UNIVERSITY OF VERONA

DEPARTMENT OF

Computer Science

DOCTORAL PROGRAM IN

Computer Science

XXXVIII cycle (2022-2025)

Wearable-Based Artificial Intelligence for Freezing of Gait Detection in Parkinson's Disease

S.S.D. IINF - 05/A

Coordinator: **Prof. Rosalba Giugno**

Co-Supervisor: **Prof. Graziano Pravadelli**
Prof. Florenc Demrozi

Candidate: **Michele Tebaldi**

To my mother,
for her endless, unwavering support and love.
Thank you for everything.
Always.

Abstract

Freezing of Gait (FoG) is one of the most disabling motor symptoms of Parkinson's disease (PD), characterized by brief episodes in which patients are unable to initiate or continue walking. Its unpredictable nature and strong variability across individuals make automated detection a challenging yet crucial task for clinical monitoring and personalized treatment. This thesis aims to develop robust and generalizable AI-based models for the detection of FoG using wearable inertial sensors. The research follows a unified trajectory: first, it analyzes the current landscape of IMU-based FoG datasets to identify structural limitations and establish methodological guidelines; second, it investigates activity recognition of key motor tasks—such as turning and sit-to-stand transitions—that often precede FoG episodes, providing a framework for contextualized detection. Building on these foundations, the work explores transfer learning to overcome data scarcity and reduce costs in terms of time and work, and introduces a Mixture-of-Experts strategy to enhance adaptability across heterogeneous sensor configurations and patient populations. Overall, this study contributes to the advancement of scalable and clinically meaningful approaches for Freezing of Gait detection in Parkinson's disease, paving the way toward personalized, wearable-based monitoring systems.

Keywords: Parkinson's Disease, Freezing of Gait (FoG), Wearable Sensors, Inertial Measurement Units (IMU), Human Activity Recognition (HAR), Deep Learning, Machine Learning, Transfer Learning, Fine-Tuning, Mixture-of-Experts (MoE), Cross-Dataset Generalization, Sensor Heterogeneity, Personalization.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Rosalba Giugno, for her continuous support and guidance throughout my research journey. From the very beginning, she welcomed me with trust and encouragement, allowing me the freedom to explore my own path while always remaining a steady point of reference. Her mentorship, respectful of my independence, has been fundamental to my growth as a researcher. Even when my work extended beyond her primary area of expertise, she supported my choices with openness and confidence.

I would like to express my sincere gratitude to my co-supervisor, Prof. Graziano Pravadelli, for giving me the opportunity to embark on this doctoral journey and for his constant trust in my potential. From the very beginning, he believed in me and supported me, even during moments of uncertainty and self-doubt. His confidence and encouragement have been a steady source of motivation throughout my PhD.

I am also deeply grateful to my co-supervisor, Prof. Florenc Demrozi, for his continuous support and valuable guidance. His technical expertise and rigorous approach have significantly contributed to the development of this work. I particularly appreciate his encouragement in pushing me to actively engage with the research community, including participating in international conferences. As a student with a disability, his support in creating these opportunities — together with hosting me at the University of Stavanger during my external research period — has been especially meaningful. His commitment to inclusion, combined with his dedication to high research standards, has had a lasting impact on both this work and my personal growth as a researcher.

I would also like to sincerely thank Prof. Luigi Borzì for his valuable collaboration, which enriched this work both scientifically and on a personal level. His support and perspective have been truly meaningful throughout this journey.

I gratefully acknowledge the financial support provided by the Piano Nazionale di Ripresa e Resilienza (PNRR) and the partnership with Fondazione Giuseppe Manni, which made this research possible.

I would like to express my deepest gratitude to my mother for her unconditional love, strength, and support. I am aware of all that she has done so that I could grow, find my path, and become the person I am today. This achievement bears also her mark.

Finally, I would like to thank my family, friends, and colleagues for their constant support, encouragement, and understanding throughout this journey.

Contents

Chapter 1

Introduction and Background

1.1 Parkinson's Disease and Freezing of Gait

Parkinson's disease (PD) is a progressive neurodegenerative disorder and one of the leading causes of disability among older adults worldwide. Its incidence and prevalence have increased substantially over recent decades, reflecting global population ageing and improvements in diagnostic accuracy [1, 2]. Epidemiological estimates indicate that PD affects approximately one to two percent of individuals over the age of sixty, with higher prevalence in men than in women [3, 4]. Although most cases are idiopathic, PD arises from a complex interplay between genetic predisposition and environmental exposures such as pesticides, heavy metals, and air pollution [5–7]. Familial forms account for less than ten percent of all cases and are associated with mutations in genes regulating protein aggregation, mitochondrial quality control, and lysosomal function, including *SNCA*, *LRRK2*, *PRKN*, *PINK1*, and *GBA* [8, 9]. Despite these advances, the majority of PD cases remain sporadic, underlining the multifactorial and heterogeneous nature of the disease.

From a neuropathological perspective, PD is characterized by the progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta, leading to a marked depletion of striatal dopamine [10]. The pathological hallmark is the presence of intraneuronal inclusions known as Lewy bodies and Lewy neurites, composed primarily of misfolded and aggregated alpha-synuclein [11, 12]. The resultant dopamine deficiency disrupts the basal ganglia–thalamo–cortical circuits that regulate voluntary movement, causing an imbalance between excitatory and inhibitory signaling pathways [13, 14]. This circuit-level dysfunction underlies the classical motor symptoms of PD: bradykinesia, rigidity, resting tremor, and postural

instability [15].

Beyond dopaminergic loss, PD involves widespread neurodegeneration across multiple neurotransmitter systems, including noradrenergic, serotonergic, and cholinergic pathways [16, 17]. These alterations contribute to the emergence of non-motor symptoms such as cognitive impairment, sleep disturbances, depression, and autonomic dysfunction. The mechanisms driving neuronal death are multifactorial and include mitochondrial dysfunction, oxidative stress, impaired protein clearance, and chronic neuroinflammation [18, 19]. Misfolded alpha-synuclein aggregates disrupt mitochondrial respiration and axonal transport, exacerbating oxidative injury and synaptic failure [20]. According to the Braak staging hypothesis, pathological alpha-synuclein deposition follows a characteristic ascending pattern—beginning in peripheral structures such as the olfactory bulb and enteric nervous system, then spreading to the brainstem, midbrain, and finally cortical areas [12, 21]. This spatial progression mirrors the clinical evolution of PD, in which early non-motor symptoms often precede overt motor impairment by years.

Clinically, PD is a highly heterogeneous disorder. In its prodromal phase, subtle non-motor symptoms such as hyposmia, constipation, REM sleep behavior disorder, and depression can manifest long before motor onset [22, 23]. Motor signs typically emerge after approximately 50–60% of dopaminergic neurons have degenerated in the substantia nigra [24]. Bradykinesia—defined as slowness and decrement in movement amplitude—is the cardinal motor feature, often accompanied by rigidity and resting tremor. As the disease advances, axial symptoms such as postural instability and gait disturbances become increasingly disabling. Pharmacological treatment with levodopa and dopaminergic agonists provides symptomatic relief but does not modify disease progression [25]. Over time, long-term therapy leads to motor fluctuations and dyskinesias due to pulsatile dopamine receptor stimulation and maladaptive synaptic plasticity [26, 27]. For patients with advanced disease and refractory motor complications, deep brain stimulation (DBS) of the subthalamic nucleus or globus pallidus internus can effectively improve motor control and quality of life [10, 28].

Among the most disabling motor complications of PD is *freezing of gait* (FoG), a transient and episodic phenomenon in which patients are unable to initiate or maintain forward stepping despite the intention to walk. During a FoG episode, patients frequently describe the sensation as if their feet were "glued to the floor," often accompanied by trembling of the legs or rapid shuffling movements [29, 30]. FoG typically occurs during turning, gait initiation, or when navigating narrow spaces, and can be precipitated by stress, time pressure, or cognitive dual-tasking [31, 32]. Its

expression varies considerably between patients—some experience only occasional episodes under specific circumstances, while others develop frequent and severe freezing that leads to falls and loss of independence [33]. FoG is closely linked to disease duration and dopaminergic treatment; it often emerges in advanced stages but can also appear early in the disease course [34]. Neuroimaging and electrophysiological studies suggest that FoG reflects disruptions within cortical–basal ganglia–brainstem networks responsible for motor planning and executive control, particularly involving the supplementary motor area and pedunculopontine nucleus [35–37].

Clinically, FoG poses a major challenge because of its unpredictable and context-dependent nature. It contributes substantially to falls, reduced mobility, and decreased quality of life [32, 33]. Standard clinical assessments such as the Freezing of Gait Questionnaire (FOG-Q) and task-based evaluations during hospital visits provide valuable information but have significant limitations. FoG episodes are often absent during examinations, and self-reports may be biased by poor recall or underestimation [38]. Laboratory-based gait analyses and instrumented walkways can quantify spatiotemporal parameters but are restricted to short, artificial observation periods. Consequently, the real-world frequency, severity, and temporal patterns of FoG remain poorly captured by conventional tools. These limitations underscore the need for continuous, objective monitoring strategies that can characterize FoG dynamics in naturalistic conditions, bridging the gap between clinical evaluation and real-life patient experience.

1.2 From Clinical Assessment to Wearable Sensing

Clinical assessment of Parkinson's disease (PD) has traditionally relied on structured rating scales, standardized motor tests, and patient-reported diaries. The Unified Parkinson's Disease Rating Scale (UPDRS) and its revised version, the Movement Disorder Society–UPDRS (MDS–UPDRS), remain the gold standards for quantifying disease severity across motor and non-motor domains [39]. Although widely adopted, these scales provide only intermittent "snapshots" of patient status during scheduled clinic visits and are influenced by observer subjectivity and situational variability. Patient diaries, such as the Hauser diary used to document medication-related motor fluctuations, offer complementary information but depend heavily on compliance and recall accuracy, which can be unreliable in advanced stages [40].

The dynamic nature of PD symptoms poses major challenges for such episodic assessments. Tremor, bradykinesia, and gait disturbances fluctuate over medication

cycles, fatigue levels, and environmental contexts. Freezing of gait (FoG), in particular, is unpredictable and context-dependent—often triggered by turning, gait initiation, or cognitive dual-tasking—and thus rarely captured during clinical observation [34, 38]. These limitations highlight the need for objective, continuous measurements capable of tracking motor behavior in real-world conditions [41, 42].

Over the past decade, wearable technologies have emerged as a promising solution, enabling quantitative and longitudinal monitoring of PD motor symptoms with minimal intrusion. By continuously recording physiological and biomechanical signals, wearables generate high-resolution data that can complement clinical evaluation, facilitate early diagnosis, and guide individualized therapy. Among the available technologies—ranging from inertial sensors to electrophysiological and autonomic monitors—inertial measurement units (IMUs) have become the cornerstone of PD monitoring due to their compactness, low power consumption, and ability to capture motion dynamics accurately [42, 43].

IMUs integrate tri-axial accelerometers and gyroscopes, and sometimes magnetometers, to measure linear acceleration, angular velocity, and orientation. When attached to specific body locations such as the ankles, wrists, trunk, or lumbar region, IMUs record precise kinematic information on gait cycles, postural transitions, and tremor frequency. These signals can be processed to derive temporal and spatial gait parameters—stride length, cadence, and variability—as well as fine-grained measures of limb coordination and balance. Multi-sensor configurations enable whole-body motion analysis comparable to laboratory-grade motion-capture systems while remaining suitable for daily-life use [42, 44]. The ability to monitor continuously and unobtrusively in natural environments has made IMUs the reference modality for digital biomarkers of PD progression.

Beyond IMUs, several complementary sensing technologies provide a more comprehensive view of PD pathophysiology. *Electrocardiography (ECG)* sensors record cardiac electrical activity, allowing assessment of autonomic nervous-system function and heart-rate variability, which are often altered in PD [45]. *Electromyography (EMG)* measures muscle activation and helps quantify rigidity, tremor amplitude, and coordination, especially when combined with inertial data [46]. *Electroencephalography (EEG)* captures cortical oscillations and has revealed abnormal beta-band synchronization and impaired cortical–basal-ganglia coupling related to bradykinesia and FoG [47]. *Pressure-sensing insoles* detect plantar pressure distribution and ground-reaction forces, providing insight into gait asymmetry, stride timing, and balance impairment [48]. Finally, *electrodermal-activity (EDA)* sensors

measure skin conductance changes that reflect sympathetic activation, offering potential biomarkers of stress and non-motor symptoms [49, 50]. Together, these sensing modalities enable multimodal characterization of PD's complex motor and autonomic manifestations.

