $\{\mathbf{f}_n\}$ [213]. Each data point is selected with equal probability $1/N$, independent of whether it has been selected before or not. As a result, some instances may be repeated in the bootstrap sample, while others may be omitted.

  During inference, Bagging aggregates the predictions from all trained learners. For classification, the final prediction corresponds to the class receiving the majority of votes [212]. Breiman [211] demonstrated that Bagging can substantially reduce generalization error compared to using a single base learner. The detailed computational steps are outlined below.

Computation Process of the Bagging Algorithm:

Input:

- Gait dataset: $\{\mathbf{f}_n, t_n\}_{n=1}^N$, where N is the number of subjects for each classification cohort and $t_n \in \{1, 2\}$ is the class label;
- Weak learner model (decision tree): $h(\mathbf{f}_n)$;
- Number of weak learners: I.

Procedure:

1. For $i = 1, 2, \ldots, I$:
    (a) Generate a bootstrap sample $\{\mathbf{f}_n^{bd}\}$ from the original training data;
    (b) Train the i-th weak learner $h_i(\mathbf{f}_n^{bd})$ on this bootstrap sample;
    (c) Use the trained model $h_i(\mathbf{f}_n; \mathbf{f}_n^{bd})$ to predict class labels for input patterns;
2. End loop.

Output:

$$H_{\text{Bagging}}(\mathbf{f}_n) = \arg \max_{t \in \{1,2\}} \sum_{i=1}^{I} \left[h_i(\mathbf{f}_n; \mathbf{f}_n^{bd}) = t \right] \tag{5.1.16}$$

The final ensemble prediction H_{Bagging} assigns to each input $\mathbf{f}_n$ the class t that receives the highest number of votes across the ensemble of I learners.

- **Logistic Regression**: Logistic Regression (LR) is a statistical modeling approach used to estimate the probability of a binary outcome based on one or more explanatory variables [214]. For a given set of predictors, the probability of observing the event (e.g., class label 1) is modeled as

$$P(\mathbf{x}) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{i=1}^{p} \beta_i \, PC_i\right)}}, \tag{5.1.17}$$

where β_0 is the intercept term, β_i represents the coefficient associated with the i-th explanatory variable, and p is the total number of predictors.

The parameters $\{\beta_i\}$ are estimated using the method of maximum likelihood, which seeks the set of coefficients that maximize the likelihood of observing the given outcomes in the data.

The model estimates the log-odds (i.e., the natural logarithm of the odds ratio) that an instance belongs to one of the two classes — with class 1 representing control subjects and class 0 indicating disorder group subjects. For the last cohort, class 1 was used to indicate the orthopedic group and class 0 indicating the neurological disorder group.

A decision threshold of 0.5 was applied: if $P(\mathbf{x}) \geq 0.5$, the instance is classified as a control; otherwise, it is classified as a PD subject.

- **Discriminant Analysis**: A classification technique that separates data based on linear combinations of predictor variables. Linear Discriminant Analysis (LDA) assumes normal distribution of classes and maximizes class separability. In binary classification problems, where data resides in an n-dimensional space, Linear Discriminant Analysis (LDA) seeks to project the data onto a one-dimensional subspace defined by a projection vector $\mathbf{w}$. The goal is to find a direction that optimally separates the two classes by maximizing the ratio of between-class variance to within-class variance [215].

Let $I_y = \{i : y_i = y\}$, with $y \in \{-1, +1\}$, denote the set of indices corresponding to training samples in class y. The class separation in the direction $\mathbf{w} \in \mathbb{R}^n$ is quantified by the following objective function:

$$F(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}, \tag{5.1.18}$$

where $\mathbf{S}_b$ represents the between-class scatter matrix, defined as

$$\mathbf{S}_b = (\mu_{-1} - \mu_{+1})(\mu_{-1} - \mu_{+1})^\top, \tag{5.1.19}$$

and μ_y is the mean vector of the samples in class y, given by

$$\mu_y = \frac{1}{|I_y|} \sum_{i \in I_y} \mathbf{x}_i. \tag{5.1.20}$$

The within-class scatter matrix $\mathbf{S}_w$ is computed as the sum of the individual class scatter matrices:

$$\mathbf{S}_w = \mathbf{S}_{-1} + \mathbf{S}_{+1}, \tag{5.1.21}$$

with each class scatter matrix defined by

$$\mathbf{S}_y = \sum_{i \in I_y} (\mathbf{x}_i - \mu_y)(\mathbf{x}_i - \mu_y)^\top. \tag{5.1.22}$$

The optimal projection vector $\mathbf{w}$ that maximizes the class separability criterion $F(\mathbf{w})$ is obtained as

$$\mathbf{w} = (\mathbf{S}_{-1} + \mathbf{S}_{+1})^{-1}(\mu_{-1} - \mu_{+1}). \tag{5.1.23}$$

5.2 Performance Metrics

To obtain a robust estimation of overall classification performance, all selected models were trained and tested using a 5-fold cross-validation approach. Accuracy, F1 score, and AUC were used as performance metrics to evaluate the models. The assessment were made based on out-of-sample performance. Each metric was calculated separately for each classification cohort.

Provided below are the confusion matrix-based definition for each of those metrics.

F1 Score

$F1$ is a metric that strikes a balance between precision and recall, making it suitable for imbalanced datasets. The score is essentially the harmonic mean of *Precision* and *Recall*:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5.2.1}$$

where Precision and Recall are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{5.2.2}$$

$$Recall = \frac{TP}{TP + FN}. \tag{5.2.3}$$

Let us look at the confusion matrix-based representation of $F1$ score for a better understanding of the metric:

$$F1 = \frac{2TP}{2TP + FP + FN}, \tag{5.2.4}$$

where TP is true positives, FN is false negatives, FP is false positives, and TN is true negatives. The range of $F1$ score is constrained in the interval $[0, 1]$. For the $F1$ score, the minimum is reached for $TP = 0$, that is, when all the positive samples are misclassified and the maximum is achieved for $FN = FP = 0$, i.e., for perfect classification ($F1 = 1$).

Accuracy

Accuracy measures the proportion of correctly classified instances out of the total instances. Accuracy is one of the most commonly used evaluation metrics for binary classification problems. It measures the proportion of correct predictions—both true positives (TP) and true negatives (TN)—out of the total number of instances. While accuracy provides a quick snapshot of overall performance, it can be misleading in imbalanced datasets where one class dominates; a model predicting only the majority class could still yield high accuracy despite poor detection of the minority class. This is where the $F1$ score becomes particularly valuable. The $F1$ score balances precision (the proportion of true positives among predicted positives) and recall (the proportion of true positives (TP) among actual positives ($TP+FN$)), offering a harmonic mean that emphasizes the model's ability to correctly identify the positive class. Let us look at the expression of accuracy in terms of confusion matrix-based components:

$$\mathit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{5.2.5}$$

AUC Score

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a widely used performance metric for binary classification problems. It evaluates a model's ability to distinguish between the two classes across all possible classification thresholds. The ROC curve is a plot of the *True Positive Rate (TPR)* against the *False Positive Rate (FPR)* at various threshold settings. These are mathematically defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

where TP is true positives, FN is false negatives, FP is false positives, and TN is true negatives.

The AUC score represents the area under the ROC curve and can be interpreted as the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance by the classifier.

In this study, the `pROC` package in `R` was used to compute the AUC. The `roc()` function in `pROC` ranks the predicted probabilities and calculates TPR and FPR at each unique threshold. The AUC is then approximated using the trapezoidal rule:

$$\text{AUC} \approx \sum_{i=1}^{n-1} (\text{FPR}_{i+1} - \text{FPR}_i) \cdot \left(\frac{\text{TPR}_i + \text{TPR}_{i+1}}{2} \right)$$

.

The AUC value ranges from 0.5 (no discrimination, equivalent to random guessing) to 1.0 (perfect classification). A higher AUC indicates better discriminatory performance of the model.

5.3 Comparative Results

5.3.1 BART's Performance

We consider three binarization based on the previously mentioned cohorts. BART got deployed in each of the cases. First, we assess the predictive performance of cross-validated BART on Healthy vs Neurological binarization. The assessment is made based on the out-of-sample performance. The accuracy and the F1 score for this cohort are 0.97 and 0.98, respectively. The Receiver Operating Characteristic (ROC) curve measures the model's predictive ability for the two groups, with an area under the curve (AUC) of 0.99. The optimal threshold is calculated at 0.40 with a sensitivity of 1.0 and a specificity of 0.92. Next, we examine the performance of BART in the Healthy vs. Orthopedic cohort. The out-of-sample area-under-the-curve (AUC) for this cohort is 0.94. The accuracy and F1 scores are calculated to be 0.86 and 0.82. The optimal threshold is calculated at 0.58 with a sensitivity of 0.88 and a specificity of 0.83. Finally, we investigate the performance of BART in the Orthopedic vs. Neurological cohort. This cohort's 5-fold cross-validated area-under-the-curve (AUC) is 0.98. The accuracy and F1 scores are calculated to be 0.94 and 0.95, respectively. The optimal threshold is calculated at 0.38, and the corresponding sensitivity and specificity are 0.91 and 1, respectively. Next, we benchmark BART's performance against traditional machine learning (ML) models across the three cohorts. We use identical proportions (as that of BART's) of training and testing datasets, along with the same cross-validation procedures, to evaluate the out-of-sample predictive performance of these models on the selected dataset. The obtained results are provided in Table 5.1.

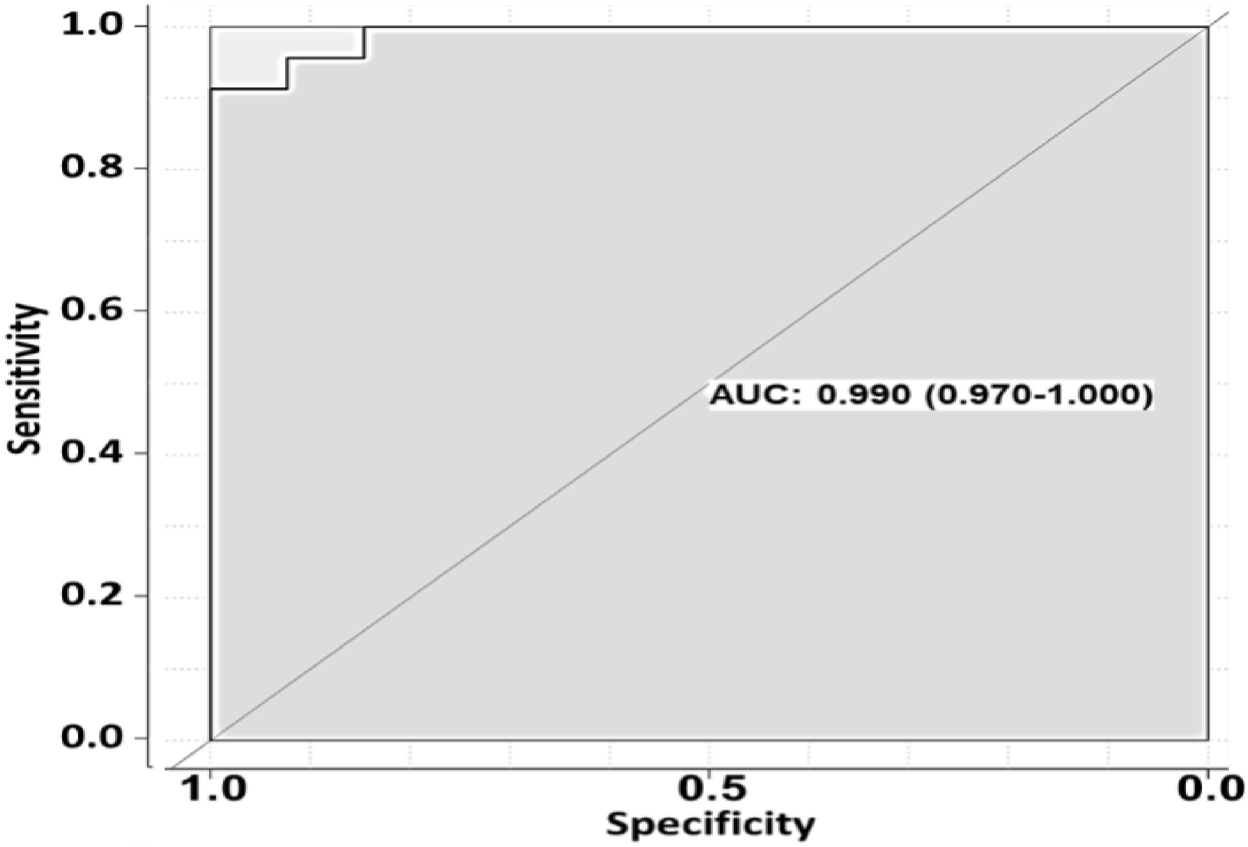


Figure 5.1: The receiver operating characteristics (ROC) curves comparing the patient-level out-of-sample predictive performance of BART for the Healthy and Neurological comparison.

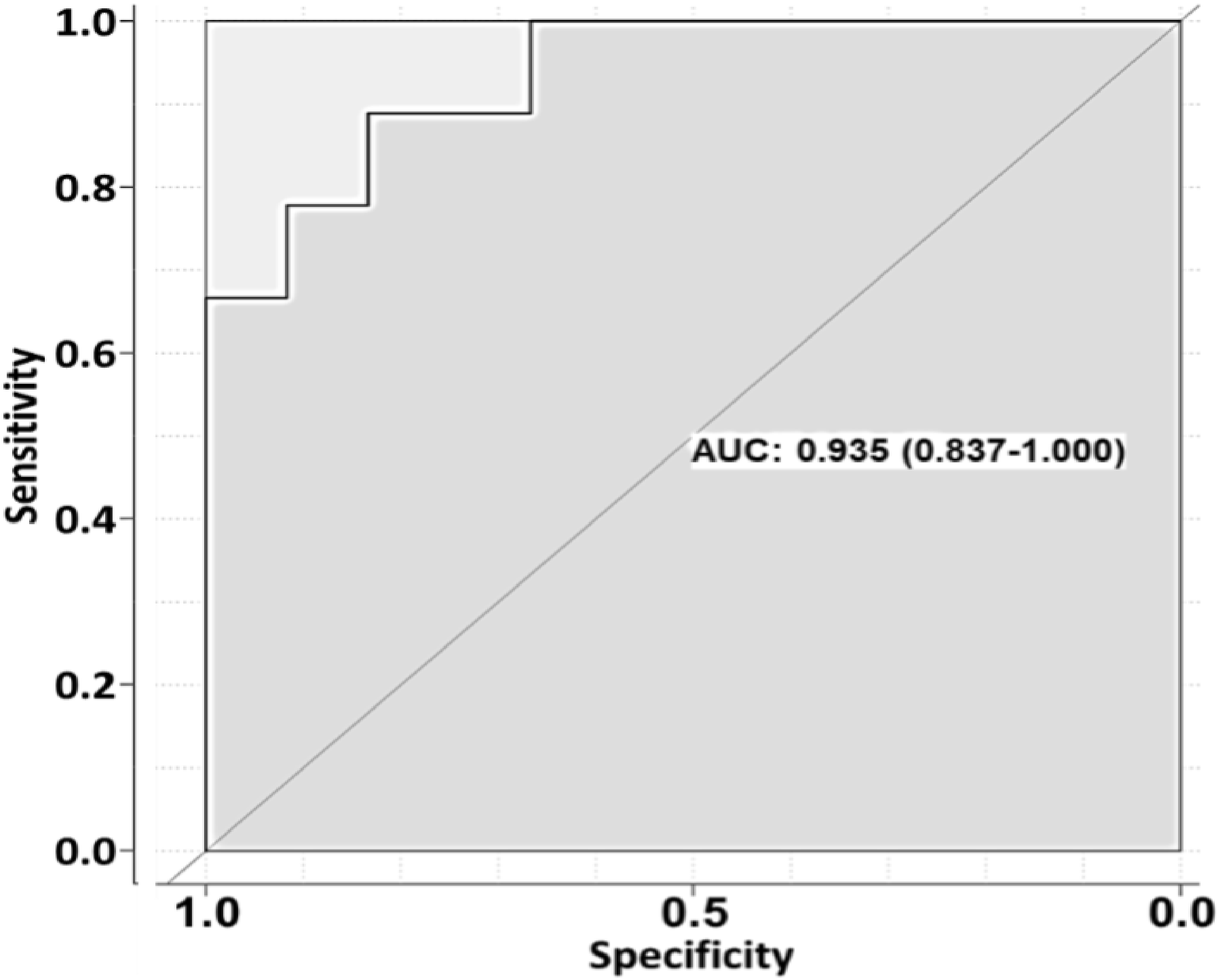


Figure 5.2: The receiver operating characteristics (ROC) curves comparing the patient-level out-of-sample predictive performance of BART for the Healthy and Orthopedic comparison.