Despite their promise, wearable systems face challenges before full clinical translation. Data variability caused by differences in sensor placement, calibration, and user compliance can compromise measurement reliability [42, 43]. Real-world recordings introduce noise and artifacts from uncontrolled environments, demanding robust signal-processing and quality-control pipelines [41, 45]. Moreover, the large data volumes generated by long-term monitoring necessitate automated analytical methods to extract clinically meaningful features. These needs have accelerated the integration of machine learning (ML) and artificial intelligence (AI) into PD research, enabling automatic recognition of motor patterns and detection of pathological events [41, 51]. Such approaches represent the next step toward objective, scalable, and patient-centered management of the disease [52, 53].

The development of wearable sensing systems therefore marks a paradigm shift in PD care—from subjective, episodic evaluations to continuous, quantitative observation. This transition is particularly critical for complex phenomena such as freezing of gait, whose manifestations depend on dynamic motor context and are rarely observed in clinical settings. The following section introduces AI-driven analytical frameworks that unlock the full potential of wearable data for precise and clinically meaningful detection of FoG and other motor abnormalities.

1.3 Artificial Intelligence for Motor Symptom Analysis and Freezing of Gait Detection

The rapid growth of wearable sensing and digital health technologies has generated unprecedented amounts of high-resolution data describing human movement, physiology, and behavior. Extracting clinically meaningful information from these continuous signals requires computational approaches capable of modeling their complexity and variability. In Parkinson's disease (PD), where motor symptoms fluctuate dynamically and manifest differently across individuals, the integration of artificial intelligence (AI) has become central to developing quantitative, objective biomarkers. By learning patterns from multimodal sensor data, AI enables automated recognition of motor states, early detection of pathological events, and personalized

disease management strategies [41, 52].

Traditional analysis methods in PD research relied on statistical summaries or rule-based thresholds applied to features such as tremor amplitude or stride variability. Although interpretable, these approaches lacked flexibility and struggled to handle nonstationary, noisy signals from real-world settings. Machine learning (ML) offers a more general framework, allowing models to infer complex relationships between input features and clinical labels directly from data. In the early 2000s, conventional ML algorithms were introduced to classify PD motor states using signals from inertial measurement units (IMUs), electromyography (EMG), and other wearable sensors [44, 54, 55]. The standard pipeline consisted of signal preprocessing (denoising, segmentation), handcrafted feature extraction (time, frequency, and time–frequency domains), and supervised classification using algorithms such as support vector machines (SVM), random forests (RF), k-nearest neighbors (k-NN), or logistic regression [**?**, 56]. These methods achieved good accuracy for distinguishing tremor or "on/off" medication states and for quantifying bradykinesia and dyskinesia severity. However, their generalizability was often limited by dependence on task-specific features and subject variability.

A fundamental challenge in ML for PD lies in the generation of reliable ground-truth labels. Motor states and freezing events are typically annotated through video recordings synchronized with wearable sensors and reviewed by clinical experts [38, 57]. This process is time-consuming and inherently subjective, introducing inconsistencies between raters and across studies. Variability in experimental protocols, sensor placements, and acquisition hardware further complicates model comparison and reproducibility. These issues underscore the importance of standardized data-collection frameworks and robust cross-validation strategies to ensure external validity.

Evaluation metrics are critical for assessing model performance in clinical contexts. Accuracy alone can be misleading when class distributions are imbalanced—as is common in FoG detection, where non-freezing periods dominate the data. Alternative metrics such as sensitivity (recall), specificity, precision, F1-score, and area under the receiver operating characteristic curve (ROC-AUC) provide more balanced evaluation [58, 59]. In time-series tasks, overlap-based metrics such as segment-wise F1 or intersection-over-union (IoU) better capture temporal alignment between predicted and true events. Transparent reporting of these measures is essential for reproducibility and comparison across studies.

Despite the promise of classical ML, these algorithms depend heavily on manually

engineered features derived from domain expertise. To overcome these constraints, deep learning (DL) methods have been increasingly adopted for wearable-based PD monitoring. DL models consist of multiple layers of nonlinear transformations that automatically learn hierarchical representations of data, thereby eliminating the need for handcrafted features. Convolutional neural networks (CNNs) capture local temporal and frequency-specific patterns in sensor signals, while recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures model long-range temporal dependencies [51,60,61]. Hybrid CNN–LSTM models combine both feature hierarchies and temporal context, achieving state-of-the-art performance in activity recognition and PD motor-state classification [52,62]. These architectures can also fuse data from multiple sensors (e.g., IMUs, EMG, ECG) to exploit complementary information about motor and autonomic function.

Deep learning's ability to model nonlinear relationships and complex temporal structures has substantially improved automated PD monitoring. Models trained on IMU data can now classify walking, turning, or sitting transitions, detect tremor episodes, and estimate clinical motor scores with increasing accuracy [41, 51]. Moreover, once trained, these networks can operate in real time on embedded hardware, opening pathways for wearable-based decision support and closed-loop systems [61,63]. Nonetheless, the power of DL comes at a cost: its effectiveness depends on large, diverse datasets to prevent overfitting. In PD, where collecting labeled data is expensive and burdensome, dataset size and heterogeneity remain major bottlenecks. Transfer learning, data augmentation, and self-supervised learning have thus emerged as critical strategies to improve generalization across patients and recording conditions [51,62].

Parallel to methodological progress, the field has witnessed a conceptual shift toward explainability and interpretability. Clinicians require insight into how models make decisions to trust AI-generated predictions [64]. Visualization techniques such as saliency maps, attention mechanisms, and layer-wise relevance propagation allow interpretation of which temporal segments or sensor channels contribute most to model outputs. These advances not only enhance clinical credibility but also reveal physiological correlates of learned representations, bridging the gap between AI and neuroscience.

AI-based analysis has been applied to a wide range of PD motor symptoms. Tremor detection algorithms typically analyze accelerometer signals to identify oscillations within the characteristic 4–6 Hz band [43, 54]. Bradykinesia and hypokinesia are assessed through movement amplitude and velocity parameters extracted from limb

and trunk sensors [42,44]. Dyskinesia, often induced by dopaminergic therapy, is modeled using spectral power and movement intensity features [54,55]. Multisensor fusion further enables estimation of axial motor symptoms, gait instability, and postural transitions, providing objective biomarkers that correlate strongly with clinical scales such as the UPDRS [41,42]. Together, these applications demonstrate AI's capacity to extend clinical observation beyond the confines of laboratory or hospital settings.

Among PD symptoms, freezing of gait (FoG) represents the most challenging for automated detection. FoG manifests as a sudden, short-lived inability to move the feet forward, despite the intention to walk, often accompanied by trembling of the legs or rapid shuffling [29,30]. Early computational approaches exploited the "freeze index," which quantifies the ratio of spectral power in the 3–8 Hz "freezing band" to that in the 0.5–3 Hz "locomotion band" [57]. While simple and interpretable, this method was highly sensitive to walking speed and sensor positioning. Later ML-based approaches used handcrafted features from acceleration and angular-velocity signals combined with classifiers such as SVMs or decision trees to detect FoG episodes [44,55]. These techniques improved sensitivity but struggled with inter-patient generalization due to heterogeneous gait patterns.

The adoption of DL has revolutionized FoG detection. CNNs and LSTMs trained directly on raw IMU signals can automatically learn discriminative features representing both gait dynamics and contextual transitions [51,53,60,65]. Multimodal networks integrating IMU and EMG data capture both kinematic and muscular signatures of freezing, while attention-based architectures highlight salient temporal regions associated with pre-freeze events, improving interpretability [53]. Recent work has extended the focus from detection to prediction—forecasting FoG episodes several seconds before onset by identifying prodromal gait changes [36,37,47]. These predictive systems could enable timely feedback or cueing interventions to prevent freezing.

Physiologically, FoG reflects disruptions across the cortical–basal ganglia–brainstem network, involving the supplementary motor area, subthalamic nucleus, and pedunculopontine nucleus [35, 37]. Neuroimaging and EEG studies indicate that abnormal oscillatory coupling between these regions impairs the transition from motor planning to execution, producing intermittent "motor blocks" [36, 47]. Incorporating such neuroscientific insights into AI architectures—through multimodal integration or physiologically constrained modeling—represents an emerging frontier in the field.

Despite substantial progress, several limitations hinder clinical deployment of AI-based FoG detection. Existing datasets remain small and heterogeneous, collected under varying conditions, sensor placements, and annotation standards [**?**]. The lack of large-scale, standardized, open datasets restricts generalization and benchmarking. Inter-patient variability further complicates modeling: freezing episodes differ in duration, intensity, and triggers among individuals. In addition, ethical and regulatory considerations regarding data privacy and algorithmic transparency must be addressed before widespread clinical adoption [66].

To overcome these barriers, research is increasingly exploring data-efficient and adaptive AI frameworks.

Transfer learning has therefore emerged as a key strategy to mitigate data scarcity and variability in wearable-based PD monitoring. In its simplest form, transfer learning allows a model trained on a source domain—typically a large and heterogeneous dataset—to be reused for a target domain with limited labeled data [51, 62]. The underlying assumption is that early layers of deep networks learn generic representations of human motion, such as rhythmicity, smoothness, and coordination, that are transferable across related motor tasks or sensor configurations. By reusing these learned features, researchers can significantly reduce both the computational cost and the amount of annotated data required for model development.

Two principal transfer paradigms are commonly employed: *feature extraction* and *fine-tuning*. In feature extraction, a pretrained model acts as a fixed feature encoder; the last layers are removed, and the extracted embeddings are used to train a lightweight classifier (e.g., logistic regression, support vector machine) on the target data. This approach is particularly advantageous when the source and target domains are moderately similar but differ in scale or annotation density, such as adapting a model trained on general human-activity datasets to Parkinsonian gait data [51]. Fine-tuning, on the other hand, involves re-training some or all of the pretrained layers on the target dataset with a reduced learning rate. This gradual adaptation allows the network to retain generic motion representations while adjusting to disease-specific characteristics such as tremor amplitude, stride variability, or the transient oscillations preceding freezing episodes [62].

Cross-domain fine-tuning has proven particularly useful in PD research. Models initially trained on large open datasets like WISDM, PAMAP2, or Opportunity—containing diverse human activities from healthy subjects—can serve as foundations for detecting PD-specific motor states [60, 61]. Subsequent fine-tuning on smaller clinical datasets enhances sensitivity to pathological movement patterns

without requiring extensive labeled data collection. In the context of FoG detection, transfer learning can also be applied across sensor placements or device types: a model trained with ankle-mounted IMUs may be fine-tuned to operate with waist or wrist sensors by re-adjusting feature representations at intermediate network layers.

A related concept is *domain adaptation*, which explicitly aligns the feature distributions of source and target domains. Techniques such as adversarial domain adaptation or maximum mean discrepancy regularization aim to minimize representation shift between datasets collected under different conditions or from different patient cohorts. These methods are particularly valuable when sensor hardware, sampling rates, or patient demographics vary, ensuring that the learned embeddings remain invariant to such nuisance factors [62].

In parallel, *few-shot* and *meta-learning* approaches are gaining attention for personalized PD monitoring. Instead of retraining a global model for each new patient, meta-learning algorithms learn initialization parameters that can be rapidly fine-tuned using only a few examples of individual movement data. This paradigm is promising for clinical deployment, where data availability per subject is inherently limited and inter-individual variability is high.

Overall, transfer learning and fine-tuning represent crucial steps toward practical, data-efficient AI for Parkinson's disease. They enable leveraging of the growing ecosystem of open human-motion datasets, facilitate adaptation to new patients' characteristics, and drastically reduce the labeling burden on clinical experts. Along with adaptive ensemble methods such as mixture-of-experts architectures, these strategies pave the way toward generalizable and clinically trustworthy AI models for real-world FoG detection.

1.4 Mixture-of-Experts: evolution and wearable sensing perspectives

The Mixture-of-Experts (MoE) paradigm originates from statistical ensemble modeling, where complex input–output mappings are decomposed into specialized submodels whose contributions are combined through an input-dependent gating function. The foundational formulation by Jacobs et al. introduced the adaptive mixture-of-local-experts architecture, in which multiple expert networks are jointly trained with a gating network that partitions the input space and assigns mixture weights dynamically [67]. This early probabilistic view framed MoE as a conditional

mixture model, encouraging specialization while maintaining end-to-end differentiability.

Although conceptually powerful, early MoE models were limited by computational constraints and scalability issues. The resurgence of MoE in deep learning was driven by the need to increase model capacity without proportionally increasing computational cost. Shazeer et al. introduced sparsely-gated MoE layers, enabling conditional computation by activating only a small subset of experts for each input [68]. This formulation significantly increased parameter capacity while preserving tractable inference complexity. Subsequent developments such as GShard [69] and Switch Transformers [70] demonstrated the feasibility of scaling MoE architectures to hundreds of billions or trillions of parameters through sparse routing and expert parallelism.

Beyond large language models, recent work has repositioned MoE as a general framework for addressing the challenges of big data environments. Gan et al. provide a comprehensive analysis of MoE from a big-data perspective, highlighting its advantages in high-dimensional sparse data modeling, heterogeneous multisource data fusion, real-time online learning, and interpretability [71]. In this view, MoE executes a divide-and-conquer strategy: expert subnetworks specialize on distinct subspaces or regimes, while the gating mechanism performs dynamic coordination and fusion.

From a structural standpoint, MoE consists of three core components: (1) a set of expert subnetworks $f_i(x;\theta_i)$; (2) a gating network $G(x)$ that computes expert relevance scores; and (3) an aggregation function combining expert outputs. In sparse implementations, the gating mechanism selects only the top-k experts:

$$G(x) = \mathrm{Softmax}(\mathrm{KeepTopK}(H(x), k)),$$

where $H(x)$ denotes the raw gating logits and k controls sparsity [68, 71]. The final prediction is computed as:

$$F(x) = \sum_{i \in \mathrm{TopK}} G_i(x)\, f_i(x).$$

This conditional activation enables scalability, improves resource utilization, and promotes specialization across heterogeneous data regimes. A recurring practical challenge is expert under-/over-utilization (routing imbalance), which is typically mitigated through routing regularization and load-balancing objectives [68, 70].

Although MoE has been most visible in large-scale language and vision models [71] , its architectural motivation aligns closely with wearable inertial sensing,

where data distributions shift across subjects, sensor placements, activity contexts, and acquisition conditions. In human activity recognition (HAR)—a domain methodologically adjacent to clinical gait monitoring—recent studies have begun to use MoE explicitly to improve representation transfer and robustness for IMU-based classification. Ray et al. propose an IMU-HAR pipeline that leverages multimodal pretraining (video, pose, and synthetic IMU) and then introduces a downstream MoE classifier to improve recognition performance when only limited real IMU data are available [72]. Complementing this, Roy and Girdzijauskas explore "mixing temporal experts" for HAR, motivating MoE as a mechanism to capture activity-specific temporal dynamics that are difficult to model with a single monolithic temporal encoder [73]. Beyond static expert pools, continual learning settings for wearable sensing have also adopted MoE to address distribution shift over time: Rahman et al. introduce a dynamic MoE approach for sensor-based HAR that grows experts incrementally for new tasks and supports flexible expert composition, aiming to reduce parameter growth and training cost compared to deploying one large fixed architecture everywhere [74]. Finally, broader time-series modeling trends indicate increasing interest in sparse MoE for scaling temporal foundation models; for example, Time-MoE proposes a large-scale time-series forecasting architecture using sparse expert activation to increase capacity while controlling inference cost [75]. Together, these works support MoE as a natural architectural candidate for wearable IMU pipelines, particularly when the primary bottlenecks are heterogeneity and generalization rather than in-lab accuracy alone — a setting that closely matches real-world Parkinsonian gait monitoring and freezing-of-gait detection.

1.5 Contextual Overview and Thesis Structure

Freezing of gait (FoG) represents one of the most complex and disabling motor symptoms of Parkinson's disease. Its episodic, context-dependent nature makes it difficult to assess reliably through traditional clinical examinations, which capture only short and highly controlled observations of patient behavior. As a result, much of the real-world variability and temporal dynamics of FoG remain poorly characterized.

Wearable sensing technologies have emerged as a powerful tool to address this limitation. Inertial measurement units (IMUs) and other portable sensors enable continuous and unobtrusive monitoring of movement in naturalistic environments, generating detailed kinematic data that describe gait dynamics over extended periods. These wearable systems provide the technological foundation for transforming

subjective clinical observations into quantitative digital biomarkers of motor function.

However, the large volumes of time-series data produced by wearable sensors require computational approaches capable of identifying meaningful patterns within complex and noisy signals. Artificial intelligence and machine learning methods have therefore become central to the analysis of wearable data in Parkinson's disease. By learning representations of motor behavior directly from sensor signals, these methods can automatically detect motor symptoms such as tremor, bradykinesia, dyskinesia, and freezing episodes. In particular, deep learning architectures have shown strong potential for modeling the temporal structure of human movement and capturing subtle transitions in gait dynamics associated with FoG.

Within this broader context, human activity recognition (HAR) provides an important methodological foundation. Originally developed to classify everyday movements using wearable sensors, HAR frameworks offer signal-processing pipelines, feature representations, and modeling strategies that are directly applicable to clinical gait analysis. Many FoG detection approaches can therefore be viewed as specialized HAR problems, where the objective is to identify pathological motor states rather than general activities.

Despite these advances, an important limitation emerges when moving from activity recognition to FoG detection. While HAR models can achieve high performance in recognizing structured motor activities, their ability to generalize across individuals is more limited when applied to pathological events such as freezing episodes. FoG is not only rare and highly imbalanced, but also strongly subject-specific, with substantial variability in its manifestation across patients.

In this context, transfer learning and fine-tuning strategies provide a natural extension. By first learning general representations of motor behavior at the population level and subsequently adapting them to individual subjects, these approaches aim to bridge the gap between global modeling and patient-specific variability. Fine-tuning enables models to retain shared knowledge about gait dynamics while adjusting to the unique motor signatures of each individual, offering a practical pathway toward personalized FoG detection in real-world settings.

However, movement data collected through wearable sensors are inherently heterogeneous. Motor patterns vary substantially across patients due to differences in disease progression, physical condition, and behavioral context. In addition, variations in sensor placement, device characteristics, and daily-life environments introduce further complexity into the recorded signals. These factors make it difficult for a single monolithic model to capture all possible patterns of Parkinsonian gait.

Architectural paradigms such as mixture-of-experts (MoE) have recently gained attention as a way to address this challenge. By combining multiple specialized models through input-dependent routing mechanisms, MoE architectures allow different components of the system to focus on distinct movement patterns from different anatomical checkpoints. Such approaches provide a flexible framework for modeling complex and variable motor data collected from wearable devices.

Together, advances in wearable sensing, machine learning, and scalable model architectures are reshaping how Parkinsonian motor symptoms can be studied and monitored. These developments provide the methodological and technological context within which new approaches for automated FoG detection can be explored.

The remainder of this thesis is organized as follows. Chapter 2 presents the scientific background and reviews the related literature on wearable sensing, activity recognition, and machine learning approaches for Parkinson's disease and Freezing of Gait detection. Chapter 3 outlines the research objectives that guide the work presented in this thesis. Chapter 4 provides an overview of the wearable datasets relevant to Parkinson's disease motor analysis, discussing their sensing configurations, experimental protocols, and annotation strategies. Building on this dataset landscape, Chapter 5 introduces the methodological framework adopted in this thesis together with the experimental evaluations and results used to assess the proposed approaches.

Chapter 6 discusses the results in a broader clinical and methodological context, highlighting the implications, limitations, and future directions for wearable-based detection of Freezing of Gait in Parkinson's disease.

Finally, Chapter 7 summarizes the main findings of the thesis and outlines concluding remarks together with perspectives for future research.

Chapter 2

Background and Related Work

The automatic detection of Freezing of Gait (FoG) using wearable sensors has emerged as an active research topic at the intersection of clinical neuroscience, wearable sensing, and machine learning. While substantial progress has been made in each of these fields, translating methodological advances into reliable and clinically meaningful FoG detection systems remains challenging. Several factors contribute to this difficulty, including the intrinsic variability of Parkinsonian motor behavior, the episodic and context-dependent nature of freezing events,

This chapter reviews the scientific literature through three methodological perspectives that are particularly relevant for FoG detection using wearable technologies. First, Human Activity Recognition (HAR) approaches are examined to understand how wearable sensor data can be used to model motor activities in people with Parkinson's disease. Second, studies employing transfer learning and fine-tuning strategies are reviewed in order to explore how machine learning models can operate effectively despite limited annotated data. Third, research related to Mixture-of-Experts (MoE) architectures is analyzed to assess their potential for managing heterogeneous sensing conditions and complex motion patterns.

By organizing the literature along these three dimensions, this chapter aims to clarify the current state of the art while identifying methodological limitations that remain open in the field. These observations provide the context necessary for understanding the research directions explored in the subsequent chapters.

2.1 Human Activity Recognition in Parkinson's Disease

Human Activity Recognition (HAR) represents one of the foundational methodological frameworks for analyzing wearable motion data. Although HAR research has been extensively developed in the context of healthy populations, its application to clinical conditions such as Parkinson's disease remains comparatively limited. For this reason, the literature reviewed in this section focuses on studies that explicitly evaluate activity recognition approaches in people with Parkinson's disease (PwPD), with the goal of understanding how established HAR methodologies translate to pathological movement patterns.

Despite the rapid evolution of Human Activity Recognition (HAR) methodologies, their translation to clinical populations such as people with Parkinson's disease (PwPD) remains comparatively limited [76, 77]. The majority of HAR research has been developed and validated on healthy, often young, participants, whose movement patterns are typically regular and repeatable. Such settings do not reflect the variability, compensatory behaviors, and motor impairments that characterize PD. Only a restricted number of investigations have explicitly addressed HAR within this clinical context, and these works represent important but still isolated efforts in a domain that continues to face substantial methodological challenges.

Albert et al. [78] collected data from 18 healthy subjects (13 young and 5 older adults) and 8 PwPD, using a smartphone carried in the front pocket with a sampling rate of 20 Hz. Participants performed basic activities including sitting, standing, holding the phone, and walking. Signals were segmented into 10 s windows, from which temporal and spectral features were derived. Support Vector Machines (SVM) and Sparse Multinomial Logistic Regression (SMLR) were evaluated under random sample-wise 10-fold cross-validation. The SVM achieved the best performance, reaching 96.1% accuracy in healthy subjects and 92.2% in PwPD. However, when the classifier was trained on healthy individuals and tested on PwPD, accuracy dropped dramatically to 60.3%. Similarly, a leave-one-subject-out (LOSO) strategy reduced performance to 86% in healthy subjects and 75.1% in PwPD, emphasizing the impact of inter-subject variability.

Aswar et al. [79] investigated 8 healthy participants and 8 PwPD equipped with three IMUs placed on the chest, forearm, and lower leg, sampled at 62.5 Hz. Subjects executed 21 predefined exercises. To mitigate imbalance, the dataset was augmented

through SMOTE and divided using a random 70/30 sample-wise split. Among the tested models, a Random Forest classifier achieved the highest performance, with F1-scores of 94.2% in healthy subjects and 96% in PwPD. Nevertheless, cross-testing revealed severe generalization limitations: training on healthy data and evaluating on PD reduced the F1-score to approximately 45–50%.

Davidashvilly et al. [80] combined information from two public datasets of healthy young individuals (PAMAP2 and MHEALTH) with a proprietary cohort of 14 PwPD. Accelerations from wrist and ankle sensors were processed to recognize between 6 and 18 activities, with only a small overlap across datasets. A one-dimensional convolutional neural network (1D-CNN) was enhanced through a Domain Adversarial Neural Network (DANN) to reduce distribution mismatch. While LOSO validation yielded accuracies between 80% and 86% in the healthy datasets, cross-testing on PwPD resulted in values between 65% and 72%, again highlighting the difficulty of transferring models to pathological gait.

More recently, Gonçalves et al. [81] focused exclusively on PwPD, instrumenting 18 participants with a single lower-back IMU sampling at 100 Hz. Seven daily activities were considered, including walking, turning, and postural transitions. Raw signals were normalized and provided to an optimized 1D-CNN architecture. Using a random hold-out split, the authors reported an accuracy of 91.1% and an F1-score of 90.6%. Although encouraging, these results rely on sample-wise separation and therefore may not fully reflect subject-independent performance.

Taken together, these studies consistently show that models developed on healthy individuals do not easily generalize to PwPD. Even within PD-only cohorts, recognition performance is typically lower, mirroring the heterogeneity of motor manifestations and the influence of disease progression. Furthermore, most investigations are based on small samples and limited sensor configurations, which restrict the representation of the wide spectrum of motor behaviors encountered in clinical practice.