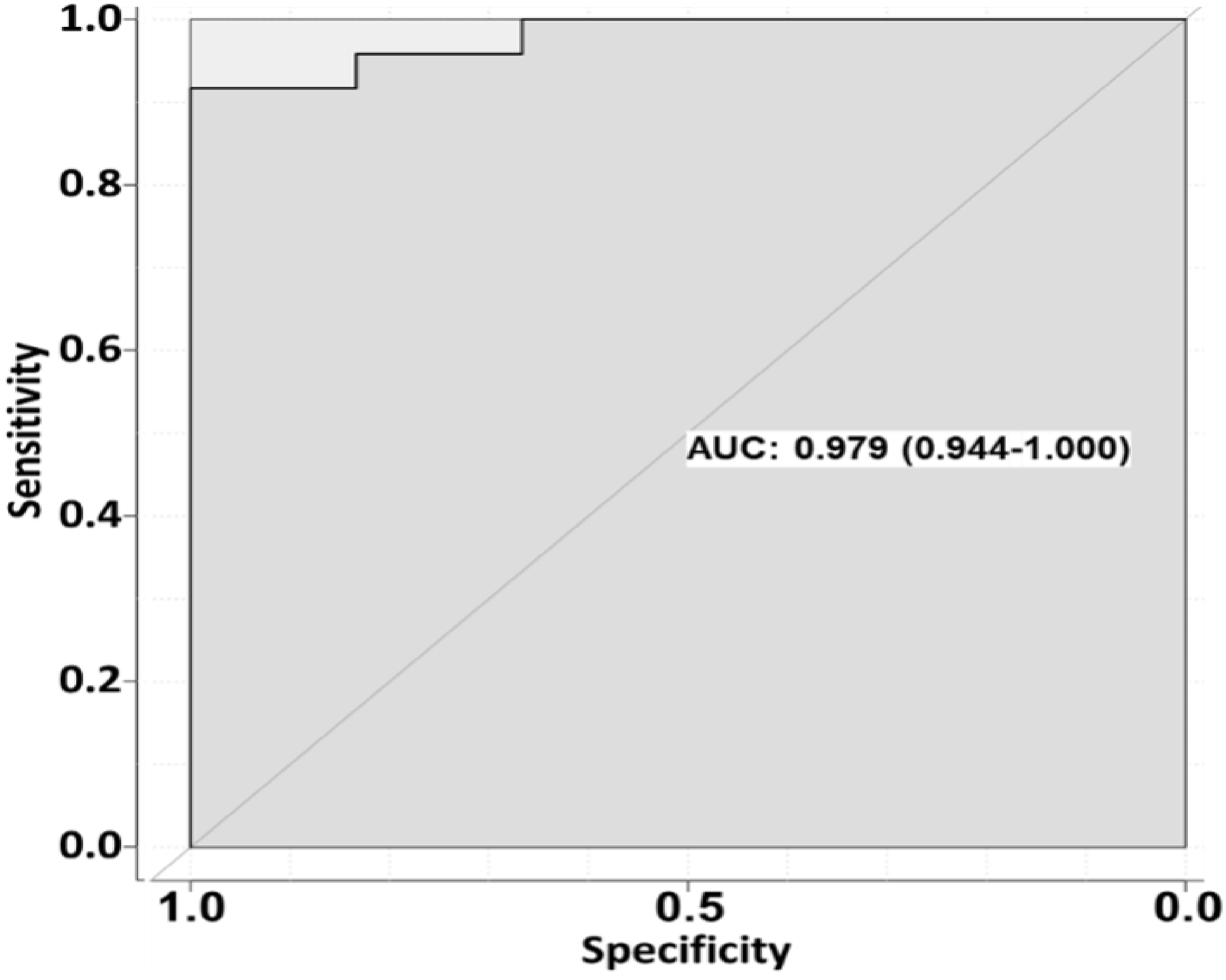


Figure 5.3: The receiver operating characteristics (ROC) curves comparing the patient-level out-of-sample predictive performance of BART for the Orthopedic and Neurological comparison.

We performed a benchmark comparison with the standard ML (kernel-based/tree-based) classification models. They include SVM (Support vector Machine), KNN (K-Nearest Neighbours), Decision Trees, Logistic Regression, NN (Neural Nets), Naïve Bayes, Ensemble-Bagged Trees and Discriminant. All classification tasks were carried out for each cohort. We assessed the classification performance for all three cohorts based on the following metrics: accuracy, F1 score, sensitivity, specificity, and area-under-the-curve (AUC) scores. Receiver operating curves (ROCs) and confusion matrices evaluated the patient-level model performance in each classification cohort. For the AUC scores, 95 % confidence intervals (CI) were calculated for each ROC-analysis. All experiments were seeded to ensure reproducibility. Our models were consistently trained and tested on the same datasets. The same split ratio and cross-validation folds were used as well for model training and to assess out-of-sample performance on the test set.

Table 5.1 presents a comparison of the model's performance with benchmark traditional machine learning (ML) algorithms across the three classification cohorts. The prediction performance is notably strong for both Healthy vs. Neurological and Orthopedic vs. Neurological groups. We benchmarked the model performance with traditional machine-learning models (ML) such as SVM, decision-tree-based models, logistic regressions, etc. BART's performance was either the same or better than the other traditional models. The performance declined for Healthy vs. Orthopedic cohort. However, BART outperformed all traditional machine learning (ML) algorithms for this cohort (see Table 5.1). This decline in performance may be due to factors such as higher collinearity among predictors, more noise than signal, or fewer distinguishing biological characteristics among the features.

BART consistently outperformed traditional ML models, achieving the highest accuracy, F1 score, and AUC across all three classification cohorts.

Table 5.1: Performance comparison of BART vs. Traditional ML models

Models	**Healthy vs. Neurological**			**Healthy vs. Orthopedic**			**Orthopedic vs. Neurological**		
	Accuracy	F1 Score	AUC	Accuracy	F1 Score	AUC	Accuracy	F1 Score	AUC
SVM-Linear	93.8	0.92	0.98	79.8	0.81	0.84	92.7	0.88	0.97
Naive Bayes	87.6	0.80	0.92	77.4	0.80	0.83	92.7	0.88	0.93
KNN	93.2	0.88	0.94	75.0	0.75	0.82	90.4	0.80	0.87
Neural Network	92.7	0.88	0.96	75.0	0.73	0.83	94.9	0.82	0.97
Ensemble - Bagged Trees	96.0	0.93	0.97	78.0	0.79	0.85	94.9	0.94	0.98
Logistic Regression	87.6	0.80	0.89	72.6	0.74	0.81	93.8	0.89	0.94
Discriminant Analysis	93.2	0.88	0.96	72.6	0.73	0.77	93.8	0.90	0.97
BART	**97.0**	**0.98**	**0.99**	**86.0**	**0.82**	**0.94**	**94.0**	**0.95**	**0.98**

5.4 SHAP Analysis

To further validate our results, we performed SHapley Additive exPlanations (SHAP) [193] analysis to assess feature importance. The findings confirmed that BART's feature importance rankings align well with traditional ML models, reinforcing its reliability in predictive modeling.

The most commonly used interpretable machine learning (ML) techniques are post hoc explanation methods, which are highly flexible and generally model-agnostic since they are applied after a predictive model has been designed and trained. In biomedical engineering applications, 'feature importance' methods are widely utilized. These approaches quantify the contribution of each input variable to the model's predictions by assigning an importance score. A higher absolute feature importance score indicates a more significant influence on the model's output. As discussed in Subsection 4.2.3, feature importance can be computed using various methods, including gain, split counts, and permutation, among others. SHAP analysis, introduced by Lundberg et al. [193], aims to unify the concept of feature importance, providing a comprehensive framework for interpreting ML model predictions. The novelty of SHAP analysis lies in: 1) The identification of a new class of additive feature importance measures, and 2) Theoretical results demonstrating the existence of a unique solution within this class that satisfies a set of desirable properties. Although these properties are well-known in classical Shapley value estimation methods, they were previously unrecognized in other additive feature attribution approaches. The first desirable property is *local accuracy.* This property asserts that when approximating the original model for a given input, local accuracy ensures that the explanation model must, at a minimum, replicate the output of the original model for the corresponding simplified input. The second property is the *missingness.* The missingness property states that if simplified inputs indicate feature presence, then any features absent in the original input should have no influence on the model's prediction. And, the third property is *consistency.* The consistency property asserts that if a model is modified such that the contribution of a specific simplified input increases or remains unchanged, regardless of other inputs, the attribution assigned to that input should not decrease. The SHAP analysis introduced an enhanced unified framework that ensures other feature importance methods, such as LIME, DeepLIFT, Layer-Wise Relevance Propagation, and Classic Shapley Value Estimation, do not violate the three properties mentioned above.

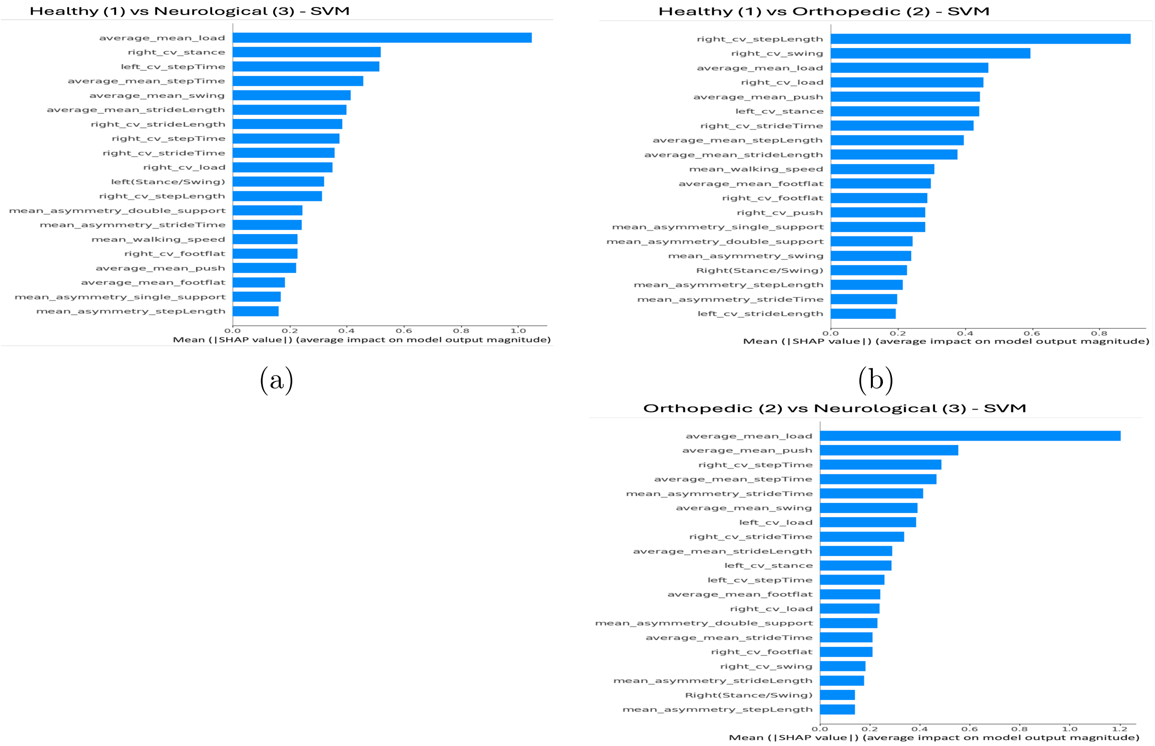
In our analysis, we have considered SHAP analysis as a benchmarking tool for feature importance based on the selected traditional ML models. This enables us to investigate the faithfulness and consistency of feature importance due to BART across the three classification cohorts.

From the SHAP feature importance plots (Fig. 5.4 - Fig. 5.10), we can observe the following:

(a) **Healthy vs. Neurological Disorder**: All ML algorithms, except discriminant analysis, identified Average mean load Phase as the most important feature for distinguishing this group from the healthy cohort. This finding aligns with BART's permutation-based and sparsity-induced prior approach for identifying key features.

(b) **Healthy vs. Orthopedic Disorder**: We observed variation in feature importance among the selected ML models. Bagged trees (which outperformed the other ML models) and discriminant analysis assigned the highest feature importance to average walking speed. Additionally, traditional ML models performed poorly compared to BART for this cohort. This may be due to the high correlation among features in this group, which poses challenges for standard SHAP analysis;

(c) **Orthopedic vs. Neurological Disorder**: All ML algorithms, except discriminant analysis (where Average mean load Phase was ranked as the second most important variable), identified Average mean load Phase as the most important feature for distinguishing these two groups. Again, this finding is consistent with BART's permutation-based and sparsity-induced prior approach for feature selection.