Evaluation methodology represents an additional source of concern. Random shuffling strategies risk mixing information from the same individual across training and testing partitions, potentially inflating performance metrics [82]. Subject-independent validation remains less common, despite being essential for assessing real-world usability. Overall, current evidence indicates that achieving robust, transferable HAR in PD requires both richer datasets and methodological frameworks explicitly designed to cope with variability.

2.2 Transfer Learning and Fine-Tuning Approaches for Freezing of Gait

A systematic methodology was adopted to examine the literature comprehensively. This involved utilizing the research query outlined in Table 2.1 through Scopus. The query is specifically designed to capture relevant studies that propose methodologies based on fine tuning or transfer learning aiming to recognize the freezing of gait. Furthermore, it is important to note that we did not discriminate by the used sensing technology. This means that the query output could also represent other types of sensing technologies, rather than wearables, such as video, audio, or physiological sensor-based research.

Operator	**Keywords**
	freezing of gait OR FoG
AND	(fine tuning OR fine-tuning OR transfer learning)

Table 2.1. Defined research for fine-tuning approaches on FoG detection using IMUs

Through this approach, 14 relevant papers were identified, but only four delved specifically into the recognition of FoG through fine-tuning or transfer learning.

In [83], the authors explored the use of unlabeled data in the fine-tuning step. In particular, starting from the publicly available Daphnet FoG dataset [84], which comprises data collections of 10 participants wearing three acceleration sensors, authors analyzed the models' performances over different window sizes of pre-FoG - from 0.75 to 9 s - and three different fine-tuning approaches. They achieved an average FoG predictions of 94.7%, 82.9%, and 68.1% in different patients, respectively, 1 second, 3 seconds, and 5 seconds before the FOG onset. They demonstrated that Recurrent Neural Networks (RNNs) can succeed in this task. However, their choice to shuffle data randomly and to include all FoG instances of their experiments in the training set can be questionable and pose challenges regarding the model performance metrics, as they point out.

Another work where the Daphnet dataset was exploited is [85]. In this study, authors used data from 5 participants and time windows of 1, 3, and 5 seconds to train an Long Short-Term Memory (LSTM) network in two ways, first using the same model in the training and fine-tuning step, then freezing the model after the training and adding one layer in the fine-tuning step. Noteworthy, they also used different

portions of the target patient data, from 10% to 50%, to simulate the scarcity of data that can often afflict these studies. The second fine-tuning approach achieved the best performances, with over 90% of prediction accuracy even in the presence of 10% of target user data in 1-second windows. However, the choice of such a limited pool of patients and the 50-50 random split into train and set might have biased the results.

Deviating from the strand of LSTM, in [86] the authors developed a one-class classification approach to predict FoG as an "anomaly in the gait" in contrast to normal gait (NG). They used data from 7 participants wearing two ankle sensors providing tri-axial acceleration and angular velocity in a leave-one-out cross-validation training of a CNN model. In this phase, both FoG and NG data are used. The layers of the model were then frozen apart, the last three layers, which are retrained on target patient data featuring only NG. This approach manages to obtain an overall sensitivity of 63% but a specificity of 98.6%. Despite the innovative design, the limited analysis to only one-time window (2s) and the limited visual FoG labeling task to only one doctor undermine the robustness of their studies. Also, they randomly split data between training and test.

In [87] authors propose two different CNN models trained respectively on IMU (two sensors on shank and forearm, with accelerometer and gyroscope) and EEG data on eight participants from a multi-modal public dataset. The two models are combined to give a unique output and are trained on a time window of 0.128s in different prediction horizons (1, 2, 3, 4, and 5s before FoG). To train the two models, they evaluated two approaches: the first is a standard stratified 5-Folds cross-validation, and the second instead applies firstly the same cross-validation on data of 7 subjects, then fine-tunes the model on the remaining subject. The ensemble model obtained mean accuracy and mean F1-score of 80.4% and 0.72 in the first approach and peaks at 92.1 % accuracy and 0.85 F1-score in 1s horizon after fine-tuning through the second.

In summary, all the works presented here share some features:

- reduced dataset size,
- random split of data, which forces data from the same patient to be both in training and test partitions and
- use of public datasets (except [86]).

On the other hand, our work proposes an extensive and exhaustive approach to appropriately address the weaknesses and the challenges of model fine-tuning for the

recognition of FoG.

2.3 Mixture-of-Experts Architectures for Wearable Sensor-Based FoG Detection

To investigate whether Mixture-of-Experts (MoE) architectures have already been explored for wearable-based detection of Freezing of Gait (FoG), a targeted literature search was conducted using the Scopus database 2.2. The query combined terms related to MoE architectures (e.g., *mixture-of-experts*, *gating networks*) with keywords associated with inertial sensing technologies (*IMU*, *accelerometer*, *gyroscope*) and Parkinson's disease or FoG. Despite the growing interest in deep learning approaches for FoG detection, this search returned no publications explicitly applying Mixture-of-Experts models to FoG recognition from wearable inertial sensors.

Field	**Query**
TITLE-ABS-KEY	(("mixture of experts" OR "mixture-of-experts" OR MoE OR "expert mixture*" OR "gating network*") AND (IMU OR "inertial measurement unit*" OR "inertial sensor*" OR "accelerometer*" OR "gyroscop*") AND ("Parkinson*" OR "PD") AND ("Freezing-of-gait" OR "FoG" OR "Freezing"))

Table 2.2. Defined research for MoE frameworks for FoG detection using wearable sensors

Given the absence of studies directly addressing this problem, the scope of the review was broadened in two directions. First, works combining MoE architectures with inertial sensing technologies were examined, regardless of the specific application domain. Second, studies employing MoE approaches within Parkinson's disease–related contexts were considered, even when they do not rely on wearable sensing modalities. This strategy makes it possible to understand how expert-specialization frameworks have been applied in related fields and to identify methodological principles that could potentially be transferred to FoG detection.

Within the domain of inertial sensing, MoE architectures have been explored as a way to manage heterogeneous motion signals or distinct movement regimes. For example, previous studies have applied expert-specialization frameworks to tasks such as trajectory reconstruction from IMU data, where different experts are

trained to model motion patterns characterized by different dynamics or sensing conditions in handwriting gestures [88]. Similarly, MoE-based ensemble strategies have been used in human–machine interaction systems combining physiological signals and inertial measurements, where multiple specialized models are aggregated to improve robustness in the presence of noisy or heterogeneous sensor inputs [89]. Although these works address problems unrelated to Parkinson's disease, they illustrate how MoE architectures can effectively decompose complex sensor signals into complementary expert representations.

MoE frameworks have also been investigated in research related to Parkinson's disease, albeit in contexts not involving wearable inertial sensing. For instance, mixture-of-experts approaches have been used to model heterogeneous relationships between biological or behavioral variables and Parkinson's disease outcomes, such as interactions between diet, microbiome composition, and disease status [90]. In another line of work, MoE architectures have been applied to speech analysis tasks involving Parkinsonian speech, where experts trained on different datasets are dynamically combined to improve generalization across heterogeneous corpora [91]. More broadly, MoE strategies have recently been integrated into medical deep learning systems for neurological disorders, enabling multi-task analysis of brain imaging data by dynamically selecting specialized submodels for different diagnostic objectives [92].

Although these studies demonstrate the versatility of MoE architectures for handling heterogeneous biomedical data and complex sensor signals, none specifically address the challenges associated with wearable-based detection of Freezing of Gait in Parkinson's disease. In particular, the literature lacks MoE architectures designed to manage variability arising from different sensor placements on the body. This gap highlights the need for adaptive architectures capable of learning complementary representations from multiple anatomical viewpoints. The approach proposed in this thesis addresses this limitation by introducing a Mixture-of-Experts framework in which each expert specializes in motion patterns captured at a specific body location, enabling improved robustness across heterogeneous sensing configurations.

2.4 Summary

This chapter reviewed the scientific literature from three methodological perspectives relevant to the automatic detection of Freezing of Gait (FoG) using wearable sensing technologies: activity recognition in Parkinson's disease, learning strategies for limited data scenarios, and computational models capable of handling heterogeneous

sensing conditions.

The analysis of Human Activity Recognition (HAR) studies highlights that, despite the maturity of activity recognition methods in general populations, their application to people with Parkinson's disease remains comparatively limited. Existing investigations frequently rely on small cohorts and simplified experimental settings, while evaluation methodologies do not always reflect subject-independent deployment scenarios. Moreover, models trained on healthy individuals often struggle to generalize to pathological gait patterns, emphasizing the importance of approaches specifically tailored to Parkinsonian motor variability.

Research on transfer learning and fine-tuning demonstrates the potential of adaptive learning strategies for addressing data scarcity in FoG detection. By transferring knowledge from reference datasets or pretrained models, these techniques enable the development of systems that can operate with limited labeled data and adapt to individual patients. However, the current literature remains constrained by small datasets, limited cross-cohort validation, and methodological inconsistencies in evaluation protocols.

Finally, the review of Mixture-of-Experts (MoE) architectures shows that expert-specialization frameworks have been successfully applied in a variety of domains involving heterogeneous data and complex signal structures.

Although these studies demonstrate the versatility of MoE architectures for handling heterogeneous biomedical data and complex sensor signals, their application to wearable-based detection of Freezing of Gait in Parkinson's disease has not yet been investigated. More broadly, the literature has not explored the potential of expert-specialization frameworks for modeling the variability and complexity of wearable motion signals associated with Parkinsonian gait.

This absence is notable because MoE architectures are specifically designed to decompose complex problems into specialized submodels that cooperate through adaptive routing mechanisms. Such characteristics suggest that these architectures may offer a promising framework for addressing the challenges posed by wearable-based monitoring of Parkinsonian motor symptoms.

Taken together, these observations reveal several open challenges in the development of robust FoG detection systems based on wearable sensor data. Addressing these challenges requires methodological frameworks capable of modeling complex motor behavior, learning effectively from limited annotated data, and generalizing across diverse real-world conditions.

Chapter 3

Aims of the thesis

As discussed in the previous sections, Freezing of Gait (FoG) represents one of the most disabling and complex motor symptoms of Parkinson's disease (PD). It is characterized by sudden and transient episodes in which patients are unable to initiate or continue walking despite the intention to move. These episodes occur unpredictably and are often triggered by specific situations, such as turning, walking through narrow spaces, or starting movement. FoG strongly impacts patients' mobility and independence, making its early detection and continuous monitoring essential for both clinical management and research.

As also introduced earlier, wearable sensors—particularly inertial measurement units (IMUs)—have emerged as powerful tools for monitoring motor function in PD. When coupled with artificial intelligence (AI) techniques, they enable the automatic recognition of motor activities and symptoms, offering an opportunity to move beyond subjective clinical observation toward continuous, objective, and data-driven monitoring in real-life settings. However, despite rapid progress in this field, several challenges still hinder the effective and generalizable detection of FoG from wearable data.

Obtaining high-quality data on Freezing of Gait is inherently challenging. Capturing freezing episodes requires prolonged monitoring of patients and careful annotation by clinical experts, making data acquisition expensive and time-consuming. Consequently, the amount of labeled data available for training machine learning models is often limited, which complicates the development of robust and generalizable detection systems.

Within this context, the overarching goal of this Ph.D. thesis is to advance the automatic detection of Freezing of Gait in Parkinson's disease through the development of AI-based models that are robust, generalizable, and clinically

meaningful. The research follows a coherent progression of steps that collectively address the main limitations of the state of the art — from understanding data foundations to designing adaptive computational solutions.

More specifically, the work pursues the following aims:

- **To systematically review the principal wearable datasets available for Freezing of Gait analysis.**

  A first step toward robust FoG detection consists in understanding the data landscape on which machine learning models are trained and evaluated. Existing wearable datasets differ in terms of cohort size, sensor technologies, acquisition protocols, and annotation procedures. Such differences influence both the type of motor behaviors captured and the methodological assumptions underlying model development.

  Therefore, this thesis begins with a detailed review of the principal datasets used for wearable-based FoG detection. The objective is to characterize their structure, sensing configurations, activity protocols, and annotation strategies, highlighting both their strengths and limitations. By clarifying the commonalities and discrepancies among available resources, this analysis provides the empirical foundation for the methodological choices explored in the remainder of the thesis.

- **To investigate activity recognition as a structural foundation for FoG analysis.**

  The literature reviewed in the previous chapter highlights that Freezing of Gait rarely occurs in isolation but rather emerges within specific motor contexts, such as gait initiation, turning, or transitions between postures. These activities constitute the behavioral framework within which freezing episodes develop. Modeling the structure of everyday motor activities is therefore essential for understanding the temporal dynamics of Parkinsonian gait and for identifying situations that increase the likelihood of freezing.