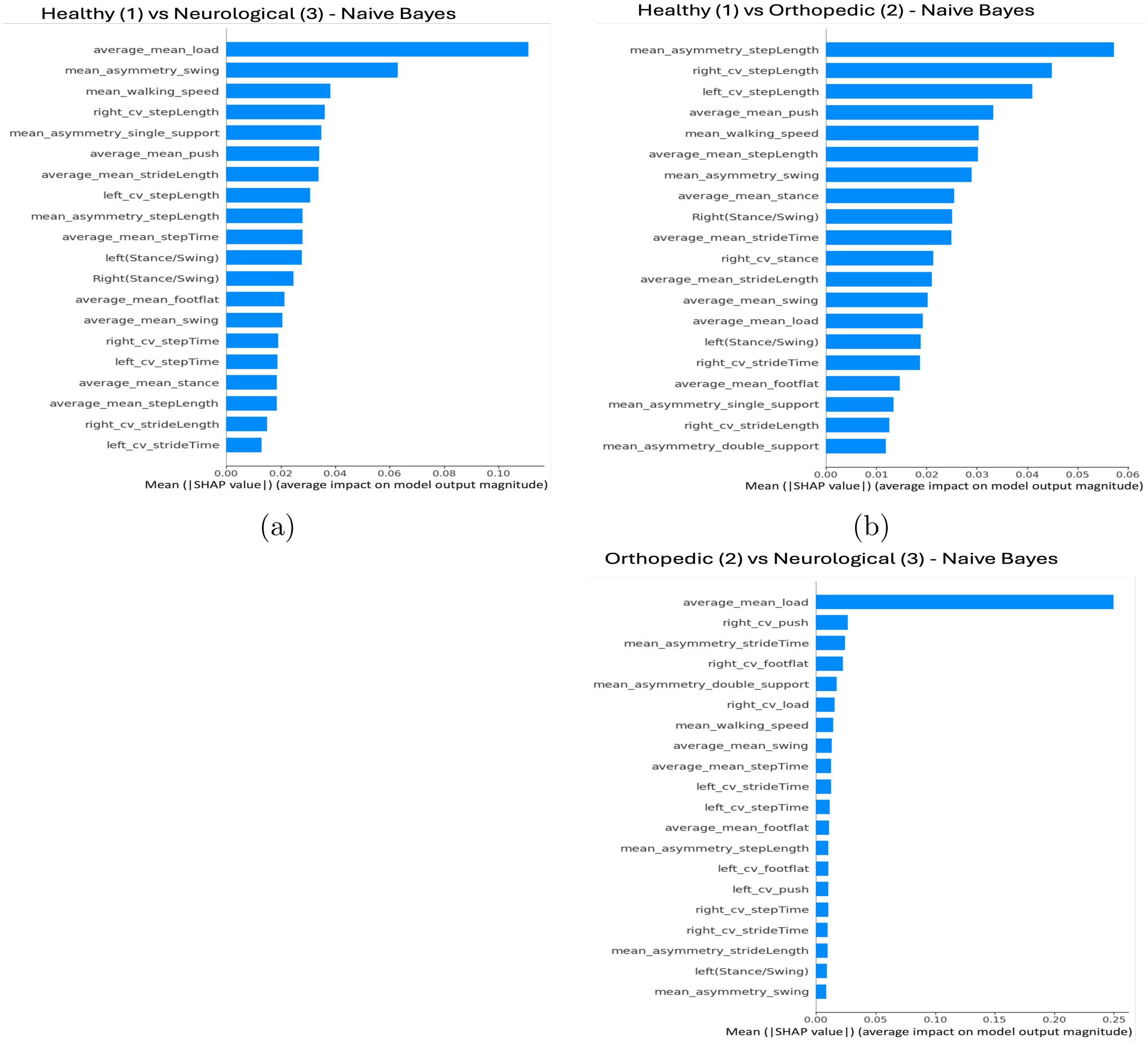


(a)

(b)

(c)

Figure 5.4: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.

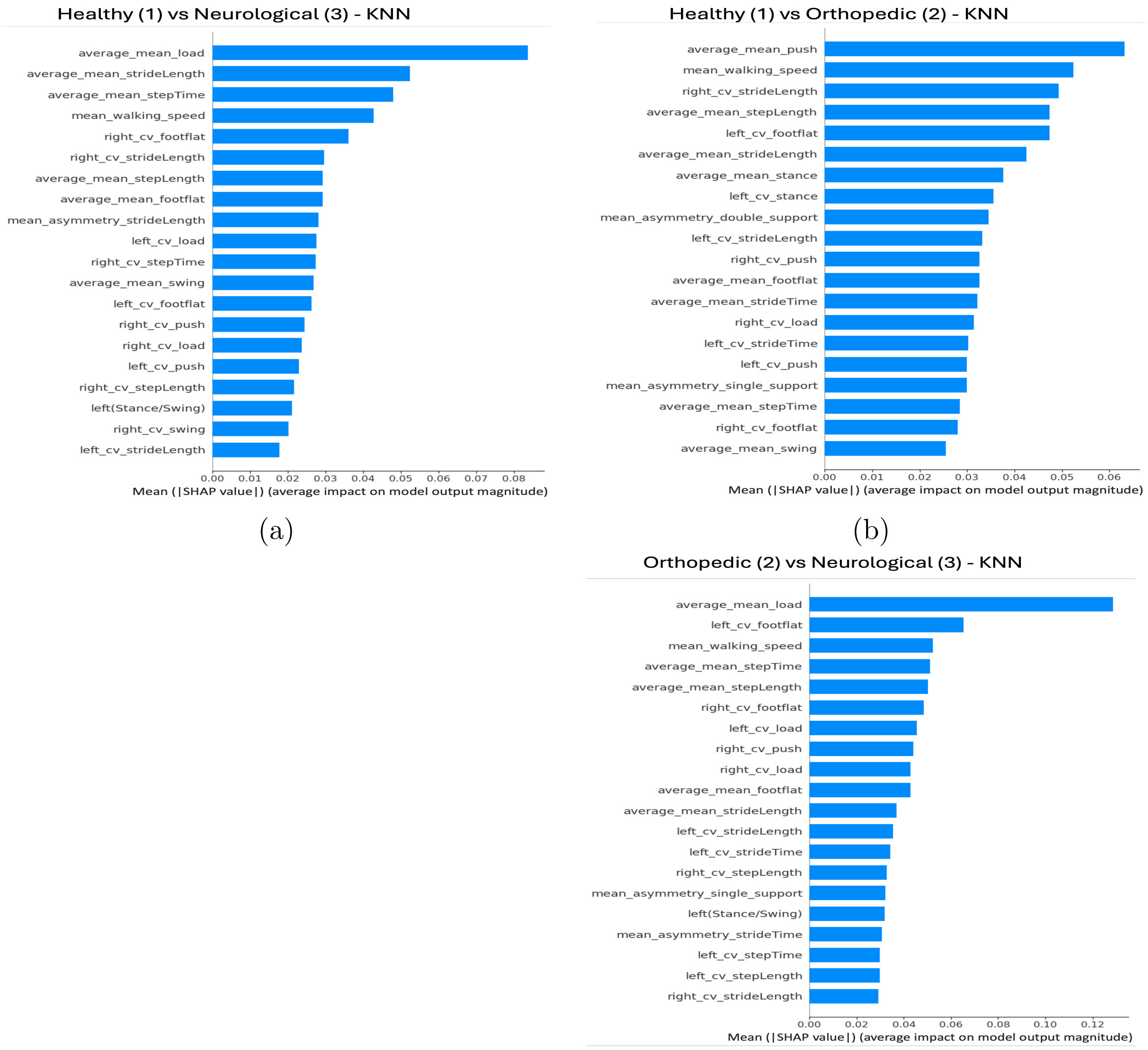


(a)

(b)

(c)

Figure 5.5: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.



(a)

(b)

(c)

Figure 5.6: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.

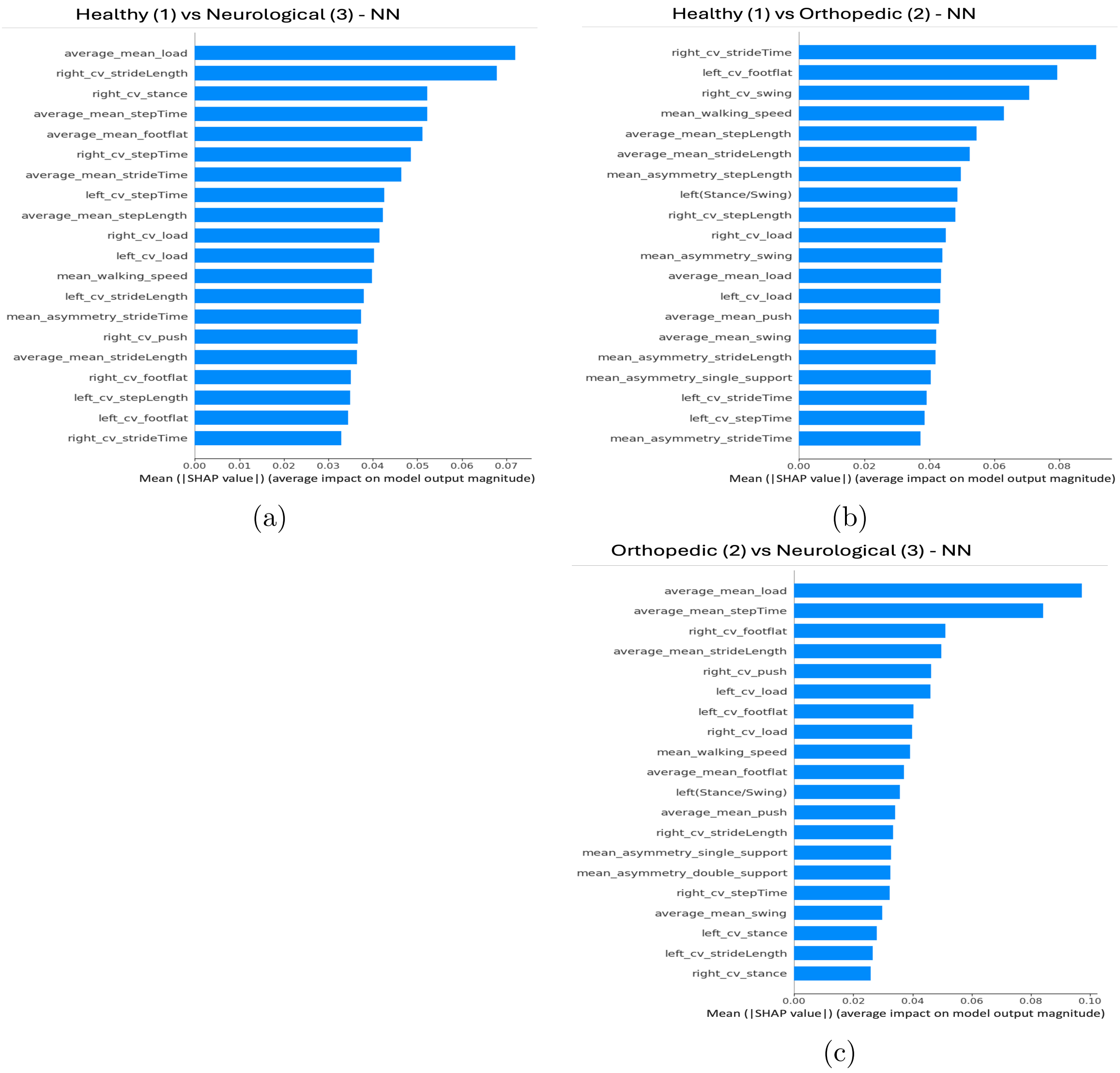
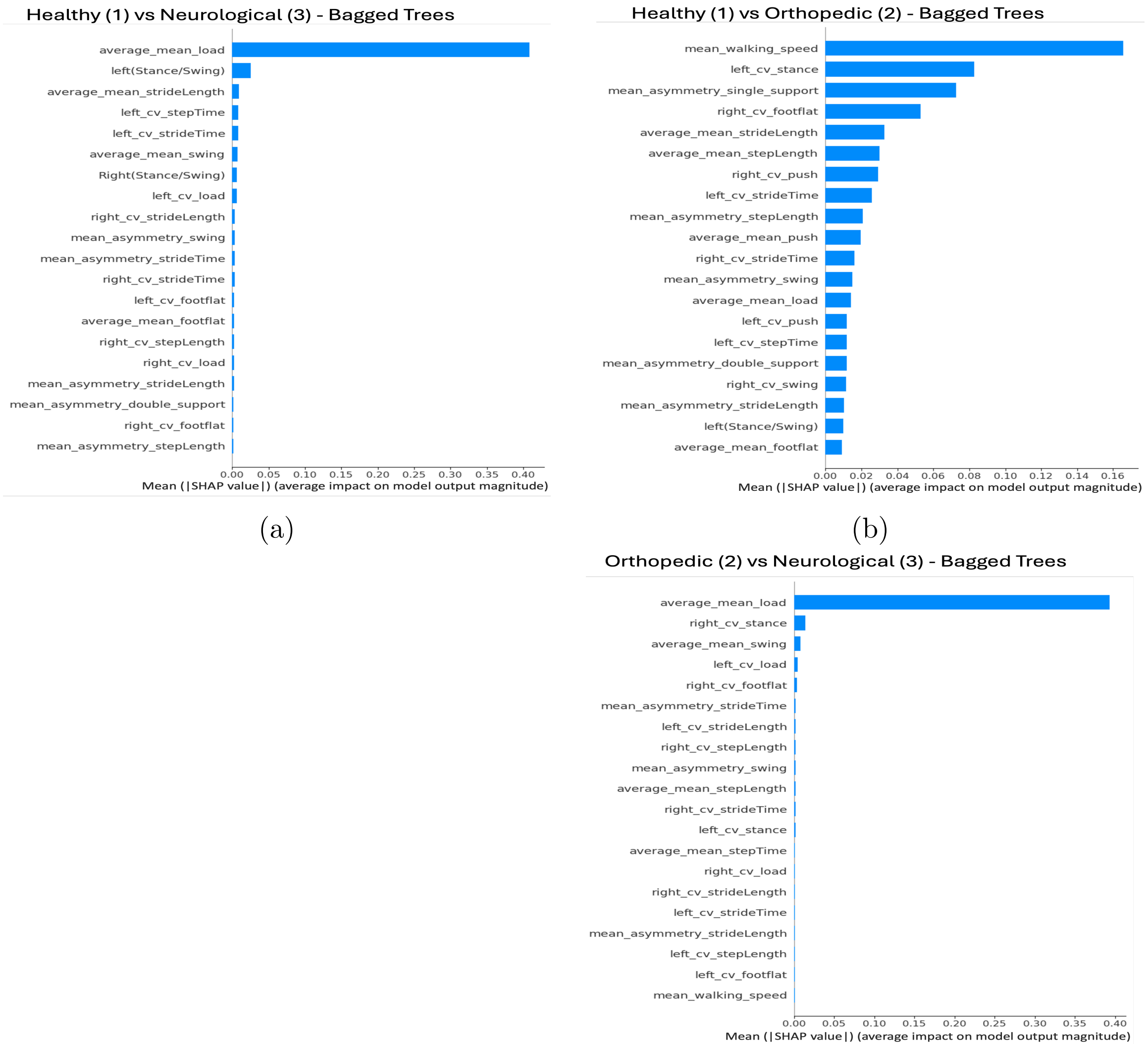


Figure 5.7: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.

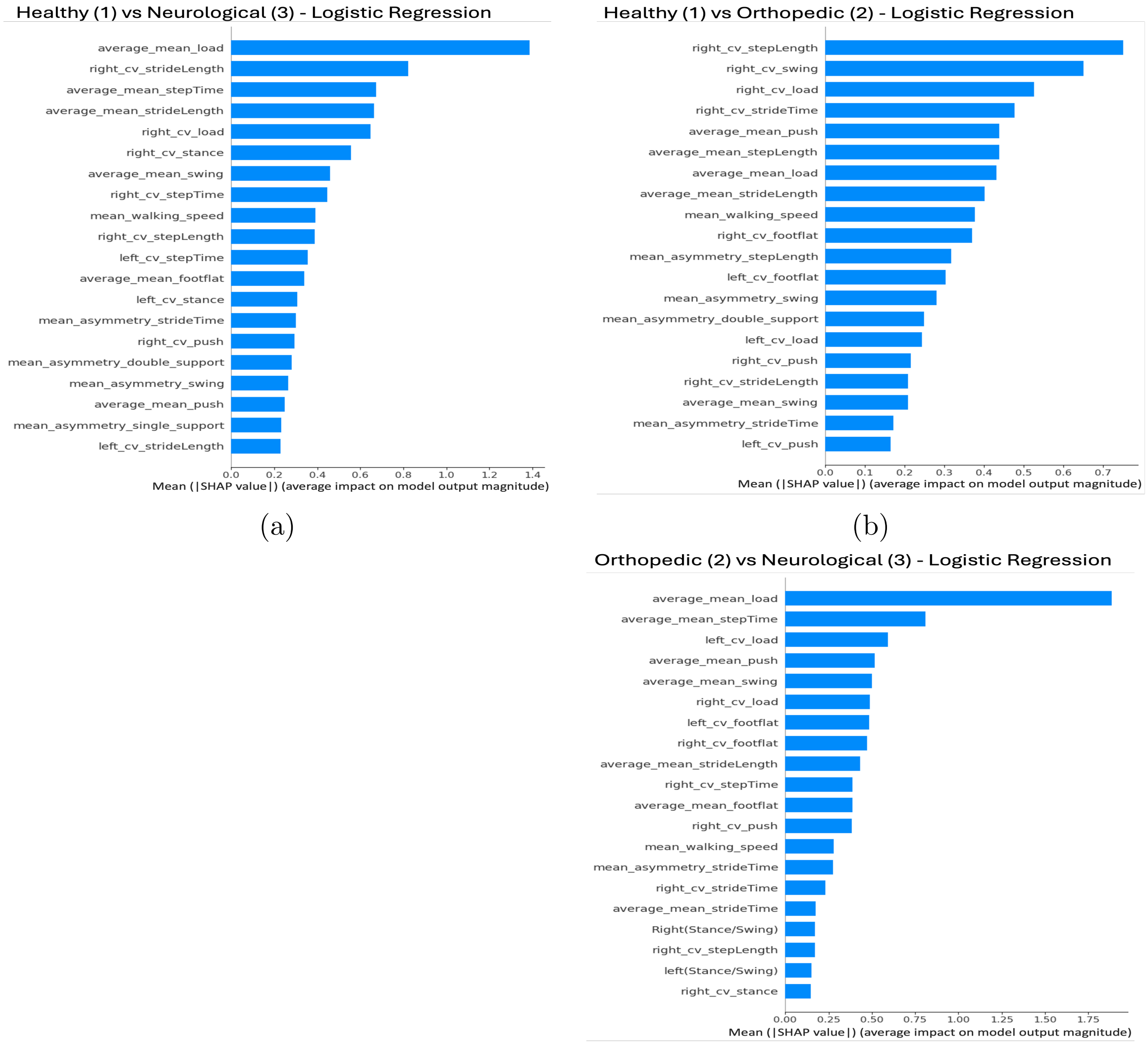


(a)

(b)

(c)

Figure 5.8: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.

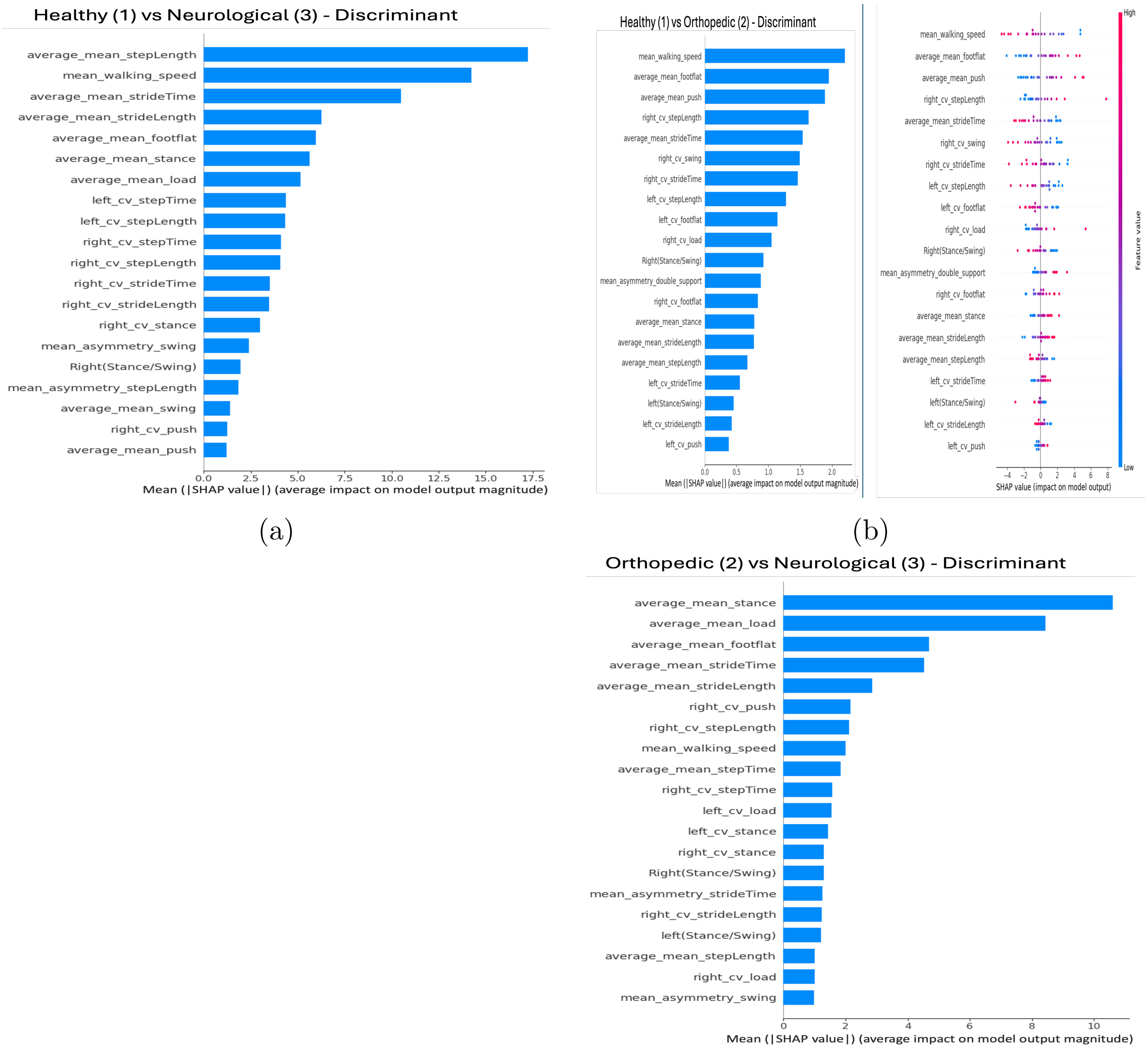


(a)

(b)

(c)

Figure 5.9: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.



(a)

(b)

(c)

Figure 5.10: Visualization of SHAP-based model explainability for benchmark machine learning (ML) models across the three cohorts. We performed a SHAP analysis based on our benchmarking ML models for investigating the consistency in feature importance among BART and the traditional ML models.

Chapter 6

Concluding Remarks

This thesis presents a unified framework that advances the intersection of statistical modeling and interpretable machine learning in the context of outlier detection and human gait analysis. Across three core chapters, we addressed challenges ranging from algorithmic robustness in industrial Internet-of-Things (IoT) applications to modeling and predicting gait variability in clinical and healthy populations. We shall now go chapter-by-chapter and discuss the concluding remarks that expands on our contributions, and also points out limitations and future directions for the addressed problems.

In Chapter 2, we acknowledged that significant efforts have been made to develop outlier detection algorithms in the IoT industry. However, most existing algorithms face a trade-off between accuracy and computational complexity. The widely used the R-PCA (Recursive Principal Component Analysis) algorithm provides a real-time and computationally efficient solution, but suffers from theoretical ambiguity, leading to reduced classification accuracy. In this study, we propose an improved outlier detection algorithm that encompasses accuracy, real-time detectability, reproducibility, and computational efficiency. Our proposed improvement focuses on the distributional assumption of the squared prediction error (SPE) scores in R-PCA. While R-PCA assumes Gaussianity for computational efficiency, it compromises accuracy in outlier detection. Through extensive simulations, we demonstrate the non-Gaussian nature of SPE scores and propose a new data-driven scheme that yields superior results compared to R-PCA. The reproducibility of the proposed scheme is emphasized, and we provide an improved version of R-PCA algorithm. Additionally, we discuss how this framework can be adapted to other PCA-based outlier algorithms that utilize either Rao's or Hawkins' test statistics for calculating the SPE scores. However, there are further issues worth considering. It should be noted that some PCA-based outlier detection schemes assuming Gaussianity, including the proposed method, are sensitive to the proportion of outliers. As a future direction, we aim to develop a more robust and cost-effective PCA-based outlier detection scheme capable of capturing a wider range of outliers or anomalies, even in higher-dimensional settings. This advancement would enhance the overall efficacy and applicability of outlier detection in the IoT-based systems.

In Chapter 3, our study utilizing Beta regression models has provided valuable insights into the complex relationship between demographic factors, gait parameters, and gait variability, contributing to a better understanding of the human gait index. By examining the main and interaction effects of various predictors on the GI, we have explained how age, gender, BMI, KA_{max}, $\text{SL}_{\text{norm(h)}}$, WS, and StPh/SwPh influence gait patterns. The applications of our findings extend to clinical practice, where clinicians can have informed decision-making processes, particularly in designing personalized rehabilitation programs to optimize gait outcomes. By tailoring interventions based on identified predictors of gait variability, clinicians can enhance mobility and functional outcomes for patients with gait abnormalities. Additionally, our study offers insights into monitoring and evaluating the effectiveness of interventions over time, thereby facilitating continuous improvement in patient care.

However, our study has some limitations. First, the study population comprised clinically healthy individuals, restricting the generalizability of findings to those with gait disorders or underlying medical conditions. Additionally, potential confounding factors such as comorbidities, medication use, walking conditions, surface characteristics, ethnicities, psychological factors, and cognitive function were not accounted for. Future research should address these limitations by including diverse populations and broader confounders to enhance generalizability. While our study provided valuable insights, relying solely on Beta regression models may limit capturing the full complexity of gait patterns. Alternative methods, such as quantile regression models or generalized additive models for location, scale, and shape (GAMLSS), could offer improved flexibility in handling bounded outcomes and heteroscedasticity. However, these methods face practical challenges, including increased computational complexity and interpretability issues [123], [224]. Future studies should evaluate these models' feasibility and computational efficiency while leveraging data from wearable sensors and long-term monitoring to enrich our understanding of gait dynamics.

In closing, our study advances the field of gait assessment by uncovering the intricate relationship between demographic factors, gait parameters, and gait variability using a single-model-based approach. By addressing specific factors contributing to GI variability and considering interaction effects among predictors, clinicians can develop targeted interventions to optimize gait outcomes and improve patient outcomes. Consequently, continued research in this area will further refine our understanding of gait patterns and enhance the effectiveness of interventions aimed at improving mobility and function in the general population, as well as patients with gait abnormalities.

Finally, in Chapter 4 and 5, our study provides an interpretable machine-learning framework that delivers patient-level predictions and offers explanations and insights into the model's diagnostics and associated predictions. Our BART-based framework generates interpretable predictions with bench-marked performance based on wearable sensors (IMU). BART is a tree-ensemble-based model that puts regulations on the tree depth in terms of a prior, which prevents overfitting and yields reliable results. This combination of accuracy and interpretability would enables physicians to receive highly reliable predictions while also gaining insight into the factors driving those predictions.

Another key focus of this work was to interpret the patient-level predictions by examining the individual contributions of each gait feature and providing biological insights into their importance. The average mean load, identified as the most important feature for distinguishing between the healthy and neurological cohorts in our gait analysis, reflects the physical stress exerted by individuals during their stride [216], [217]. Additionally, one observation from comparing the ALEs of the three variables is that the ranges of the ALE main effect function for the Average mean load feature were wider, meaning higher probability masses were assigned compared to the other features. So that concludes that the Average mean load Phase was the most important predictor, at least in terms of its main effects. This is consistent with the fact that the Average mean load Phase received the highest variable inclusion proportions in the model (as indicated by both the permutation-based test and the sparsity-induced prior-based model), as shown in the Figs. 4.5-4.10. An additional SHAP analysis conducted on the bench-marked ML models also confirms the ranking of this feature. A lower mean load in healthy individuals indicates efficient weight distribution and balance, which are critical for fluid, well-coordinated movements [218]. This non-linear impact—where a higher load aligns with neurological impairment—suggests that individuals with neurological disorders may experience altered gait mechanics due to weakened muscle function, poor motor coordination, or increased joint stress [175]. The elevated load in the Neurological cohort could be indicative of compensatory mechanisms, such as increased muscle force or joint pressure, that individuals use to maintain stability, often due to underlying motor deficits associated with neurological conditions [175]. Thus, the average mean load could serve as a valuable marker for identifying gait abnormalities linked to neurological disorders and tailoring interventions focused on improving load distribution and overall gait efficiency.

The permutation-based procedure also captured the Average Walking Speed and Average Stride length as important gait features for classifying healthy patients from the neurological cohort. The neurological cohort of this data was composed of 4 cohorts: hemispheric stroke, Parkinson's disease (PD), toxic peripheral neuropathy and radiation-induced leukoencephalopathy. Although we lack further information on the neurological subgroups, our results align with findings from previous leading studies on these subgroups. Previous Meta-analysis and systematic reviews [176], [177] revealed that walking speed, stride length, swing time and hip excursion were reduced in people with PD compared with healthy control. Previous review-based studies [179] showed that the average walking speed reported for subjects with stroke was less than the able-bodies subjects. The same conclusions were drawn for the stride length and cadence for such subjects based on 17 studies [219]. Various interventions, such as walking training with cadence cueing, can improve the average walking speed and stride length in stroke patients [179].

For the healthy vs orthopedic cohort, Average step length came out as one of the important features based on the permutation-based procedure (see, Fig. 4.7). The Average Step Length is a key indicator of lower limb function, often reflecting the ability to achieve full stride due to muscle strength, joint mobility, and balance [175].

In orthopedic patients, reduced step length is commonly observed and can signal limitations due to pain, joint stiffness, or compensatory gait patterns developed to avoid discomfort [180]. Clinically, a shorter step length might suggest weakened muscles around the hip, knee, or ankle joints, as well as balance challenges [181], [182]. Identifying this pattern allows clinicians to focus on exercises or interventions targeting flexibility and strength to restore a more natural gait pattern and improve stride efficiency. Additionally, average walking speed, a robust measure of functional mobility, endurance, and overall lower body strength [183], emerged as an important feature in both permutation-based and sparsity-induced prior approaches (see, Fig. 4.7 and Fig. 4.9). Reduced walking speed in orthopedic patients can indicate both physical limitations and cautionary behavior to minimize joint impact or avoid falls [184]. In practice, a higher walking speed is often linked to better balance and dynamic stability, enabling more efficient and safe ambulation [185]. Clinicians often use walking speed as a primary metric to assess the impact of orthopedic interventions, as improvements here typically correlate with enhanced patient confidence, reduced pain, and better musculoskeletal health [220]. Therefore, emphasizing walking speed in orthopedic patients' treatment and rehabilitation programs can support improvements in overall mobility and independence. Also, The Average Mean Asymmetry in Single Support, identified as an important feature by the permutation-based approach, is clinically significant as it reflects the body's ability to distribute weight evenly during the gait cycle, a key factor in stable and efficient movement [186], [187]. Lower asymmetry levels in single support indicate balanced weight-bearing, which is typical in healthy individuals with optimal joint function and muscle coordination [190]. In contrast, increased asymmetry suggests possible orthopedic issues, such as joint pain, muscle weakness, or postural imbalances, which often lead patients to favour one leg over the other to minimize discomfort or reduce strain and this adaptation can worsen over time, potentially leading to further musculoskeletal imbalances [191]. As a result, monitoring asymmetry levels can provide valuable insights into the severity of orthopedic impairment and may guide targeted interventions aimed at restoring symmetry, thereby enhancing balance, reducing strain on compensatory muscles, and improving overall gait stability.