  In this thesis, activity recognition is treated not merely as an auxiliary task but as a modeling layer that enables contextual reasoning for subsequent FoG detection. By analyzing wearable inertial signals during different motor activities, we investigate how accurately IMU-based systems can characterize movement primitives and transitions. The objective is to extract representations that remain stable across subjects while preserving sensitivity to clinically

meaningful variability. Such representations can later support FoG-oriented models by providing contextual information about the ongoing motor activity.

- **To cope with data scarcity by leveraging fine-tuning approaches for FoG detection.**

  Another challenge emerging from the literature concerns the limited availability of annotated freezing events for training machine learning models. Freezing episodes occur intermittently and unpredictably, and their identification typically requires careful annotation by clinical experts. As a result, many existing studies rely on relatively small amounts of labeled data, which can limit the generalization capability of learned models.

  This thesis addresses this issue through a two-stage learning strategy. First, a model is trained to capture general FoG dynamics from a structured dataset where experimental conditions and annotation procedures are well defined. This stage aims to learn invariant representations that describe freezing behavior at the population level. In a second step, the pretrained model is fine-tuned using data from a single target individual. Rather than requiring extensive subject-specific recordings, this approach investigates how limited calibration data can adapt a global representation to personal motor signatures. The objective is not only to improve detection accuracy for individual patients but also to explore a realistic pathway toward deployable FoG monitoring systems.

- **To develop adaptive architectures that generalize across body locations and sensing configurations.**

  One of the strongest sources of variability emerging from the reviewed datasets concerns sensor placement. Signals acquired at the ankle, lower back, wrist, or thigh describe the same motor event through different biomechanical perspectives. Models trained on a single configuration tend to internalize that viewpoint, making them fragile when transferred to alternative placements.

  To address this issue, the thesis proposes a Mixture-of-Experts (MoE) framework in which each expert specializes in interpreting motion patterns from a specific body location. Instead of forcing a single model to homogenize heterogeneous inputs, specialization allows each expert to learn the dynamics, noise characteristics, and discriminative cues typical of its anatomical viewpoint. A gating mechanism then combines these experts, dynamically weighting their contributions according to the available sensors.

This design enables knowledge learned from multiple datasets to become reusable in another whenever similar body positions are present, even if protocols, subjects, or environments differ. The approach therefore promotes modularity, facilitates integration of new datasets, and supports scalable deployment scenarios in which wearable configurations may change over time.

In summary, this thesis aims to establish a unified methodological framework for AI-driven detection of Freezing of Gait, grounded in the practical realities of the current dataset ecosystem. The work progresses from contextual modeling of motor activities, to efficient exploitation of high-quality annotations, and finally to adaptive integration of motion information captured from different body locations.

For each of these objectives, the thesis first analyzes the available wearable datasets in order to understand their sensing configurations, activity protocols, and annotation strategies. This datasets' review provides the empirical foundation for the methodological developments that follow.

Subsequently, the thesis introduces the proposed computational approaches and evaluates them experimentally, illustrating how contextual modeling of motor activities, data-efficient learning strategies, and position-aware integration of wearable motion signals can contribute to more robust monitoring of Parkinson's disease.

3.1 Publications related to the thesis contributions

The work presented in this thesis builds upon a set of research contributions developed during the Ph.D. These publications are aligned with the main objectives outlined before and are discussed in detail in the following chapters.

3.1.1 Dataset analysis and characterization

The analysis of wearable datasets for Parkinson's disease presented in Chapter 4 is grounded in the following contributions:

1 **M. Tebaldi** , L. Borzì, G. Olmo, R. Giugno, G. Pravadelli and F. Demrozi, *Exploring Parkinson's Disease Datasets: Key Findings*, IEEE EMBC, 2025 [93].

2 L. Borzì, F. Demrozi, R. A. Bacchin, C. Turetta, **M. Tebaldi**, L. Sigcha, S. Zolfaghari, D. Rinaldi, G. Fazzina, G. Balestro, A. Picelli, G. Pravadelli, G.

Olmo, S. Tamburin, L. Lopiano and C. A. Artusi, *A multi-level annotated sensor dataset of gait freezing manifestations and severity in Parkinson's disease*, Scientific Data, 2026 [94].

The first work investigates the structure and variability of wearable datasets, with a focus on sensing configurations, acquisition protocols, and annotation strategies, providing the empirical basis for the analysis developed in Chapter 5.

The dataset landscape is further enriched by the inclusion of the recently released *FoG-STAR* dataset, which provides multi-level annotations of FoG episodes, motor activities, and symptom manifestations from multiple sensor locations. This resource reflects recent efforts toward more structured and context-aware wearable datasets.

3.1.2 Activity recognition as a structural layer for FoG analysis

The investigation of Human Activity Recognition (HAR) in Parkinson's disease, presented in Chapter 5.1, builds upon:

3 L. Borzì, **M. Tebaldi**, F. Demrozi, *AI-based recognition of daily activities in Parkinson's disease using wearable inertial sensors*, FAIEMA 2025 [95].

This work demonstrates that clinically relevant motor activities can be recognized with high accuracy (up to 91–92% F1-score) under a subject-independent protocol using a large-scale wearable dataset. These findings support the use of activity recognition as a structural layer for contextualizing FoG detection.

3.1.3 Data-efficient learning through fine-tuning for FoG detection

The study of fine-tuning strategies for subject-specific FoG detection, presented in Chapter 5.2, builds upon:

4 **M. Tebaldi**, G. Pravadelli, F. Demrozi, R. Giugno and C. Turetta, *Enhancing Freezing of Gait Detection in Parkinson's Through Fine-Tuned Deep Learning Models*, IEEE ICDH, 2024 [96].

This work investigates transfer learning approaches for adapting FoG detection models to individual patients using limited data. The results show that, while population-level models capture general freezing patterns, fine-tuning improves alignment with subject-specific motor signatures, highlighting both the potential and the limitations of personalized adaptation.

3.1.4 Mixture-of-Experts modeling for heterogeneous sensor configurations

The Mixture-of-Experts (MoE) framework presented in Chapter 5.3 represents an exploratory methodological contribution developed during the Ph.D. While no dedicated publication is currently associated with this work, it constitutes an initial investigation into position-aware modeling for wearable FoG detection. Further development and validation of this approach are ongoing, and its dissemination in a peer-reviewed venue will be evaluated in future work.

Chapter 4

Datasets' landscape

As outlined in the research objectives, the development of reliable computational methods for detecting and analyzing motor symptoms in Parkinson's disease depends strongly on the characteristics of the data used for training and evaluation. Wearable sensing technologies have enabled the collection of detailed motion signals in both controlled experiments and real-world environments, providing new opportunities for quantitative and data-driven analysis of Parkinsonian motor behavior.

In this context, wearable-based datasets represent a fundamental resource for the study of gait impairments, postural transitions, and complex motor phenomena such as freezing of gait (FoG). However, these datasets differ substantially in terms of sensing modalities, experimental protocols, annotation strategies, and clinical characterization. Such differences reflect the diversity of research goals across studies but also introduce methodological challenges when comparing results or developing generalizable algorithms.

Publicly available datasets play a particularly important role by supporting reproducibility, enabling cross-study comparisons, and facilitating benchmarking of computational approaches. Over the past decade, several datasets have been released, ranging from early laboratory-based collections focused on specific gait tasks to more recent multimodal or semi-free-living recordings capturing a broader spectrum of daily activities. Despite this progress, the available resources remain heterogeneous and often emphasize different aspects of Parkinsonian motor behavior.

This chapter therefore presents a structured overview of wearable-based datasets relevant to Parkinson's disease motor analysis, with particular emphasis on gait and FoG-related phenomena. Rather than reviewing sensing technologies or clinical concepts—which were introduced in Chapter 1—the focus here is on the datasets themselves as methodological resources. Each dataset is examined in terms of

cohort characteristics, sensing setup, experimental context, and annotation strategy, highlighting both their strengths and their limitations.

In addition to surveying existing resources, this chapter introduces the FOGSTAR dataset (Freezing of Gait Severity, Tasks, Activities, and Ratings) [94]. This dataset was developed to address several limitations observed across prior studies and is positioned within the broader dataset landscape described in this chapter.

4.1 Dataset Inclusion Criteria and Scope

The dataset landscape for Parkinson's disease motor assessment is broad and heterogeneous, spanning diverse sensing technologies, clinical objectives, and experimental contexts. To ensure coherence and relevance, this chapter adopts a set of explicit inclusion criteria that define the scope of datasets considered and clarify the perspective from which they are analyzed.

First, only datasets based on wearable sensing technologies are included. Specifically, the chapter focuses on datasets employing body-worn inertial sensors (e.g., accelerometers, gyroscopes, inertial measurement units), either as standalone modalities or in combination with complementary sensing systems such as pressure insoles, physiological sensors, or vision-based reference systems. Datasets relying exclusively on non-wearable modalities—such as optical motion capture, force platforms, imaging techniques, or questionnaire-only assessments—are excluded, as they fall outside the wearable-centered methodological framework of this thesis.

Second, all included datasets involve participants diagnosed with Parkinson's disease and contain recordings of motor behavior relevant to mobility or gait-related symptoms. While some datasets also include healthy control cohorts, only the Parkinson's disease data are considered for the purposes of this chapter. The primary focus is placed on datasets addressing gait impairments, postural transitions, and freezing of gait, although datasets targeting broader motor symptomatology (e.g., bradykinesia or dyskinesia) are included when wearable-derived movement signals constitute a central component of the data.

Third, datasets are required to provide time-resolved sensor data suitable for algorithmic analysis. This includes continuous or segmented time series recorded during structured tasks, semi-controlled protocols, or free-living activities. Datasets offering only aggregated summary scores or coarse outcome measures without access to the underlying sensor signals are not considered. Annotation strategies may vary across datasets—from frame-level event labels to window-based motor

state ratings—but some form of temporally aligned reference information must be available.

Fourth, the chapter emphasizes publicly accessible or well-documented datasets that have been described in peer-reviewed literature. While access conditions may differ (e.g., fully open, controlled access, or application-based release), all datasets included here are sufficiently documented to allow independent reuse, methodological replication, or benchmarking. Proprietary datasets without public descriptions are excluded.

Within these boundaries, no restrictions are imposed on experimental setting or recording duration. As a result, the chapter encompasses datasets acquired in controlled laboratory environments, semi-structured clinical protocols, and real-world or free-living contexts, as well as datasets spanning short scripted tasks, session-based recordings, and multi-day monitoring. This breadth is intentional and reflects the diverse methodological trade-offs present in the field.

Finally, it is important to note that this chapter is dataset-centric rather than algorithm-centric. The inclusion criteria are defined with the goal of characterizing the data resources available to the research community, not of evaluating specific detection or classification approaches. Methodological implications arising from dataset properties are discussed later in the chapter, while algorithmic development and experimental validation are addressed in subsequent chapters.

4.2 Datasets insights

This subsection presents a structured overview of the datasets that satisfy the inclusion criteria defined above. Each resource is described individually, focusing on cohort composition, sensing configuration, experimental protocol, and annotation strategy. The aim is to provide a concise yet informative reference that clarifies the practical characteristics of each dataset before moving to the comparative analyses developed later in the chapter.

4.2.1 Weargait-PD

The WearGait-PD dataset [97] includes data from 60 PwPD (25 female, 35 male, mean age 66 $\pm$ 9 years) and 65 age-matched controls (40 female, 25 male, mean age 76 $\pm$ 9 years). PwPD presented with a mean modified Hoehn and Yahr (H&Y) stage of 2.1 $\pm$ 0.5 and Movement Disorder Society–Unified Parkinson's Disease

Rating Scale (MDS-UPDRS) Part III scores ranging of 24 ± 10. Inclusion criteria included a clinical diagnosis of PD for patients and no neurological disorders for healthy subjects, with exclusions made only for conditions preventing safe performance of tasks. Demographic and clinical information were collected for each participant, including age, sex/gender, height, weight, race, years since diagnosis, medication types and doses, time since last dose, physical/occupational therapy, presence and configuration of deep brain stimulation (DBS), and clinical rating scales.

Sensors: Each participant was equipped with multiple synchronized systems, described in the following.

- **Wearable inertial sensors:** 13 wireless Xsens MTw Awinda IMUs were placed at the mid-forehead, sternum (xiphoid process), lower back (L4/L5), wrists, lateral thighs, lateral shanks, ankles, and dorsum of each foot. The IMUs captured 3D acceleration, gyroscope, magnetometer, and orientation with a sampling rate of 100 Hz.
- **Smart pressure insoles:** Moticon OpenGo insoles in both shoes recorded plantar pressure (16 sensors per insole) and inertial data (3D acceleration and gyroscope) with a sampling rate of 100 Hz.
- **Instrumented walkway:** ProtoKinetics Zeno Walkway (16 ft) recorded footfalls and pressure distribution using 18,432 sensors at 100 Hz.
- **Video-recording:** GoPro HERO10 cameras captured frontal and lateral video views at 1080p and 50 fps. All devices were synchronized via TTL pulse (IMUs and walkway), acoustic signal converted to TTL for insoles, and visual LED for video. All data streams were temporally aligned post-hoc and validated.

Sensor Type	Location
IMU	Mid forehead, sternum, lower back (L4/L5)
	Bilateral wrists, thighs, shanks, ankles, feet dorsum
Insole	Plantar surface (16 pressure sensors per foot)
Walkway	Entire 16–ft mat area
Camera	Frontal and lateral angles (50 fps)

Table 4.1. Sensor Placement Summary

Tasks: Participants were asked to perform a standardized sequence of tasks, listed and descripted in Tab. 4.36 Each task was recorded with all sensor modalities and annotated frame-by-frame with general (e.g., walk, turn, sit) and clinical events (e.g., freezing, staggering).

Task name	Acronym	Description
SelfPace	SP	Walk at a self-selected pace across the mat
HurriedPace	HP	Walk at a hurried pace
SelfPace_mat	SPm	On-mat self-paced walk
HurriedPace_mat	HPm	On-mat hurried walk
SelfPace_-matTURN	SPmT	On-mat self-paced walk with alternating turns
TandemGait	TG	Tandem heel-to-toe walk
Timed Up and Go	TUG	Standard TUG task with chair
Balance	B	Six static stance conditions with eyes open/closed and varying foot positions
SelfPace_doorpat	SPdoorpat	Walk through a mock doorway
FreeWalk	FW	Walk through hallways and perform real-world mobility maneuvers

Table 4.2. List of pre-scripted tasks performed by participants.

The dataset is available in `CSV` and `MAT` formats:

- CSV files: One per task per participant, containing synchronized time-series from all sensors.
- MAT files: One per participant, with eight task-specific structures storing annotations, walkway, insole, and IMU data.
- Clinical spreadsheets: Metadata for each participant including MDS-UPDRS item scores and demographics.

Each time series is sampled at 100 Hz and contains up to 346 variables. Missing data is represented as `NaN`. Annotations use standardized labels (e.g., ´´unlabeled" when no event was active).
All data were preprocessed, quality-checked, and made publicly available via the Synapse repository (ID: syn52540892).

4.2.2 Multimodal

The Multimodal Freezing of Gait (FoG) dataset [98] includes data from 12 individuals with Parkinson's disease (6 female, 6 male, mean age 69.1 $\pm$ 7.9 years), all of whom experienced freezing of gait during off-medication states. Participants had a mean disease duration of 9.3 $\pm$ 6.8 years and exhibited clinically significant FoG episodes, with Unified Parkinson's Disease Rating Scale (UPDRS) Part III scores of 45.0 $\pm$ 16.0. Inclusion criteria required a clinical diagnosis of idiopathic PD and the presence of FoG during off-time, while exclusion criteria included other neurological, orthopedic, or sensory impairments affecting gait.

Sensors: Each participant was equipped with a synchronized multimodal sensing system, as described below.

- **Electroencephalography (EEG):** 25 EEG channels were recorded using a 32-channel wireless MOVE system (Brain Products GmbH) following the international 10–20 system, with TP9 and TP10 as reference electrodes. Signals were sampled at 1000 Hz.
- **Electromyography (EMG):** Three EMG channels were acquired from the gastrocnemius muscle of the right leg and tibialis anterior muscles of both legs, sampled at 1000 Hz using the same MOVE system.
- **Wearable inertial sensors:** Four inertial measurement units (TDK MPU6050) were mounted on the lateral tibia of both legs, the fifth lumbar vertebra (L5), and the left wrist. Each IMU recorded 3D acceleration and angular velocity at 500 Hz.
- **Skin conductance (SC):** Skin conductance was recorded from the index and middle fingers of the left hand using integrated LM324-based sensors at 500 Hz.
- **Video-recording:** All experiments were video-recorded to enable expert annotation of freezing episodes. Multimodal data streams were time-aligned using hardware timestamps and synchronized post-hoc.

Tasks: Participants performed two walking paradigms designed to elicit freezing of gait, each repeated twice, yielding four tasks in total. Tasks involved walking through narrow corridors, bypassing obstacles, and executing 180-degree turns, as well as turning within confined square regions. Participants initiated each trial from a seated

Sensor Type	Location
EEG	Frontal, central, parietal, occipital scalp regions (10–20 system)
EMG	Right gastrocnemius; bilateral tibialis anterior
IMU	Bilateral shanks, L5 (waist), left wrist
SC	Index and middle fingers (left hand)
Camera	Frontal view for annotation

Table 4.3. Sensor Placement Summary

position and completed sit-to-stand and stand-to-sit transitions during the tasks. All tasks were performed without external intervention and recorded continuously with all sensor modalities. FoG episodes were annotated by two neurologists based on synchronized video recordings.

Task name	Acronym	Description
CorridorWalk	CW	Walk through narrow corridor with obstacles and 180° turns
CorridorWalk_rep	CW2	Repetition of corridor walking task
SquareTurn	ST	Walk and turn inside confined square area
SquareTurn_rep	ST2	Repetition of square turning task

Table 4.4. List of pre-scripted tasks performed by participants.

The dataset is available in raw and preprocessed formats:

- **Raw data:** EEG, EMG, IMU, and SC signals stored in modality-specific formats (EEG/EMG in BrainVision format; IMU and SC in CSV files).
- **Processed data:** Task-wise synchronized and filtered signals stored in `TXT` files with unified sampling rate of 500 Hz.
- **Annotations:** Binary labels indicating freezing of gait (1) or non-freezing (0) for each time sample.

Each time series contains up to 60 variables including EEG channels, EMG channels, inertial data, skin conductance, and labels. Missing sensor channels for some participants are filled with zeros. Signals were band-pass filtered and notch-filtered to remove power line interference.

All data were preprocessed, quality-checked, and made publicly available via the Mendeley Data repository. For the purposes of this study, we focused exclusively on the inertial sensor data collected from the shank- and waist-mounted IMUs, excluding EEG, EMG, and skin conductance modalities. This decision was made to ensure comparability with wearable-only sensing approaches and to improve feasibility for long-term ambulatory monitoring.

The dataset contains a total duration of 3 h 42 min of valid recordings, including 88 min of freezing episodes across 334 annotated FoG events. FoG event durations range from 1 to 201 s, with a strong class imbalance between freezing and non-freezing segments.

4.2.3 CuPiDMultimodal

The CuPiD multimodal dataset [99] includes data from 18 individuals with Parkinson's disease, of whom 11 exhibited freezing of gait (FoG) episodes during the experimental protocol. Participants had a mean age of 68.9 $\pm$ 10.2 years and a mean disease duration of 8.8 $\pm$ 4.6 years, with Unified Parkinson's Disease Rating Scale (UPDRS-III) scores ranging from 25 to 56 and Freezing of Gait Questionnaire (FOG-Q) scores from 4 to 30, corresponding to Hoehn and Yahr stages II–IV. Inclusion criteria required a clinical diagnosis of PD and self-reported FoG, while exclusion criteria included psychiatric comorbidities, history of stroke, traumatic brain injury, or other neurological disorders.

Sensors: Each participant was equipped with a synchronized multimodal sensing platform, described below.

- **Electrocardiography (ECG):** ECG and 3D acceleration were recorded using an Actiwave wearable sensor at a sampling rate of 512 Hz. The ECG electrodes were attached to the chest.

- **Skin conductance (SC):** Skin conductance and 3D acceleration were recorded using a Shimmer wearable sensor at a sampling rate of 51.2 Hz. SC electrodes were attached to the index and middle fingers of one hand.

- **Wearable inertial sensors:** A network of nine IMUs was attached to multiple body segments (feet, shanks, thighs, waist, and trunk) to capture full-body kinematics during walking.

- **Video-recording:** Two synchronized camera systems (mobile HDR and fish-eye cameras) recorded all sessions to enable expert frame-level annotation of FoG episodes. All data streams were synchronized post-hoc using timestamps.

Sensor Type	Location
ECG	Chest
SC	Index and middle fingers (hand)
IMU	Bilateral feet, shanks, thighs, waist, trunk
Camera	Frontal and wide-angle views

Table 4.5. Sensor Placement Summary

Tasks: Participants performed a standardized protocol designed to provoke freezing of gait, including turning, walking through narrow passages, and cognitively demanding walking tasks. The protocol comprised the following task categories:

Task name	Acronym	Description
Ziegler Protocol	ZP	Walking task including 360° and 180° turns and narrow passage traversal, performed with and without cognitive load
Figure Eight	FE	Repeated figure-eight walking patterns in a confined area, with and without cognitive load
Straight Walk	SW	Straight-line walking with turns, with and without narrow corridor constraints
Circles	C	Circular walking with random 180° and 360° turns
Hospital Tour	HT	Free walking in hospital corridors including crowds, obstacles, and involuntary stops
Baseline	BL	Seated rest periods between walking tasks

Table 4.6. List of pre-scripted tasks performed by participants.

The dataset is organized into synchronized multimodal recordings:

- **Physiological signals:** ECG and skin conductance time series with associated accelerometry.
- **Inertial data:** 3D acceleration and angular velocity from nine IMUs.
- **Annotations:** Frame-level FoG labels derived from expert review of synchronized video recordings.

ECG data are sampled at 512 Hz and SC data at 51.2 Hz. Data streams are time-aligned and segmented using sliding windows of 3 s with 0.5 s overlap. FoG episodes are annotated with frame-level precision (40 ms).
All data were preprocessed, quality-checked, and annotated by two expert clinicians. For the purposes of this study, we focused exclusively on the physiological sensing modalities, namely ECG and skin conductance, omitting inertial and video data. This choice was motivated by the goal of investigating autonomic nervous system correlates of freezing of gait and enabling future wrist- or chest-worn prediction systems.

The dataset contains a total of 184 annotated FoG episodes with durations ranging from 0.12 to 98.88 s (mean: 8.84 s). Approximately 65% of FoG events last less than 5 s. FoG events are unevenly distributed across participants, with the majority occurring during turning tasks.

4.2.4 O'Day

The O'Day IMU dataset [100] includes data from 7 individuals with Parkinson's disease (3 female, 4 male, age range 51–68 years) who experienced freezing of gait (FoG) during a standardized walking protocol. Participants had disease durations ranging from 5 to 15 years and performed between 5 and 14 walking trials across 2 to 6 clinic visits. All participants completed the protocol off dopaminergic medication and off deep brain stimulation. Inclusion criteria required a clinical diagnosis of Parkinson's disease and the ability to walk independently without an assistive device. Written informed consent was obtained under protocols approved by the Stanford University Institutional Review Board.

Sensors: Each participant was equipped with a set of synchronized wearable inertial sensors, described below.

- **Wearable inertial sensors:** Inertial measurement units (APDM Wearable Technologies, Inc.) were placed on the chest, lumbar region (L5), and the dorsum of both feet and ankles for all participants (6 IMUs). A subset of participants (n=4) additionally wore sensors on the head, posterior wrists, and lateral thighs, yielding a total of 11 IMUs. Each IMU recorded tri-axial acceleration and angular velocity at a sampling rate of 128 Hz.

- **Video-recording:** All walking trials were video-recorded to allow expert

annotation of freezing of gait episodes. IMU and video data streams were synchronized post-hoc using timestamps.

Sensor Type	Location
IMU	Chest, lumbar (L5), bilateral ankles, bilateral feet
IMU (subset)	Head, bilateral wrists, bilateral thighs
Camera	Frontal view for annotation

Table 4.7. Sensor Placement Summary

Tasks: Participants completed a freeze-eliciting walking protocol consisting of repeated turns and obstacle navigation. Each walking trial included two ellipses and two figure-eight trajectories around tall barriers, designed to provoke freezing of gait through frequent directional changes and narrow passages. Each participant completed multiple trials per visit across multiple visits, resulting in 60 unique walking trials in total. All tasks were recorded continuously with synchronized IMU and video data and annotated for FoG events by experienced raters.