Finally, we investigated the feature importance for the orthopedic cohort against the neurological disorder cohort and the Average mean load Phase come out to be as the most important one. In this case, the average mean load is a key indicator of neuromuscular control and stability during movement. As observed in the Neurological cohort, elevated mean load values likely reflect compensatory adjustments in response to motor deficits, such as increased muscle force or joint pressure, used to maintain balance. This trend differentiates neurological gait patterns from orthopedic ones, where structural limitations affect step dynamics rather than load. Additionally, the sparsity-induced prior-based BART framework also identified alteration in gait variability in certain gait parameters related to the loading phases. ALE plots revealed prolonged variation in flat feet (left foot), stride time (both feet), and stance phase (left) were associated with the neurological cohort (see Fig. 4.14). Flat foot is often associated with pain and can significantly impact walking speed and balance, increasing the risk of falls [188], [189]. This was consistent with observations in the healthy vs. neurological cohort, where Fig. 4.11 indicated that the neurological cohort had a slower average walking speed and a shorter average stride length. Additionally, the risk of falls is higher in PD patients due to gait freezing, which is associated with increased variability in foot flatness. However, the importance of these features was largely overshadowed by the average loading phase, resulting in a much narrower range for the individual (ALE) main effect function compared to that of the average loading phase. Identifying such load discrepancies can guide interventions aimed at improving stability and reducing compensatory strain in neurological conditions. Another observation from comparing the individual effects (through the Accumulated Local Effects (ALE) plots) of the variables is that the range of the individual (ALE) main effect function for the Average mean load Phase feature was wider, indicating higher probability masses assigned compared to the other features in both the healthy vs neurological and orthopedic vs neurological cohorts. This suggests that the Average mean load Phase was the most important predictor, at least in terms of its main effects. This finding is consistent with the fact that the Average mean load Phase received the highest variable inclusion proportions in the model, as indicated by both the permutation-based test and the sparsity-induced prior-based model, as shown in the chapter 4, for both classification cohorts. Additionally, SHAP analysis based on the selected ML model revealed that the Average mean load Phase ranked as the most important feature (based on the mean absolute SHAP value) in determining the model's prediction (see Fig. 5.4-5.10) for these two cohorts.

We acknowledge that this study has limitations and that further investigation are needed to enhance the understanding of wearable-sensor-based gait predictions for both healthy and pathological cohorts. Firstly, the IOPL dataset [142] lacks information on cohort sizes and demographic characteristics of the four sub-cohorts within the neurological disorder group and the two sub-cohorts within the orthopedic disorder group. Conducting a subgroup analysis would offer better insights into the nuances of patient-level gait characteristics among these pathological sub-cohorts and the healthy population. Secondly, difference analysis revealed statistically significant differences in some demographic variables, such as age and BMI, across the three groups. Our study did not account for age-specific or BMI-specific effects (fixed or random) in the model, which may influence the predictions. Thirdly, we acknowledge limitations in

BART's performance for the healthy vs. orthopedic classification, which lagged behind the performance of the other two classifications. This may be due to high noise within this cohort's data or substantial collinearity among predictors. Although the identified feature importance was biologically plausible, it did not align as consistently with the SHAP analysis as it did for the other two classifications. Additionally, our work has not addressed the uncertainty in permutation-based feature importance for BART in classification tasks, which could be an interesting area for future research. Also, we did not consider a "causal" approach in our study to understand the gait features and their effects in the classification of healthy and pathological groups using BART. Future research should focus on exploring the causality of various gait disorders and their nuances compared to healthy gait patterns using wearable-sensor data.

In conclusion, this study focuses on understanding gait predictions for healthy and different pathological groups based on wearable sensors. Data-driven gait predictions for patient outcomes are being developed and applied more frequently [153], [155], [156], [157], [158]. However, black-box models that generate gait predictions without proper explanations are difficult for physicians to trust, as they offer little direction for informed patient care. These predictions serve the purpose of identifying optimal courses of action without necessitating an in-depth understanding of the underlying mechanisms. BART not only explains the decision paths for these predictions but also assigns credit to each input feature (via VIP) contributing to the prediction, capturing various aspects of local explanations. This enables us to understand the model's global behavior during prediction. Our work is an introspective study on using interpretable ML models like BART, which go beyond linearity and mere prediction, to understand patient-level gait characteristics that differentiate healthy subjects from those in a disorder group.

Appendix A

Chapter 3

The log-likelihood function for the class of beta regression can be written as follows:

$$l_i(\beta, \theta) = \sum_{i=1}^{n} l_i(\mu_i, \phi_i), \tag{A.0.1}$$

where

$$l_i(\mu_i, \phi_i) = \log\Gamma(\phi_i) - \log\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i - 1)\log(y_i) + [(1-\mu_i)\phi_i - 1]\log(1-y_i). \tag{A.0.2}$$

The components of the score vector derived from differentiating the log-likelihood function for $r = 1, \ldots, k$ are as follows

$$U_r(\beta, \theta) = \frac{\partial l(\beta, \theta)}{\partial \beta_r} = \sum_{i=1}^{n} \phi_i(y_i^* - \mu_i^*)\frac{d\mu_i}{d\eta_{1i}}\frac{d\eta_{1i}}{d\beta_k}, \tag{A.0.3}$$

where

$$\frac{d\mu_i}{d\eta_{1i}} = \frac{1}{g_1'(\mu_i)}, \quad y_i^* = \log\frac{y_i}{1-y_i}, \quad \mu_i^* = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i), \tag{A.0.4}$$

and $\psi(\cdot)$ is the digamma function. The other component of the score vector based on the precision sub-model is:

$$U_R(\beta,\theta) = \frac{\partial l(\beta,\theta)}{\partial\theta_R} = \sum_{i=1}^{n}\left\{\mu_i(y_i^*-\mu_i^*) + \psi(\phi_i) - \psi((1-\mu_i)\phi_i) + \log(1-y_i)\right\}\frac{d\phi_i}{d\eta_{2i}}\frac{d\eta_{2i}}{d\theta_R}, \tag{A.0.5}$$

where $\frac{d\phi_i}{d\eta_{2i}} = \frac{1}{g_2'(\phi_i)}$, and $R = 1,\ldots,h$. The vectors are defined as $y^* = (y_1^*,\ldots,y_n^*)^T$, $\mu^* = (\mu_1^*,\ldots,\mu_n^*)^T$, and the matrices are given by

$$T_1 = \mathrm{diag}\left(\frac{d\mu_i}{d\eta_{1i}}\right), \quad T_2 = \mathrm{diag}\left(\frac{d\phi_i}{d\eta_{2i}}\right), \quad \Phi = \mathrm{diag}(\phi_i),$$

with $\mathrm{diag}(\mu_i)$ denoting the $n\times n$ diagonal matrix with elements μ_i, for $i = 1,\ldots,n$. Additionally,

$$\upsilon_i = \mu_i(y_i^*-\mu_i^*) + \psi(\phi_i) - \psi((1-\mu_i)\phi_i) + \log(1-y_i).$$

Hence, we can write the $(k+h)\times 1$ dimensional score vector $U(\zeta)$ as

$$U(\zeta) = \begin{bmatrix} U_\beta(\beta,\theta)^T \\ U_\theta(\beta,\theta)^T \end{bmatrix},$$

where

$$U_\beta(\beta,\theta) = \tilde{X}^T\Phi T_1(y^*-\mu^*), \tag{A.0.6}$$

$$U_\theta(\beta, \theta) = \tilde{Z}^T T_2 \upsilon. \tag{A.0.7}$$

Linear Beta Regression Model with Fixed Precision: For a linear regression model with fixed precision parameters, we define

$$g_1(\mu_i) = \eta_{1i} = x_i^T \beta, \quad g_2(\phi_i) = \eta_{2i} = \phi_i = \phi,$$

where $\phi > 0$ is a constant, leading to $\tilde{X} = X$ and $\tilde{Z} = 1$. Here, X represents the matrix of covariates with rows given by x_i^T, and the parameters $\beta \in \mathbb{R}^k$ and $\phi \in (0, \infty)$. Consequently, the score vector simplifies as

$$U_\beta(\beta, \theta) = \phi X^T T(y^* - \mu^*), \tag{A.0.8}$$

$$U_\phi(\beta, \theta) = \sum_{i=1}^{n} \upsilon_i, \tag{A.0.9}$$

where $T = \operatorname{diag}\left(\frac{d\mu_i}{d\eta_i}\right)$, and y_i^*, μ_i^* and υ_i are as defined in the "Likelihood and Method of Estimation" section.

Linear Beta Regression Model with Variable Dispersion: This entails using the same expressions

$$g_1(\mu_i) = \eta_{1i} = x_i^T \beta, \quad g_2(\phi_i) = \eta_{2i} = z_i^T \theta,$$

where $\beta \in \mathbb{R}^k$ and $\theta \in \mathbb{R}^h$. In this model, the matrices $\tilde{X}$ and $\tilde{Z}$, are equivalent to X and Z respectively, where X represents the matrix of covariates with rows given by x_i^T, and Z represents the matrix of covariates with rows given by z_i^T. The score vector remains identical to the one described in Eqs. (A.0.6) and (A.0.7).

Table A.1: Model discrimination

Measure for model discrimination	VDBR	FPBR	Logit-transformed Linear Regression
AIC	-774.30	-650.20	-252.55
BIC	-718.50	-613.96	-216.31

Bibliography

[1] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[2] T. Bouwmans and E. H. Zahzah, “Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance,” *Computer Vision and Image Understanding*, vol. 122, pp. 22–34, 2014.

[3] A. Das, S. Misra, S. Joshi, J. Zamberno, G. Memik and A. Choudhary, “An efficient FPGA implementation of principle component analysis based network Intrusion Detection System,” *Design, Automation and Test in Europe*, pp. 1160–1165, 2008.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[5] M. F. Elrawy, T. K. Abdelhamid, and A. M. Mohamed, “IDs in telecommunication network using PCA,” *International Journal of Computer Networks & Communications*, vol. 5, no. 4, pp. 147–157, 2013.

[6] S. C. Chan, H. C. Wu, and K. M. Tsui, “Robust recursive eigendecomposition and subspace-based algorithms with application to fault detection in wireless sensor

networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1703–1718, 2012.

[7] R. Gnanadesikan and J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, vol. 28, no. 1, p. 81, 1972.

[8] F. Harrou, Y. Sun, and S. Khadraoui, "Amalgamation of anomaly-detection indices for enhanced process monitoring," *Journal of Loss Prevention in the Process Industries*, vol. 40, pp. 365–377, 2016.

[9] D. M. Hawkins, *Identification of Outliers*. London: Chapman and Hall, 1980.

[10] D. M. Hawkins and L. P. Fatti, "Exploring multivariate data using the minor principal components," *The Statistician*, vol. 33, no. 4, p. 325-338, 1984.

[11] H. Huang, H. Al-Azzawi, and H. Brani. "Network traffic anomaly detection," *arXiv preprint*, 2014.

[12] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: A new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[13] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[14] R. Kwitt and U. Hofmann, "Unsupervised anomaly detection in network traffic by means of robust PCA," in Proc. *Int. Multi-Conf. on Computing in the Global Information Technology (ICCGI'07)*, pp. 37-37, 2007.

[15] A. Lakhina, M. Crovella, and C. Diot, “Diagnosing network-wide traffic anomalies,” in *Proc. of the 2004 conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 219-230, 2004.

[16] Y. Liu, L. Zhang, and Y. Guan, “Sketch-based streaming PCA algorithm for network-wide traffic anomaly detection,” in *Proc. IEEE 30th Int. Conf. Distributed Computing Systems.(ICDCS)*, Genova, Italy, 2010, pp. 807–816.

[17] B. Mertens, M. Thompson, and T. Fearn, “Principal component outlier detection and Simca: A synthesis,” *The Analyst*, vol. 119, no. 12, pp. 2777-2784, 1994.

[18] H. Om and A. Kundu, “A hybrid system for reducing the false alarm rate of anomaly intrusion detection system,” in 2012 1st International Conference on Recent Advances in Information Technology (RAIT), pp. 131-136, 2012.

[19] C. R. Rao, “The use and interpretation of principal component analysis in applied research,” *Sankhya*, Series A (1961-2002), pp. 329–358, 1964.

[20] F. E. Satterthwaite, “An approximate distribution of estimates of variance components,” *Biometrics Bulletin*, vol. 2, no. 6, p. 110, 1946.

[21] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang, “Principal component-based anomaly detection scheme,” *Foundations and Novel Approaches in Data Mining*, pp. 311–329, 2006.

[22] T. Yu, X. Wang, and A. Shami, “Recursive principal component analysis-based data outlier detection and sensor data aggregation in IOT Systems,” *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2207–2216, 2017.

[23] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1-13, 2020.

[24] D. Chicco, "Ten quick tips for machine learning in Computational biology," *BioData Mining*, vol. 10, no. 1, p. 35, 2017.

[25] Y. Qiao, B. Zhang, and Z. Zhang, "Unsupervised anomaly detection for IOT data based on robust adversarial learning," in *Proc. IEEE 24th Int Conf. on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems &amp; Application (HPCC/DSS/SmartCity/DependSys)*, pp. 2324-2330, 2022.

[26] P. Zhou and J. Feng, "Outlier-robust tensor PCA," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2263-2271.

[27] H. Wang, G. Ni, J. Chen, and J. Qu, "Research on rolling bearing state health monitoring and life prediction based on PCA and internet of things with multi-sensor," *Measurement*, vol. 157, p. 107657, 2020.

[28] A. I. Faisal, T. Mondal, and M. J. Deen, "Systematic development of a simple human gait index," *IEEE Reviews in Biomedical Engineering*, pp. 1–13, 2023.

[29] M. A. Samara, I. Bennis, A. Abouaissa, and P. Lorenz, "A survey of outlier detection techniques in IOT: Review and Classification," *Journal of Sensor and Actuator Networks*, vol. 11, no. 1, p. 4, 2022.

[30] I. T. Jolliffe, "*Principal Component Analysis for Special Types of Data*," *Springer* New York, pp. 338-372, 2002..

[31] J. Liu, J. Wang, X. Liu, T. Ma, and Z. Tang, “MWRSPCA: Online fault monitoring based on moving window recursive sparse principal component analysis,” *Journal of Intelligent Manufacturing*, vol. 33, no. 5, pp. 1255–1271, 2021.