Task name	Acronym	Description
Turning and Barrier Course	TBC	Walking task including ellipses and figure-eight paths around tall barriers to elicit freezing of gait

Table 4.8. List of pre-scripted tasks performed by participants.

The dataset consists of synchronized inertial and annotation data:

- **IMU signals:** Tri-axial acceleration and angular velocity for each sensor.
- **Annotations:** Binary labels indicating freezing of gait (FoG) or non-freezing for each time window.
- **Metadata:** Participant demographics and trial identifiers.

Raw IMU data were downsampled from 128 Hz to 64 Hz and segmented into overlapping 2 s windows. Each window was labeled according to the majority human-rated label within the window. Data augmentation via random rotations was applied during model training to simulate variability in sensor orientation.
All data were preprocessed, quality-checked, and made publicly available together with analysis scripts via a GitHub repository. For the purposes of this study, we

focused exclusively on the inertial sensor data, omitting video recordings and survey responses. Furthermore, we restricted the analysis to the 6-IMU configuration (chest, lumbar, and bilateral ankles and feet) to ensure consistency across all participants.

The dataset contains a total of 60 walking trials with an overall duration of 88.8 minutes, including 211 annotated freezing of gait events, corresponding to 23.9% of the total walking time. Freezing episodes vary substantially in duration across participants and are primarily elicited during turning segments of the walking protocol.

4.2.5 Daphnet

The Daphnet Freezing of Gait dataset [101] includes data from 10 individuals with Parkinson's disease (7 male, 3 female, mean age 66.4 $\pm$ 4.8 years) with a history of freezing of gait (FoG). Participants had a mean disease duration of 13.7 $\pm$ 9.7 years and a mean Hoehn and Yahr (H&Y) stage of 2.6 $\pm$ 0.65 in the ON state. Inclusion criteria required a clinical diagnosis of idiopathic Parkinson's disease, the ability to walk unassisted in the OFF medication state, and a documented history of FoG. Exclusion criteria included severe visual or hearing impairment, dementia, and other neurological or orthopedic disorders affecting gait.

Sensors: Each participant was equipped with a set of synchronized wearable inertial sensors, described below.

- **Wearable inertial sensors:** Three tri-axial accelerometers were attached to the shank (just above the ankle), the thigh (just above the knee), and the lower back (belt-mounted). Each sensor recorded 3D acceleration at a sampling rate of 64 Hz.
- **Video-recording:** All walking trials were video-recorded to allow expert frame-level annotation of freezing of gait events. Sensor and video data streams were synchronized post-hoc using timestamps.

Tasks: Participants completed a standardized walking protocol designed to elicit freezing of gait. The protocol consisted of three task categories representing common daily walking situations:

The dataset consists of synchronized inertial and annotation data:

- **Accelerometer signals:** Tri-axial acceleration from each sensor location.

Sensor Type	Location
Accelerometer	Shank (above ankle), thigh (above knee), lower back (belt)
Camera	Frontal view for annotation

Table 4.9. Sensor Placement Summary

Task name	Acronym	Description
Straight Walk	SW	Walking back and forth in a straight line along a hallway with 180° turns
Random Walk	RW	Random walking in a hall space with multiple 360° turns and instructed stops
Activities of Daily Living	ADL	Simulated daily activities including entering rooms, walking to a kitchen, retrieving an object, and returning to the start

Table 4.10. List of pre-scripted tasks performed by participants.

- **Annotations:** Frame-level labels indicating freezing of gait or non-freezing periods.

- **Metadata:** Participant identifiers and task segment information.

Each time series is sampled at 64 Hz. Signals were segmented using sliding windows of 4 s with a step size of 0.5 s for spectral feature extraction. Freezing of gait events were annotated by expert physiotherapists based on synchronized video recordings, with event onset defined as the arrest of alternating stepping and offset defined as the resumption of the stepping pattern.
All data were preprocessed, quality-checked, and made publicly available for research use. For the purposes of this study, we focused exclusively on the accelerometer signals recorded from the three body-worn sensors, omitting video recordings and feedback-related system outputs. This choice was made to ensure consistency with lightweight wearable sensing configurations and to facilitate reproducible algorithm benchmarking.

The dataset contains a total of 8 h 20 min of recordings, including 237 annotated freezing of gait events across 8 participants. FoG episode durations range from 0.5 to 40.5 s (mean 7.3 s), with approximately 50% of events lasting less than 5.4 s and over 93% shorter than 20 s.

4.2.6 IMU

The IMU dataset [102] includes data from 35 individuals with Parkinson's disease (16 female, 19 male; age range 44–84 years) with documented freezing of gait (FoG). Participants had Hoehn and Yahr stages between 2 and 4 and a mean disease duration of 8.04 $\pm$ 4.10 years. Inclusion criteria required a clinical diagnosis of idiopathic Parkinson's disease, ability to walk independently, and absence of neurological or physical dysfunctions other than those associated with PD. Exclusion criteria included prior surgery for PD and diagnosed vestibular, visual, or somatosensory impairments. All measurements were performed in the ON medication state, with participants having taken dopaminergic medication approximately 1 h before testing. Written informed consent was obtained under protocols approved by the ethics committee of the School of Physical Education and Sport at the University of São Paulo.

Sensors: Each participant was equipped with a synchronized wearable sensing system, described below.

- **Wearable inertial sensor:** A single Physilog 5 inertial measurement unit (GaitUp, Switzerland) was placed on the shank of the most affected leg. The IMU recorded tri-axial acceleration and tri-axial angular velocity at a sampling rate of 128 Hz.
- **Video-recording:** All trials were recorded using a digital camera (Sony, 30 fps) to allow expert annotation of freezing of gait episodes. IMU and video data streams were synchronized post-hoc using timestamps.

Sensor Type	Location
IMU	Shank (most affected side)
Camera	Lower-limb frontal view for annotation

Table 4.11. Sensor Placement Summary

Tasks: Participants performed a turning-in-place task designed to elicit freezing of gait. Each trial consisted of alternating 360° turns to the right and left at a self-selected pace for a duration of 2 min. Participants completed three sessions separated by approximately one month to increase the likelihood of capturing FoG episodes. One

additional static trial (10 s standing) was recorded for sensor calibration purposes. All trials were recorded continuously with synchronized IMU and video data and annotated for FoG events by two movement disorder specialists.

Task name	Acronym	Description
Turning-in-Place	TIP	Alternating 360° turns to the left and right at self-selected pace for 2 min
Calibration	CAL	Standing upright and still for 10 s

Table 4.12. List of pre-scripted tasks performed by participants.

The dataset is organized into three main components:

- **IMU data:** Tri-axial acceleration and angular velocity stored in ASCII text files, one per participant per session.
- **Annotations:** Binary labels indicating freezing of gait (1) or non-freezing (0) for each time sample.
- **Metadata:** Clinical and demographic information including UPDRS-II/III, H&Y stage, NFoG-Q, MoCA, MMSE, and medication dosage.

Each trial contains 15,360 samples (120 s at 128 Hz) and nine variables: three accelerometer axes, three gyroscope axes, time index, frame number, and a FoG flag. Signals were low-pass filtered using a fourth-order zero-phase Butterworth filter with a 60 Hz cutoff frequency. FoG onset was defined as the arrest of the alternating turning pattern and offset as the resumption of effective stepping, based on synchronized video review.
All data were preprocessed, quality-checked, and made publicly available via the Figshare repository (DOI: 10.6084/m9.figshare.14984667). For the purposes of this study, we focused exclusively on the inertial measurement unit data, omitting video recordings and clinical questionnaire data. This decision was made to ensure consistency with single-sensor wearable FoG detection approaches and to facilitate reproducible algorithm benchmarking.

The dataset contains a total of 1,611 s of freezing of gait across all participants. On average, participants experienced 3.0 ± 2.9 FoG episodes per session, although 36% of participants did not exhibit FoG during the turning task.

4.2.7 REMPARK

The REMPARK dataset [103] includes data from 21 individuals with Parkinson's disease (3 female, 18 male; mean age 69.3 ± 9.7 years) who experienced freezing of gait (FoG) during home-based monitoring. Participants were selected from a larger cohort of 92 PD patients enrolled in the REMPARK project and met inclusion criteria of Hoehn and Yahr stage greater than 2 in the OFF state, FoG Questionnaire (FoG-Q) score above 6, and a total FoG duration of at least 1 min during the recordings. Mean Hoehn and Yahr stage in the OFF state was 3.07 ± 0.43, and mean FoG-Q score was 15.8 ± 4.1. Exclusion criteria included dementia and other neurological or orthopedic conditions affecting gait. Written informed consent was obtained under protocols approved by the corresponding ethics committees of all participating clinical centers.

Sensors: Each participant was equipped with a synchronized wearable sensing system, described below.

- **Wearable inertial sensor:** A single waist-worn inertial measurement unit (9x2 IMU, 77×37×21 mm, 78 g) was placed on the left side of the waist near the body's center of mass. The sensor recorded tri-axial acceleration at a sampling rate of 200 Hz and stored data locally on a microSD card.
- **Video-recording:** All sessions were video-recorded and synchronized with the IMU data to enable expert annotation of freezing of gait episodes.

Sensor Type	Location
IMU	Left waist (near center of mass)
Camera	Frontal view for annotation

Table 4.13. Sensor Placement Summary

Tasks: Participants performed a scripted protocol in their own homes, designed to elicit freezing of gait during realistic activities of daily living. Each participant completed two sessions of approximately 20 min: one in the OFF medication state (early morning before taking dopaminergic medication) and one in the ON medication state (at least one hour after medication intake). The protocol consisted of the following task categories:

The dataset consists of synchronized inertial and annotation data:

Task name	Acronym	Description
Home Tour	HT	Guided walk through the participant's home environment
Stand Up and Go	SUG	Stand up from a chair, walk through a doorway, turn, and return to the chair (FoG provocation task)
Outdoor Walk	OW	Short walk outside the home environment
Dual Task	DT	Walking while performing a concurrent cognitive or motor task (e.g., reading or carrying an object)
False Positive Protocol	FP	Rapid repetitive upper-limb movements (e.g., brushing teeth, drawing) to test false detections (ON state only)

Table 4.14. List of pre-scripted tasks performed by participants.

- **IMU signals:** Tri-axial acceleration from the waist-mounted sensor.
- **Annotations:** Frame-level labels indicating freezing of gait or non-freezing periods, obtained by expert review of synchronized video recordings.
- **Metadata:** Clinical and demographic information including Hoehn and Yahr stage, UPDRS-III scores (ON and OFF), FoG-Q, and MMSE.

Acceleration signals were originally sampled at 200 Hz and resampled to 40 Hz for analysis. Signals were low-pass filtered with a second-order Butterworth filter (cutoff frequency 15 Hz) and segmented into overlapping windows of 3.2 s (128 samples) with 50% overlap. FoG onset was defined as the arrest of effective stepping, and offset as the resumption of alternating gait, based on synchronized video review by experienced clinicians.

All data were preprocessed, quality-checked, and annotated within the REMPARK project framework. For the purposes of this study, we focused exclusively on the waist-mounted inertial measurement unit data, omitting video recordings and additional contextual labels. This choice was motivated by the goal of developing lightweight, unobtrusive, and home-deployable wearable monitoring systems.

The dataset contains a total of 93.03 min of annotated freezing of gait across 1,321 FoG episodes. Of these, 79.7% occurred during the OFF medication state, 13.3% during the ON state, and 7.0% in intermediate motor states.

4.2.8 6MWT

The 6MWT dataset [104] includes data from 15 individuals with Parkinson's disease (6 female, 9 male; mean age 67.5 $\pm$ 7.2 years) who performed a standardized Six-Minute Walk Test (6MWT). Participants had Hoehn and Yahr stages ranging from 2 to 3 and exhibited mild to moderate gait impairments. Inclusion criteria required a clinical diagnosis of Parkinson's disease and the ability to walk independently without assistive devices. Exclusion criteria included other neurological or orthopedic disorders affecting gait.

Sensors: Each participant was equipped with a synchronized wearable sensing system, described below.

- **Wearable inertial sensors:** Three inertial measurement units (IMUs) were placed on the lower back (L5), and on the lateral aspect of both ankles. Each IMU recorded tri-axial acceleration and tri-axial angular velocity at a sampling rate of 102.4 Hz.
- **Video-recording:** All trials were video-recorded to enable expert annotation of gait and freezing-related events. Sensor and video data streams were synchronized post-hoc.