[32] J. Nonnekes, R. J. M. Goselink, E. Růžička, A. Fasano, J. G. Nutt, and B. R. Bloem, “Neurological disorders of gait, balance and posture: a sign-based approach,” *Nat. Rev. Neurol.*, vol. 14, no. 3, pp. 183–189, Mar. 2018, doi: 10.1038/nrneurol.2017.178.

[33] S. Subramaniam, S. Majumder, A. I. Faisal, and M. J. Deen, “Insole-based systems for health monitoring: current solutions and research challenges,” *Sensors*, vol. 22, no. 2, p. 438, Jan. 2022, doi: 10.3390/s22020438.

[34] F. Horst, S. Lapuschkin, W. Samek, K. R. Müller, and W. I. Schöllhorn, “Explaining the unique nature of individual gait patterns with deep learning,” *Sci. Rep.*, 2019, doi: 10.1038/s41598-019-38748-8.

[35] S. Majumder et al., “Smart homes for elderly healthcare—recent advances and research challenges,” *Sensors*, vol. 17, no. 11, p. 2496, Oct. 2017, doi: 10.3390/s17112496.

[36] P. Mandal, K. Tank, T. Mondal, C.-H. Chen, and M. J. Deen, “Predictive walking-age health analyzer,” *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 2, pp. 363–374, Mar. 2018, doi: 10.1109/JBHI.2017.2666603.

[37] N. Agoulmine, M. J. Deen, J.-S. Lee, and M. Meyyappan, “U-health smart home,” *IEEE Nanotech Mag.*, vol. 5, no. 3, pp. 6–11, Sep. 2011, doi: 10.1109/MNANO.2011.941951.

[38] Bo Jin et al., “Walking-age analyzer for healthcare applications,” *IEEE J.*

Biomed. Heal. Informatics, vol. 18, no. 3, pp. 1034–1042, May 2014, doi: 10.1109/JBHI.2013.2296873.

[39] S. Majumder, T. Mondal, and M. J. Deen, "A simple, low-cost and efficient gait analyzer for wearable healthcare applications," *IEEE Sens. J.*, vol. 19, no. 6, pp. 2320–2329, Mar. 2019, doi: 10.1109/JSEN.2018.2885207.

[40] M. Naghshvarianjahromi, S. Majumder, S. Kumar, N. Naghshvarianjahromi, and M. J. Deen, "Natural brain-inspired intelligence for screening in healthcare applications," *IEEE Access*, vol. 9, pp. 67957–67973, 2021, doi: 10.1109/ACCESS.2021.3077529.

[41] L. Filli et al., "Profiling walking dysfunction in multiple sclerosis: characterisation, classification and progression over time," *Sci. Rep.*, vol. 8, no. 1, p. 4984, Dec. 2018, doi: 10.1038/s41598-018-22676-0.

[42] A. I. Faisal, S. Majumder, T. Mondal, D. Cowan, S. Naseh, and M. J. Deen, "Monitoring methods of human body joints: state-of-the-art and research challenges," *Sensors*, vol. 19, no. 11, p. 2629, Jun. 2019, doi: 10.3390/s19112629.

[43] B. R. Bloem et al., "Measurement instruments to assess posture, gait, and balance in Parkinson's disease: Critique and recommendations," *Mov. Disord.*, vol. 31, no. 9, pp. 1342–1355, Sep. 2016, doi: 10.1002/mds.26572.

[44] A. Shcherbina et al., "The effect of digital physical activity interventions on daily step count: a randomised controlled crossover substudy of the MyHeart Counts Cardiovascular Health Study," *Lancet Digit. Heal.*, vol. 1, no. 7, pp. e344–e352, Nov. 2019, doi: 10.1016/S2589-7500(19)30129-3.

[45] J. Addington et al., "Clinical and functional characteristics of youth at clinical

high-risk for psychosis who do not transition to psychosis," *Psychol. Med.*, vol. 49, no. 10, pp. 1670–1677, Jul. 2019, doi: 10.1017/S0033291718002258.

[46] S. Majumder and M. J. Deen, "Wearable IMU-based system for real-time monitoring of lower-limb joints," *IEEE Sens. J.*, vol. 21, no. 6, pp. 8267–8275, Mar. 2021, doi: 10.1109/JSEN.2020.3044800.

[47] W. Jiang et al., "A wearable tele-health system towards monitoring COVID-19 and chronic diseases," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 61–84, 2022, doi: 10.1109/RBME.2021.3069815.

[48] S. Majumder and M. J. Deen, "A Robust Orientation Filter for Wearable Sensing Applications," *IEEE Sens. J.*, vol. 20, no. 23, pp. 14228–14236, Dec. 2020, doi: 10.1109/JSEN.2020.3009388.

[49] A. I. Faisal, T. Mondal, and M. J. Deen, "Systematic Development of a Simple Human Gait Index," *IEEE Rev. Biomed. Eng.*, vol. 17, pp. 229–242, 2024, doi: 10.1109/RBME.2023.3279655.

[50] S. B. Gonçalves, S. B. C. Lama, and M. T. da Silva, "Three decades of gait index development: A comparative review of clinical and research gait indices," *Clin. Biomech.*, vol. 96, p. 105682, Jun. 2022, doi: 10.1016/j.clinbiomech.2022.105682.

[51] M. L. McMulkin and B. A. MacWilliams, "Application of the Gillette Gait Index, Gait Deviation Index and Gait Profile Score to multiple clinical pediatric populations," *Gait Posture*, vol. 41, no. 2, pp. 608–612, Feb. 2015, doi: 10.1016/j.gaitpost.2015.01.005.

[52] M. H. Schwartz and A. Rozumalski, “The gait deviation index: a new comprehensive index of gait pathology,” *Gait Posture*, vol. 28, no. 3, pp. 351–357, Oct. 2008, doi: 10.1016/j.gaitpost.2008.05.001.

[53] L. Wang, Y. Sun, Q. Li, T. Liu, and J. Yi, “IMU-based gait normalcy index calculation for clinical evaluation of impaired gait,” *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 1, pp. 3–12, Jan. 2021, doi: 10.1109/JBHI.2020.2982978.

[54] V. Rota, L. Perucca, A. Simone, and L. Tesio, “Walk ratio (step length/cadence) as a summary index of neuromotor control of gait,” *Int. J. Rehabil. Res.*, vol. 34, no. 3, pp. 265–269, Sep. 2011, doi: 10.1097/MRR.0b013e328347be02.

[55] A. Shumway-Cook, C. S. Taylor, P. N. Matsuda, M. T. Studer, and B. K. Whetten, “Expanding the Scoring System for the Dynamic Gait Index,” *Phys. Ther.*, vol. 93, no. 11, pp. 1493–1506, Nov. 2013, doi: 10.2522/ptj.20130035.

[56] L. M. Schutte, U. Narayanan, J. L. Stout, P. Selber, J. R. Gage, and M. H. Schwartz, “An index for quantifying deviations from normal gait,” *Gait Posture*, vol. 11, no. 1, pp. 25–31, Feb. 2000, doi: 10.1016/S0966-6362(99)00047-8.

[57] M.-J. Chung and M.-J. J. Wang, “The change of gait parameters during walking at different percentage of preferred walking speed for healthy adults aged 20–60 years,” *Gait Posture*, vol. 31, no. 1, pp. 131–135, Jan. 2010, doi: 10.1016/j.gaitpost.2009.09.013.

[58] R. W. Bohannon and A. Williams Andrews, “Normal walking speed: a descriptive meta-analysis,” *Physiotherapy*, vol. 97, no. 3, pp. 182–189, Sep. 2011, doi: 10.1016/j.physio.2010.12.004.

[59] N. Veronese et al., “Association Between Gait Speed With Mortality, Cardiovascular Disease and Cancer: A Systematic Review and Meta-analysis of Prospective Cohort Studies,” *J. Am. Med. Dir. Assoc.*, vol. 19, no. 11, pp. 981-988.e7, Nov. 2018, doi: 10.1016/J.JAMDA.2018.06.007.

[60] D. Hamacher, N. B. Singh, J. H. Van Dieën, M. O. Heller, and W. R. Taylor, “Kinematic measures for assessing gait stability in elderly individuals: a systematic review,” *J. R. Soc. Interface*, vol. 8, no. 65, p. 1682, Dec. 2011, doi: 10.1098/RSIF.2011.0416.

[61] J. M. Hausdorff, “Gait dynamics in Parkinson’s disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling,” *Chaos An Interdiscip. J. Nonlinear Sci.*, vol. 19, no. 2, p. 026113, Jun. 2009, doi: 10.1063/1.3147408.

[62] B. Van Gheluwe, K. A. Kirby, and F. Hagman, “Effects of simulated genu valgum and genu varum on ground reaction forces and subtalar joint function during gait,” *J. Am. Podiatr. Med. Assoc.*, vol. 95, no. 6, pp. 531–541, Nov. 2005, doi: 10.7547/0950531.

[63] M. Gnucci, M. Flemma, M. Tiberti, M. Ricci, A. Pallotti, and G. Saggio, “Assessment of gait harmony in older and young people,” in *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, SCITEPRESS - Science and Technology Publications, 2018, pp. 155–160. doi: 10.5220/0006572701550160.

[64] M. Iosa et al., “The golden ratio of gait harmony: repetitive proportions of repetitive gait phases,” *Biomed Res. Int.*, vol. 2013, pp. 1–7, 2013, doi: 10.1155/2013/918642.

[65] V. Mikos et al., "Regression analysis of gait parameters and mobility measures in a healthy cohort for subject-specific normative values," *PLoS One*, vol. 13, no. 6, p. e0199215, Jun. 2018, doi: 10.1371/journal.pone.0199215.

[66] F. Wahid, R. Begg, N. Lythgo, C. J. Hass, S. Halgamuge, and D. C. Ackland, "A multiple regression approach to normalization of spatiotemporal gait features," *J. Appl. Biomech.*, vol. 32, no. 2, pp. 128–139, Apr. 2016, doi: 10.1123/jab.2015-0035.

[67] R. Kaur, Z. Chen, R. Motl, M. E. Hernandez, and R. Sowers, "Predicting multiple sclerosis from gait dynamics using an instrumented treadmill: a machine learning approach," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 9, pp. 2666–2677, Sep. 2021, doi: 10.1109/TBME.2020.3048142.

[68] A. Rohan, M. Rabah, T. Hosny, and S.-H. Kim, "Human pose estimation-based real-time gait analysis using convolutional neural network," *IEEE Access*, vol. 8, pp. 191542–191550, 2020, doi: 10.1109/ACCESS.2020.3030086.

[69] D. Slijepcevic et al., "Explaining machine learning models for clinical gait analysis," *ACM Trans. Comput. Healthc.*, vol. 3, no. 2, pp. 1–27, Apr. 2022, doi: 10.1145/3474121.

[70] E. J. Harris, I.-H. Khoo, and E. Demircan, "A survey of human gait-based artificial intelligence applications," *Front. Robot. AI*, vol. 8, pp. 1–28, Jan. 2022, doi: 10.3389/frobt.2021.749274.

[71] A. I. Faisal, T. Mondal, D. Cowan, and M. J. Deen, "Characterization of knee and gait features from a wearable tele-health monitoring system," *IEEE Sens. J.*, vol. 22, no. 6, pp. 4741–4753, Mar. 2022, doi: 10.1109/JSEN.2022.3146617.

[72] S. Shalev-Shwartz and S. Ben-David, "*Understanding machine learning: From theory to algorithms*," Cambridge University Press, 2014, doi: 10.1017/CBO9781107298019.

[73] A. Spanos, "Statistical inference: an overview," in *International Encyclopedia of Statistical Science*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1433–1439. doi: 10.1007/978-3-642-04898-2542.

[74] B. Efron and T. Hastie, *Computer Age Statistical Inference*. Cambridge University Press, 2016. doi: 10.1017/CBO9781316576533.

[75] D. Bzdok, "Classical Statistics and Statistical Learning in Imaging Neuroscience," *Front. Neurosci.*, vol. 11, no. OCT, p. 273651, Mar. 2016, doi: 10.3389/FNINS.2017.00543/BIBTEX.

[76] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nat. Methods*, vol. 15, no. 4, pp. 233–234, Apr. 2018, doi: 10.1038/nmeth.4642.

[77] S. Majumder, T. Mondal, and M. J. Deen, "Wearable sensors for remote health monitoring," *Sensors*, vol. 17, no. 12, p. 130, Jan. 2017, doi: doi: 10.3390/s17010130.

[78] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," *Sensors*, vol. 19, no. 9, p. 2164, May 2019, doi: 10.3390/s19092164.

[79] A. I. Faisal, "Development of a low-cost and easy-to-use wearable knee joint monitoring system," McMaster University, 2020. Accessed: Jun. 26, 2020. [Online]. Available: `https://macsphere.mcmaster.ca/handle/11375/25401`

[80] H.-J. Lee, W. H. Chang, B.-O. Choi, G.-H. Ryu, and Y.-H. Kim, "Age-related differences in muscle co-activation during locomotion and their relationship with

gait speed: a pilot study," *BMC Geriatrics*, vol. 17, no. 1, p. 44, Dec. 2017, doi: 10.1186/s12877-017-0417-4.

[81] A. Aboutorabi, M. Arazpour, M. Bahramizadeh, S. W. Hutchins, and R. Fadayevatan, "The effect of aging on gait parameters in able-bodied older subjects: a literature review," *Aging Clinical and Experimental Research*, vol. 28, no. 3, pp. 393–405, Jun. 2016, doi: 10.1007/s40520-015-0420-6.

[82] M. J. Deen, "Information and communications technologies for elderly ubiquitous healthcare in a smart home," *Personal and Ubiquitous Computing*, vol. 19, no. 3–4, pp. 573–599, Jul. 2015, doi: 10.1007/s00779-015-0856-x.

[83] K. Nisar, A. A. A. Ibrahim, L. Wu, A. Adamov, and M. J. Deen, "Smart home for elderly living using wireless sensor networks and an Android application," in *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, Oct. 2016, pp. 1–8, doi: 10.1109/ICAICT.2016.7991655.

[84] A. I. Faisal et al., "A simple, low-cost multi-sensor-based smart wearable knee monitoring system," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 1–1, Mar. 2020, doi: 10.1109/JSEN.2020.3044784.

[85] K. M. T. Goutier, S. L. Jansen, C. G. C. Horlings, U. M. Kung, and J. H. J. Allum, "The influence of walking speed and gender on trunk sway for the healthy young and older adults," *Age and Ageing*, vol. 39, no. 5, pp. 647–650, Sep. 2010, doi: 10.1093/ageing/afq066.

[86] G. T. Harding, C. L. Hubley-Kozey, M. J. Dunbar, W. D. Stanish, and J. L.

Astephen Wilson, “Body mass index affects knee joint mechanics during gait differently with and without moderate knee osteoarthritis,” *Osteoarthritis and Cartilage*, vol. 20, no. 11, pp. 1234–1242, Nov. 2012, doi: 10.1016/j.joca.2012.08.004.

[87] “Why weight matters when it comes to joint pain - Harvard Health.” Accessed: Jan. 08, 2020. [Online]. Available: `https://www.health.harvard.edu/pain/why-weight-matters-when-it-comes-to-joint-pain`

[88] M. Akram, E. Cerin, K. E. Lamb, and S. R. White, “Modelling count, bounded and skewed continuous outcomes in physical activity research: beyond linear regression models,” *International Journal of Behavioral Nutrition and Physical Activity*, vol. 20, no. 1, pp. 1–11, Dec. 2023, doi: 10.1186/S12966-023-01460-Y/FIGURES/3.

[89] R. Kieschnick and B. D. McCullough, “Regression analysis of variates observed on (0, 1): percentages, proportions and fractions,” *Statistical Modelling*, vol. 3, no. 3, pp. 193–213, Oct. 2003, doi: 10.1191/1471082X03st053oa.

[90] N. L. Johnson, “Systems of frequency curves generated by methods of translation,” *Biometrika*, vol. 36, no. 1/2, p. 149, Jun. 1949, doi: 10.2307/2332539.

[91] A. C. Atkinson, *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Clarendon Press, 1985.

[92] M. J. Barclay, C. W. Smith, and R. L. Watts, “The determinants of corporate leverage and dividend policies,” *Journal of Applied Corporate Finance*, vol. 7, no. 4, pp. 4–19, Jan. 1995, doi: 10.1111/j.1745-6622.1995.tb00259.x.

[93] N. L. Johnson, S. Kotz, and N. Balakrishnan, “Chapter 21: Beta distributions,” *Continuous Univariate Distributions, Volume 2*, 1995.

[94] J. Aitchison and S. M. Shen, "Logistic-normal distributions: some properties and uses," *Biometrika*, vol. 67, no. 2, p. 261, Aug. 1980, doi: 10.2307/2335470.

[95] C. Cox, "Nonlinear quasi-likelihood models: applications to continuous proportions," *Computational Statistics & Data Analysis*, vol. 21, no. 4, pp. 449–461, Apr. 1996, doi: 10.1016/0167-9473(95)00024-0.

[96] G. S. Maddala, "A perspective on the use of limited-dependent and qualitative variables models in accounting research," *Accounting Review*, vol. 66, no. 4, pp. 788–807, 1991. Available: `https://www.jstor.org/stable/248156`

[97] M. Hviid and B. Villadsen, "Beta distributed market shares in a spatial model with an application to the market for audit services," *Review of Industrial Organization*, vol. 10, no. 6, pp. 737–747, Dec. 1995, doi: 10.1007/BF010243.

[98] R. C. Mittelhammer, *Mathematical Statistics for Economics and Business.* New York, Springer, 1996. doi: 10.1007/978-1-4612-3988-8.

[99] P. McCullagh and J. A. Nelder, *Generalized Linear Models.* Routledge, 2019. doi: 10.1201/9780203753736.

[100] J. Brehm and S. Gates, "Donut shops and speed traps: evaluating models of supervision on police behavior," *Am. J. Pol. Sci.*, vol. 37, no. 2, p. 555, May 1993, doi: 10.2307/2111384.

[101] P. Paolino, "Maximum likelihood estimation of models with beta-distributed dependent variables," *Polit. Anal.*, vol. 9, no. 4, pp. 325–346, Jan. 2001, doi: 10.1093/oxfordjournals.pan.a004873.

[102] J. Buckley, "Estimation of models with beta-distributed dependent variables:

a replication and extension of paolino's study," *Polit. Anal.*, vol. 11, no. 2, pp. 204–205, Jan. 2003, doi: 10.1093/pan/mpg010.

[103] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *J. Appl. Stat.*, vol. 31, no. 7, pp. 799–815, Aug. 2004, doi: 10.1080/0266476042000214501.

[104] A. C. Harvey, "Estimating regression models with multiplicative heteroscedasticity," *Econometrica*, vol. 44, no. 3, p. 461, May 1976, doi: 10.2307/1913974.

[105] R. D. Cook and S. Weisberg, "Diagnostics for heteroscedasticity in regression," *Biometrika*, vol. 70, no. 1, p. 1, Apr. 1983, doi: 10.2307/2335938.

[106] G. K. Smyth, "Generalized linear models with varying dispersion," *J. R. Stat. Soc. Ser. B*, vol. 51, no. 1, pp. 47–60, 1989, doi: 10.1111/j.2517-6161.1989.tb01747.x.

[107] F. J. A. Cysneiros, G. A. Paula, and M. Galea, "Heteroscedastic symmetrical linear models," *Stat. Probab. Lett.*, vol. 77, no. 11, pp. 1084–1090, Jun. 2007, doi: 10.1016/j.spl.2007.01.012.

[108] M. Smithson and J. Verkuilen, "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables," *Psychol. Methods*, vol. 11, no. 1, pp. 54–71, Mar. 2006, doi: 10.1037/1082-989X.11.1.54.

[109] A. B. Simas, W. Barreto-Souza, and A. V. Rocha, "Improved estimators for a general class of beta regression models," *Comput. Stat. Data Anal.*, vol. 54, no. 2, pp. 348–366, Feb. 2010, doi: 10.1016/j.csda.2009.08.017.

[110] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," in *Trees - Structure and Function*, vol. 29, no. 6, Springer Science and

Business Media Deutschland GmbH, 1992, pp. 610–624. doi: 10.1007/978-1-4612-0919-5_38.

[111] R. Chattamvelli and R. Shanmugam, "Beta Distribution," in *Continuous Distributions in Engineering and the Applied Sciences – Part I*, Springer, Cham, 2021, pp. 41–56. doi: 10.1007/978-3-031-02430-6_4.

[112] B. Jørgensen, *The theory of dispersion models*, CRC Press, 1997.

[113] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, 1st Edition. New York, Chapman and Hall/CRC Press, 1979. doi: 10.1201/b14832.

[114] C. Birchenhall, W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes in C: The Art of Scientific Computing," *Econ. J.*, vol. 104, no. 424, p. 725, 1994, doi: 10.2307/2234664.

[115] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978, doi: 10.1214/aos/1176344136.

[116] T. Lencioni, I. Carpinella, M. Rabuffetti, A. Marzegan, and M. Ferrarin, "Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks," *Sci. Data*, vol. 6, no. 1, p. 309, 2019, doi: 10.1038/s41597-019-0323-z.

[117] S. Taş, S. Güneri, B. Kaymak, and Z. Erden, "A comparison of results of 3-dimensional gait analysis and observational gait analysis in patients with knee osteoarthritis," *Acta Orthop. Traumatol. Turc.*, vol. 49, no. 2, pp. 151–159, 2015, doi: 10.3944/AOTT.2015.14.0158.

[118] D. C. Kerrigan, M. K. Todd, U. Della Croce, L. A. Lipsitz, and J. J. Collins,

"Biomechanical gait alterations independent of speed in the healthy elderly: Evidence for specific limiting impairments," *Arch. Phys. Med. Rehabil.*, vol. 79, no. 3, pp. 317–322, 1998, doi: 10.1016/S0003-9993(98)90013-2.

[119] R. K. Begg and W. A. Sparrow, "Ageing effects on knee and ankle joint angles at key events and phases of the gait cycle," *J. Med. Eng. Technol.*, vol. 30, no. 6, pp. 382–389, Jan. 2006, doi: 10.1080/03091900500445353.

[120] F. Cribari-Neto and A. Zeileis, "Beta regression in R," *Journal of Statistical Software*, vol. 34, no. 2, 2010, doi: 10.18637/jss.v034.i02.

[121] P. L. Espinheira, S. L. P. Ferrari, and F. Cribari-Neto, "Influence diagnostics in beta regression," *Computational Statistics & Data Analysis*, vol. 52, no. 9, pp. 4417–4431, May 2008, doi: 10.1016/j.csda.2008.02.028.

[122] R. D. Cook, "Cook's distance," in *International Encyclopedia of Statistical Science*, Springer, 2011, pp. 301–302, doi: 10.1007/978-3-642-04898-2_189.

[123] A. A. Bohl, D. K. Blough, P. A. Fishman, J. R. Harris, and E. A. Phelan, "Are generalized additive models for location, scale, and shape an improvement on existing models for estimating skewed and heteroskedastic cost data?," *Health Services & Outcomes Research Methodology*, vol. 13, no. 1, pp. 18–38, Mar. 2013, doi: 10.1007/s10742-012-0086-x.

[124] C. Chen and Y. Wei, "Computational issues for quantile regression," *Sankhya*, vol. 67, no. 2, pp. 399–417, 2005, Available: https://www.jstor.org/stable/25053439.

[125] S. Majumder et al., "Smart homes for elderly healthcare—recent advances

and research challenges," *Sensors*, vol. 17, no. 11, p. 2496, Oct. 2017, doi: 10.3390/s17112496.

[126] P. Mandal, K. Tank, T. Mondal, C.-H. Chen, and M. J. Deen, "Predictive walking-age health analyzer," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 363–374, Mar. 2018, doi: 10.1109/JBHI.2017.2666603.

[127] N. Agoulmine, M. J. Deen, J.-S. Lee, and M. Meyyappan, "U-Health smart home," *IEEE Nanotechnol. Mag.*, vol. 5, no. 3, pp. 6–11, Sep. 2011, doi: 10.1109/MNANO.2011.941951.

[128] B. Jin et al., "Walking-age analyzer for healthcare applications," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 3, pp. 1034–1042, May 2014, doi: 10.1109/JBHI.2013.2296873.

[129] S. Majumder, T. Mondal, and M. J. Deen, "A simple, low-cost and efficient gait analyzer for wearable healthcare applications," *IEEE Sens. J.*, vol. 19, no. 6, pp. 2320–2329, Mar. 2019, doi: 10.1109/JSEN.2018.2885207.

[130] M. Naghshvarianjahromi, S. Majumder, S. Kumar, N. Naghshvarianjahromi, and M. J. Deen, "Natural brain-inspired intelligence for screening in healthcare applications," *IEEE Access*, vol. 9, pp. 67957–67973, 2021, doi: 10.1109/ACCESS.2021.3077529.

[131] L. Filli et al., "Profiling walking dysfunction in multiple sclerosis: characterisation, classification and progression over time," *Sci. Rep.*, vol. 8, no. 1, p. 4984, Dec. 2018, doi: 10.1038/s41598-018-22676-0.

[132] A. I. Faisal, S. Majumder, T. Mondal, D. Cowan, S. Naseh, and M. J. Deen,

"Monitoring methods of human body joints: state-of-the-art and research challenges," *Sensors*, vol. 19, no. 11, p. 2629, Jun. 2019, doi: 10.3390/s19112629.

[133] B. R. Bloem et al., "Measurement instruments to assess posture, gait, and balance in Parkinson's Disease: Critique and Recommendations," *Mov. Disord.*, vol. 31, no. 9, pp. 1342–1355, Sep. 2016, doi: 10.1002/mds.26572.

[134] A. Shcherbina et al., "The effect of digital physical Activity Interventions on Daily Step Count: A Randomised Controlled Crossover Substudy of the MyHeart Counts Cardiovascular Health Study," *Lancet Digit. Health*, vol. 1, no. 7, pp. e344–e352, Nov. 2019, doi: 10.1016/S2589-7500(19)30129-3.

[135] J. Addington et al., "Clinical and functional characteristics of youth at clinical high-risk for psychosis who do not transition to psychosis," *Psychol. Med.*, vol. 49, no. 10, pp. 1670–1677, Jul. 2019, doi: 10.1017/S0033291718002258.

[136] S. Majumder and M. J. Deen, "Wearable IMU-based system for real-time monitoring of lower-limb joints," *IEEE Sens. J.*, vol. 21, no. 6, pp. 8267–8275, Mar. 2021, doi: 10.1109/JSEN.2020.3044800.

[137] W. Jiang et al., "A Wearable tele-health system towards monitoring COVID-19 and chronic diseases," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 61–84, 2022, doi: 10.1109/RBME.2021.3069815.

[138] S. Majumder and M. J. Deen, "A robust orientation filter for wearable sensing applications," *IEEE Sens. J.*, vol. 20, no. 23, pp. 14228–14236, Dec. 2020, doi: 10.1109/JSEN.2020.3009388.

[139] M. Mukherjee, A. I. Faisal, N. Balakrishnan, S. Kumar, and M. J. Deen, "An inferential model for understanding the effects of demographic and gait factors and

their interactions on the human gait index: a beta regression approach," *IEEE J. Biomed. Health Inform.*, pp. 1–14, 2025, doi: 10.1109/JBHI.2025.3570664.

[140] R. P. Barrois et al., "Étude Observationnelle du Demi-tour à L'aide de Capteurs Inertiels Chez Les Sujets Victimes d'AVC et Relation Avec le Risque de Chute," *Neurophysiol. Clin. Neurophysiol.*, vol. 46, no. 4–5, p. 244, Nov. 2016, doi: 10.1016/j.neucli.2016.09.019.

[141] R. P.-M. M. Barrois et al., "Observational study of 180° turning strategies using inertial measurement units and fall risk in poststroke hemiparetic patients," *Front. Neurol.*, vol. 8, no. MAY, p. 236184, May 2017, doi: 10.3389/fneur.2017.00194.

[142] C. Truong et al., "A data set for the study of human locomotion with inertial measurements units," *Image Process. Line*, vol. 9, pp. 381–390, Nov. 2019, doi: 10.5201/ipol.2019.265.

[143] Y. Hutabarat, D. Owaki, and M. Hayashibe, "Recent advances in quantitative gait analysis using wearable sensors: a review," *IEEE Sens. J.*, vol. 21, no. 23, pp. 26470–26487, Dec. 2021, doi: 10.1109/JSEN.2021.3119658.

[144] I. T. G. de Oliveira Gondim, C. de C. B. de Souza, M. A. B. Rodrigues, I. M. Azevedo, M. das G. W. de Sales Coriolano, and O. G. Lins, "Portable accelerometers for the evaluation of spatio-temporal gait parameters in people with Parkinson's Disease: an integrative review," *Arch. Gerontol. Geriatr.*, vol. 90, p. 104097, Sep. 2020, doi: 10.1016/j.archger.2020.104097.

[145] D. Kobsar et al., "Wearable inertial sensors for gait analysis in adults with osteoarthritis—a scoping review," *Sensors*, vol. 20, no. 24, p. 7143, Dec. 2020, doi: 10.3390/s20247143.

[146] A. Tessitore et al., "Resting-state brain connectivity in patients with Parkinson's Disease and freezing of gait," *Parkinsonism Relat. Disord.*, vol. 18, no. 6, pp. 781–787, Jul. 2012, doi: 10.1016/j.parkreldis.2012.03.018.

[147] J. Crémers, K. D'Ostilio, J. Stamatakis, V. Delvaux, and G. Garraux, "Brain activation pattern related to gait disturbances in Parkinson's Disease," *Mov. Disord.*, vol. 27, no. 12, pp. 1498–1505, Oct. 2012, doi: 10.1002/mds.25139.

[148] M. Endo et al., "Data-driven discovery of movement-linked heterogeneity in neurodegenerative diseases," *Nat. Mach. Intell.*, vol. 6, no. 9, pp. 1034–1045, Aug. 2024, doi: 10.1038/s42256-024-00882-y.

[149] S. Pardoel, J. Kofman, J. Nantel, and E. D. Lemaire, "Wearable-sensor-based detection and prediction of freezing of gait in Parkinson's Disease: a review," *Sensors*, vol. 19, no. 23, p. 5141, Nov. 2019, doi: 10.3390/s19235141.

[150] S. T. Moore, H. G. MacDougall, and W. G. Ondo, "Ambulatory monitoring of freezing of gait in Parkinson's Disease," *J. Neurosci. Methods*, vol. 167, no. 2, pp. 340–348, Jan. 2008, doi: 10.1016/j.jneumeth.2007.08.023.

[151] T. Reches et al., "Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test," *Sensors*, vol. 20, no. 16, p. 4474, Aug. 2020, doi: 10.3390/s20164474.

[152] D. Rodríguez-Martín et al., "Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer," *PLoS One*, vol. 12, no. 2, p. e0171764, Feb. 2017, doi: 10.1371/journal.pone.0171764.

[153] R. San-Segundo, H. Navarro-Hellín, R. Torres-Sánchez, J. Hodgins, and F. De la

Torre, “Increasing robustness in the detection of freezing of gait in Parkinson’s Disease,” *Electronics*, vol. 8, no. 2, p. 119, Jan. 2019, doi: 10.3390/electronics8020119.

[154] L. Sigcha et al., “Deep learning approaches for detecting freezing of gait in Parkinson’s Disease patients through on-body acceleration sensors,” *Sensors*, vol. 20, no. 7, p. 1895, Mar. 2020, doi: 10.3390/s20071895.

[155] E. E. Tripoliti et al., “Automatic detection of freezing of gait events in patients with Parkinson’s Disease,” *Comput. Methods Programs Biomed.*, vol. 110, no. 1, pp. 12–26, Apr. 2013, doi: 10.1016/j.cmpb.2012.10.016.

[156] N. Mohammadian Rad, T. Van Laarhoven, C. Furlanello, and E. Marchiori, “Novelty detection using deep normative modeling for IMU-based abnormal movement monitoring in Parkinson’s Disease and autism spectrum disorders,” *Sensors*, vol. 18, no. 10, p. 3533, Oct. 2018, doi: 10.3390/s18103533.

[157] A. Salomon et al., “A Machine learning contest enhances automated freezing of gait detection and reveals time-of-day effects,” *Nat. Commun.*, vol. 15, no. 1, p. 4853, Jun. 2024, doi: 10.1038/s41467-024-49027-0.

[158] L. Sigcha et al., “Improvement of performance in freezing of gait detection in parkinson’s disease using transformer networks and a single waist-worn triaxial accelerometer,” *Eng. Appl. Artif. Intell.*, vol. 116, p. 105482, 2022, doi: 10.1016/j.engappai.2022.105482.

[159] M. H. M. Noor, A. Nazir, M. N. A. Wahab, and J. O. Y. Ling, “Detection of freezing of gait using unsupervised convolutional denoising autoencoder,” *IEEE Access*, vol. 9, pp. 115700–115709, 2021, doi: 10.1109/ACCESS.2021.3104975.

[160] N. Ben Chaabane et al., "Quantitative gait analysis and prediction using artificial intelligence for patients with gait disorders," *Sci. Rep.*, vol. 13, no. 1, p. 23099, Dec. 2023, doi: 10.1038/s41598-023-49883-8.

[161] X. Marimon et al., "Kinematic analysis of human gait in healthy young adults using IMU sensors: exploring relevant machine learning features for clinical applications," *Bioengineering*, vol. 11, no. 2, p. 105, Jan. 2024, doi: 10.3390/bioengineering11020105.

[162] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *Ann. Appl. Stat.*, vol. 4, no. 1, Mar. 2010, doi: 10.1214/09-AOAS285.

[163] W. Zhang et al., "Wearable sensor-based quantitative gait analysis in Parkinson's Disease patients with different motor subtypes," *npj Digit. Med.*, vol. 7, no. 1, p. 169, Jun. 2024, doi: 10.1038/s41746-024-01163-z.

[164] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014, doi: 10.1007/s10115-013-0679-x.

[165] K. Goel, R. Sindhgatta, S. Kalra, R. Goel, and P. Mutreja, "The effect of machine learning explanations on user trust for automated diagnosis of COVID-19," *Comput. Biol. Med.*, vol. 146, p. 105587, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105587.

[166] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," Feb. 2018, doi: `https://doi.org/10.48550/arXiv.1802.03888`.

[167] Y. V. Tan and J. Roy, “Bayesian additive regression trees and the general BART model,” *Stat. Med.*, vol. 38, no. 25, pp. 5048–5069, Nov. 2019, doi: 10.1002/sim.8347.

[168] C. J. Carlson, S. N. Bevins, and B. V. Schmid, “Plague risk in the western united states over seven decades of environmental change,” *Glob. Chang. Biol.*, vol. 28, no. 3, pp. 753–769, Feb. 2022, doi: 10.1111/gcb.15966.

[169] F. Huber and L. Rossini, “Inference in bayesian additive vector autoregressive tree models,” *Ann. Appl. Stat.*, vol. 16, no. 1, Mar. 2022, doi: 10.1214/21-AOAS1488.

[170] D. H. Nguyen, X. H. Le, D. T. Anh, S.-H. Kim, and D.-H. Bae, “Hourly streamflow forecasting using a bayesian additive regression tree model hybridized with a genetic algorithm,” *J. Hydrol.*, vol. 606, p. 127445, Mar. 2022, doi: 10.1016/j.jhydrol.2022.127445.

[171] J. Bleich, A. Kapelner, E. I. George, and S. T. Jensen, “Variable selection for BART: an application to gene regulation,” *Ann. Appl. Stat.*, vol. 8, no. 3, Sep. 2014, doi: 10.1214/14-AOAS755.

[172] A. R. Linero, “Bayesian regression trees for high-dimensional prediction and variable selection,” *J. Am. Stat. Assoc.*, vol. 113, no. 522, pp. 626–636, Apr. 2018, doi: 10.1080/01621459.2016.1264957.

[173] E. Saha, “Theory of posterior concentration for generalized bayesian additive regression trees,” Apr. 2023, doi: `https://doi.org/10.48550`.

[174] D. W. Apley and J. Zhu, “Visualizing the effects of predictor variables in black

box supervised learning models," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 82, no. 4, pp. 1059–1086, Sep. 2020, doi: 10.1111/rssb.12377.

[175] S. Y. Shin, Y. Kim, A. Jayaraman, and H.-S. Park, "Relationship between gait quality measures and modular neuromuscular control parameters in chronic post-stroke individuals," *J. Neuroeng. Rehabil.*, vol. 18, no. 1, p. 58, Dec. 2021, doi: 10.1186/s12984-021-00860-0.

[176] A. Rezaei, S. G. Bhat, C.-H. Cheng, R. J. Pignolo, L. Lu, and K. R. Kaufman, "Age-related changes in gait, balance, and strength parameters: a cross-sectional study," *PLoS One*, vol. 19, no. 10, p. e0310764, Oct. 2024, doi: 10.1371/journal.pone.0310764.

[177] A. P. J. Zanardi et al., "Gait parameters of Parkinson's Disease compared with healthy controls: a systematic review and meta-analysis," *Sci. Rep.*, vol. 11, no. 1, p. 752, Jan. 2021, doi: 10.1038/s41598-020-80768-2.

[178] D. S. Peterson, M. Mancini, P. C. Fino, F. Horak, and K. Smulders, "Speeding up gait in parkinson's disease," *J. Parkinsons. Dis.*, vol. 10, no. 1, pp. 245–253, Jan. 2020, doi: 10.3233/JPD-191682.

[179] S. J. Olney and C. Richards, "Hemiparetic gait following stroke. part I: characteristics," *Gait Posture*, vol. 4, no. 2, pp. 136–148, Apr. 1996, doi: 10.1016/0966-6362(96)01063-6.

[180] L. R. Nascimento, C. Q. de Oliveira, L. Ada, S. M. Michaelsen, and L. F. Teixeira-Salmela, "Walking training with cueing of cadence improves walking speed and stride length after stroke more than walking training alone: a systematic review," *J. Physiother.*, vol. 61, no. 1, pp. 10–15, Jan. 2015, doi: 10.1016/j.jphys.2014.11.015.

[181] R. L. McGrath, M. L. Ziegler, M. Pires-Fernandes, B. A. Knarr, J. S. Higginson, and F. Sergi, "The effect of stride length on lower extremity joint kinetics at various gait speeds," *PLoS One*, vol. 14, no. 2, p. e0200862, Feb. 2019, doi: 10.1371/journal.pone.0200862.

[182] S. Shin, R. J. Valentine, E. M. Evans, and J. J. Sosnoff, "Lower extremity muscle quality and gait variability in older adults," *Age Ageing*, vol. 41, no. 5, pp. 595–599, Sep. 2012, doi: 10.1093/ageing/afs032.

[183] N. P. Wages et al., "Relative contribution of muscle strength, lean mass, and lower extremity motor function in explaining between-person variance in mobility in older adults," *BMC Geriatr.*, vol. 20, no. 1, p.

[184] J. Schlicht, D. N. Camaione, and S. V. Owen, "Effect of intense strength training on standing balance, walking speed, and sit-to-stand performance in older adults," *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.*, vol. 56, no. 5, pp. M281–M286, May 2001, doi: 10.1093/gerona/56.5.M281.

[185] B. M. Peoples, K. D. Harrison, K. G. Santamaria-Guzman, S. E. Campos-Vargas, P. G. Monaghan, and J. A. Roper, "Functional lower extremity strength influences stepping strategy in community-dwelling older adults during single and dual-task walking," *Sci. Rep.*, vol. 14, no. 1, p. 13379, Jun. 2024, doi: 10.1038/s41598-024-64293-0.

[186] E. M. Murtagh, J. L. Mair, E. Aguiar, C. Tudor-Locke, and M. H. Murphy, "Outdoor walking speeds of apparently healthy adults: a systematic review and meta-analysis," *Sport. Med.*, vol. 51, no. 1, pp. 125–141, Jan. 2021, doi: 10.1007/s40279-020-01351-3.

[187] S. A. Alves, R. M. Ehrig, P. C. Raffalt, A. Bender, G. N. Duda, and A. N. Agres, “Quantifying asymmetry in gait: the weighted universal symmetry index to evaluate 3D ground reaction forces,” *Front. Bioeng. Biotechnol.*, vol. 8, p. 579511, Oct. 2020, doi: 10.3389/fbioe.2020.579511.

[188] A. Gouelle and F. Mégrot, “Interpreting spatiotemporal parameters, symmetry, and variability in clinical gait analysis,” in *Handbook of Human Motion*, Cham: Springer International Publishing, 2016, pp. 1–20, doi: 10.1007/978-3-319-30808-1-35-1.

[189] P. S. Sung, J. T. Zipple, J. M. Andraka, and P. Danial, “The kinetic and kinematic stability measures in healthy adult subjects with and without flat foot,” *Foot*, vol. 30, pp. 21–26, Mar. 2017, doi: 10.1016/j.foot.2017.01.010.

[190] H. B. Menz, M. E. Morris, and S. R. Lord, “Foot and ankle risk factors for falls in older people: a prospective study,” *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.*, vol. 61, no. 8, pp. 866–870, Aug. 2006, doi: 10.1093/gerona/61.8.866.

[191] S. Cabral, “Gait symmetry measures and their relevance to gait retraining,” in *Handbook of Human Motion*, vol. 1–3, Cham: Springer International Publishing, 2018, pp. 429–447, doi: 10.1007/978-3-319-14418-4-201.

[192] S. Lv, Z. Huan, X. Chang, Y. Huan, and J. Liang, “Analysis of gait symmetry under unilateral load state,” in *Communications in Computer and Information Science*, vol. 1321, Springer, Singapore, 2020, pp. 244–256, doi: 10.1007/978-981-33-4214-9-18.

[193] S. Lundberg and S.-I. Lee, “An unified approach to interpreting model predictions,” May 2017, doi: https://doi.org/10.48550/arXiv.

[194] V. Ročková and E. Saha, "On theory for BART," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, Apr. 2019, pp. 2839–2848, PMLR.

[195] A. Kapelner and J. Bleich, "bartMachine: Machine learning with bayesian additive regression trees," *J. Stat. Softw.*, vol. 70, pp. 1–40, 2016.

[196] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[197] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint* arXiv:1802.03888, 2018.

[198] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, 1st ed. Chapman and Hall/CRC Press, 1984. doi: 10.1201/9781315139470.

[199] S. Chebrolu, A. Abraham, and J. P. Thomas, "Feature deduction and ensemble design of intrusion detection systems," *Computers & Security*, vol. 24, no. 4, pp. 295–307, 2005.

[200] M. Sandri and P. Zuccolotto, "A bias correction algorithm for the gini variable importance measure in classification trees," *J. Comput. Graph. Stat.*, vol. 17, no. 3, pp. 611–628, 2008.

[201] L. Auret and C. Aldrich, "Empirical comparison of tree ensemble variable importance measures," *Chemom. Intell. Lab. Syst.*, vol. 105, no. 2, pp. 157–170, 2011.

[202] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, pp. 1–13, 2006.

[203] H. Ishwaran, “Variable importance in binary regression trees and forests,” *Electron. J. Statist.*, vol. 1, pp. 519–537, 2007.

[204] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[205] V. N. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer, 1995.

[206] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann, 1988.

[207] E. Fix, *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*, vol. 1. USAF School of Aviation Medicine, 1985.

[208] D. E. Rumelhart and PDP Research Group, *Parallel Distributed Processing*, vol. 1. Cambridge, MA: MIT Press, 1986.

[209] D. J. C. MacKay, “Bayesian interpolation,” *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.

[210] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design.* Boston, MA: PWS Publishing, 1996.

[211] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[212] G. Fumera, F. Roli, and A. Serrau, “A theoretical analysis of bagging as a linear combination of classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1293–1299, 2008.

[213] X. Cai, S. Yang, F. Zheng, M. Lu, Y. Wu, and S. Krishnan, “Knee joint vibration signal analysis with matching pursuit decomposition and dynamic weighted classifier fusion,” *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 904267, 11 pages, 2013.

[214] M. Schumacher, R. Roßner, and W. Vach, “Neural networks and logistic regression: Part I,” *Computational Statistics & Data Analysis*, vol. 21, no. 6, pp. 661–682, 1996.

[215] C. M. Bishop, *Neural networks for pattern recognition.* Oxford, U.K.: Clarendon Press, 1997.

[216] T. Schmeltzpfenning and T. Brauner, “Foot biomechanics and gait,” in *Handbook of Footwear Design and Manufacture*, Elsevier, pp. 27–48, 2013.

[217] C.H. Na et al., “Kinematic movement and balance parameter analysis in neurological gait disorders,” *J. Biol. Eng.*, vol. 18, no. 1, p. 6, Jan. 2024, doi: 10.1186/s13036-023-00398-w.

[218] O. Beauchet et al., “Guidelines for assessment of gait and reference values for spatiotemporal gait parameters in older adults: the biomathics and canadian gait consortiums initiative,” *Front. Hum. Neurosci.*, vol. 11, p. 353, Aug. 2017, doi: 10.3389/fnhum.2017.00353.

[219] R. Nakamura, T. Handa, S. Watanabe, and I. Morohashi, “Walking cycle after stroke,” *Tohoku J. Exp. Med.*, vol. 154, no. 3, pp. 241–244, 1988, doi: 10.1620/tjem.154.241.

[220] L. Majed, R. Ibrahim, M. J. Lock, and G. Jabbour, “Walking around the preferred speed: examination of metabolic, perceptual, spatiotemporal and

stability parameters,” *Front. Physiol.*, vol. 15, p. 1357172, Feb. 2024, doi: 10.3389/fphys.2024.1357172.