Sensor Type	Location
IMU	Lower back (L5), bilateral ankles
Camera	Frontal and lateral views for annotation

Table 4.15. Sensor Placement Summary

Tasks: Participants performed a single Six-Minute Walk Test in an indoor corridor. Subjects were instructed to walk continuously back and forth along a straight 20 m walkway at a comfortable self-selected pace for six minutes, turning 180° at each end of the corridor. The test was designed to capture steady-state walking as well as fatigue-related gait changes and potential freezing episodes. All data were recorded continuously with synchronized IMU and video data and annotated frame-by-frame by expert raters.

The dataset consists of synchronized inertial and annotation data:

- **IMU signals:** Tri-axial acceleration and angular velocity from three body-worn sensors.

- **Annotations:** Frame-level labels indicating gait phases and freezing of gait (FoG) episodes.

- **Metadata:** Participant demographics and clinical scores including UPDRS-III and Hoehn and Yahr stage.

Signals were sampled at 102.4 Hz and segmented into overlapping windows of 3 s with 50% overlap. Freezing of gait events were identified from synchronized video recordings and defined as brief episodes of inability to initiate or continue effective stepping.
All data were preprocessed, quality-checked, and made publicly available by the authors. For the purposes of this study, we focused exclusively on the inertial measurement unit data, omitting video recordings and additional clinical assessments. This choice was made to ensure consistency with wearable-only monitoring approaches and to facilitate reproducible benchmarking.

The dataset contains approximately 90 min of total walking data, including 176 annotated freezing of gait episodes. Freezing events occur predominantly during turning phases at corridor endpoints and exhibit large inter-subject variability.

4.2.9 tDCS

The tDCS dataset [105] is a subset of the public dataset released for the Parkinson's Disease Freezing of Gait Prediction competition hosted on Kaggle. This subset includes data from individuals with Parkinson's disease who participated in a transcranial direct current stimulation (tDCS) intervention study. Participants performed standardized walking tasks under different stimulation conditions (real and sham tDCS), with the aim of assessing the effect of non-invasive brain stimulation on gait and freezing of gait (FoG). Inclusion criteria required a clinical diagnosis of Parkinson's disease and the presence of gait impairment, while exclusion criteria included other neurological or musculoskeletal disorders affecting locomotion.

Sensors: Each participant was equipped with wearable inertial sensors, described below.

- **Wearable inertial sensors:** Inertial measurement units were placed on both ankles and on the lower back. Each IMU recorded tri-axial acceleration and tri-axial angular velocity at a sampling rate of 128 Hz.

- **Video-recording:** All trials were video-recorded to enable expert annotation of freezing of gait episodes. IMU and video data streams were synchronized post-hoc.

Sensor Type	Location
IMU	Lower back (L5), bilateral ankles
Camera	Frontal view for annotation

Table 4.16. Sensor Placement Summary

Tasks: Participants performed repeated walking trials along an indoor corridor under different stimulation conditions. Each session consisted of straight walking with frequent 180° turns at corridor endpoints to provoke freezing of gait. Trials were conducted before and after application of tDCS (real or sham), and all tasks were recorded continuously with synchronized IMU and video data and annotated for FoG events by expert raters.

Task name	Acronym	Description
Corridor Walk (pre-tDCS)	CW_{pre}	Straight walking with 180° turns before stimulation
Corridor Walk (post-tDCS)	CW_{post}	Straight walking with 180° turns after stimulation

Table 4.17. List of pre-scripted tasks performed by participants.

The dataset consists of synchronized inertial and annotation data:

- **IMU signals:** Tri-axial acceleration and angular velocity from three body-worn sensors.
- **Annotations:** Frame-level labels indicating freezing of gait (FoG) or non-freezing periods.
- **Metadata:** Subject identifiers, stimulation condition (real or sham tDCS), and trial identifiers.

Each time series is sampled at 128 Hz and stored in comma-separated value (`CSV`) files, one per trial. Each file contains the three-dimensional accelerometer and

gyroscope signals from each sensor location, together with a binary FoG label for each time sample. Missing values are represented as `NaN`.
All data were preprocessed, quality-checked, and made publicly available via the Kaggle competition repository. For the purposes of this study, we focused exclusively on the tDCS subset of the dataset and excluded all other subdatasets (e.g., DBS and medication-related recordings). Furthermore, we restricted the analysis to the inertial sensor data recorded from the lower back and ankle-mounted IMUs, omitting video recordings and non-wearable sensing modalities. This choice was made to ensure consistency across subjects and to facilitate reproducible benchmarking of wearable-based freezing of gait detection algorithms.

4.2.10 PERFORM

The PERFORM dataset [106] was collected within the European FP7 PERFORM project and comprises two complementary subsets: a short-term dataset (PERFORM-Short) recorded under controlled clinical conditions, and a long-term dataset (PERFORM-Long) acquired in real-life home environments. Both subsets include data from individuals with Parkinson's disease (PD) and were designed to support the automatic assessment of motor symptoms such as tremor, bradykinesia, levodopa-induced dyskinesia (LID), and freezing of gait (FoG).

PERFORM-Short includes 39 recordings from 24 PD patients acquired in a hospital environment. Recordings lasted approximately 15 minutes and were collected during structured protocols involving lying on a bed, rising to a seated position, standing up, walking, opening and closing doors, drinking, and performing spontaneous movements. PERFORM-Long includes long-term home recordings from 12 PD patients collected over five consecutive days, with two sessions per day (morning and afternoon) of approximately 4 hours each. Patients also completed symptom diaries every 30 minutes to provide coarse clinical labels for validation.

Sensors: Each participant was equipped with a wearable multi-sensor system composed of inertial sensors, described below.

- **Wearable inertial sensors:** Four tri-axial accelerometers were placed on the wrists and ankles to capture limb movements. One combined accelerometer and gyroscope unit was positioned at the waist to record trunk acceleration and angular velocity. All sensors were wirelessly connected to a portable data logger worn on the belt.

- **Video-recording (PERFORM-Short only):** Short-term sessions were video-recorded to enable expert annotation of motor symptoms.

Sensor Type	Location
Accelerometer	Bilateral wrists, bilateral ankles
IMU (accel + gyro)	Waist (trunk)
Camera (short-term only)	Frontal view for annotation

Table 4.18. Sensor Placement Summary

Tasks: The two PERFORM subsets differ in task structure.

PERFORM-Short followed a predefined clinical protocol consisting of lying on a bed, rising to a seated position, standing up from a chair, walking, opening and closing a door, drinking, and performing spontaneous movements. *PERFORM-Long* captured free-living activities at home during daily life without scripted tasks, including walking, standing, sitting, and other activities of daily living.

Subset	Acronym	Description
PERFORM-Short	PS	Structured in-clinic protocol including posture changes, walking, and object interaction
PERFORM-Long	PL	Unscripted home-based daily activities over multiple days

Table 4.19. List of task categories performed by participants.

The dataset consists of synchronized inertial and annotation data:

- **Inertial signals:** Tri-axial acceleration from five body-worn sensors and tri-axial angular velocity from the waist unit.
- **Annotations:** Motor symptom labels including tremor, bradykinesia, levodopa-induced dyskinesia, and freezing of gait. PERFORM-Short labels were obtained by expert neurologists using synchronized video, while PERFORM-Long labels were derived from patient diaries and clinician review.
- **Metadata:** Participant identifiers, recording session identifiers, and symptom severity levels.

Signals were segmented using sliding windows of 1–5 s depending on the symptom detection module, with overlaps ranging from 50% to 75%. PERFORM-Short was used primarily for the development and training of symptom recognition models, while PERFORM-Long served as an evaluation dataset for long-term symptom monitoring under real-world conditions.
All data were preprocessed, quality-checked, and stored within the PERFORM project infrastructure. For the purposes of this study, we focused exclusively on the wearable inertial sensor data and on freezing of gait-related labels, omitting auxiliary test devices (e.g., glove, microphone, and camera) and non-motor symptom assessments. This choice was made to ensure consistency with wearable-based monitoring approaches and to facilitate reproducible algorithm benchmarking.

The PERFORM dataset exhibits substantial variability between its two subsets. PERFORM-Short provides densely annotated, task-driven recordings suitable for supervised model training, whereas PERFORM-Long offers ecologically valid, long-duration data reflecting daily motor fluctuations

4.2.11 Daneault

The Daneault dataset [107] comprises wearable sensor data collected from 31 individuals with Parkinson's disease (PD) experiencing motor fluctuations. Participants were recruited at two clinical sites (Spaulding Rehabilitation Hospital, Boston, and Mount Sinai Hospital, New York). After data quality control, recordings from 28 subjects were retained for public release. Subjects had Hoehn and Yahr stages ranging from II to IV and were receiving levodopa therapy. Inclusion criteria required a clinical diagnosis of PD with self-reported motor fluctuations and dyskinesia, while exclusion criteria included other neurological disorders and the presence of deep brain stimulation (DBS).

Sensors: Each participant was equipped with a minimal set of wearable accelerometer-based devices.

- **Wearable inertial sensors:** Three accelerometer devices were used: a GeneActiv wristwatch worn on the most affected upper limb, a Pebble smartwatch worn on the least affected upper limb, and a smartphone worn at the waist in a belt pouch. All devices recorded tri-axial acceleration at a sampling rate of 50 Hz. No gyroscope data were collected in order to maximize battery life and enable long-term recordings.

Sensor Type	Location
Accelerometer	Most affected wrist (GeneActiv)
Accelerometer	Least affected wrist (Pebble smartwatch)
Accelerometer	Waist (smartphone)

Table 4.20. Sensor Placement Summary

Tasks: Data were collected over four consecutive days following a mixed laboratory and home protocol.

During Days 1 and 4, participants attended laboratory sessions in which they performed a predefined battery of motor tasks multiple times (6–8 repetitions every 30 minutes) across different medication states. Tasks included standing, straight walking, walking while counting, stair ascent and descent, walking through a narrow corridor, sit-to-stand transitions, finger-to-nose movements, alternating hand movements, typing, drawing, folding, drinking, organizing papers, and sitting. During each task, trained clinicians rated symptom severity or presence for tremor, bradykinesia, and dyskinesia using standardized clinical scales.

During Days 2 and 3, participants wore the sensors continuously in their home and community environments while performing unscripted daily activities. Subjects also completed symptom diaries, sleep logs, and medication intake records.

Subset	Acronym	Description
Laboratory protocol	Lab	Structured motor tasks assessed by clinicians
Home monitoring	Home	Unscripted daily-life activities with patient diaries

Table 4.21. List of task categories performed by participants.

The dataset consists of synchronized inertial and clinical annotation data:

- **Accelerometer signals:** Tri-axial acceleration from three body locations (both wrists and waist).
- **Clinical annotations:** Limb-specific symptom labels for tremor (0–4 severity), bradykinesia (presence/absence), and dyskinesia (presence/absence) obtained during laboratory tasks.
- **Metadata:** Demographic data, medication intake times, UPDRS scores, Hoehn and Yahr stage, and patient-reported symptom diaries.

All signals were resampled to a uniform sampling rate of 50 Hz and temporally aligned across devices using a synchronization procedure based on a reference movement event. Missing data segments were represented as `NaN`. Data were made publicly available through the Synapse repository.
For the purposes of this study, we focused exclusively on the wearable accelerometer data and on the laboratory-session annotations related to lower-limb motor symptoms. We excluded smartphone diary entries, survey data, and auxiliary metadata not directly related to movement analysis. This choice was made to ensure consistency with wearable-based symptom monitoring and to facilitate reproducible benchmarking.

The dataset provides both densely annotated laboratory recordings and long-duration real-world data. The laboratory subset offers high-quality ground truth labels suitable for supervised learning, while the home subset provides ecologically valid recordings reflecting daily motor fluctuations.

4.2.12 REMAP

The REMAP dataset [108] (REal-world Mobility Activities in Parkinson's disease) comprises multimodal recordings from 24 participants, including 12 individuals with Parkinson's disease (7 male, 5 female; mean age 61.25 $\pm$ 8.5 years) and 12 healthy control volunteers (3 male, 9 female; mean age 59.25 $\pm$ 13.4 years). Participants with PD had modified Hoehn and Yahr stages of 3 or less in the practically-defined "off" medication state and experienced mild to moderate motor impairment. Inclusion criteria for PD participants required a clinical diagnosis of idiopathic Parkinson's disease, while exclusion criteria included significant cognitive impairment, depression, and the use of walking aids indoors.

Sensors: Each participant was equipped with a synchronized multimodal sensing system, described below.

- **Wearable inertial sensors:** Two wrist-worn wearable devices recorded tri-axial acceleration continuously throughout the study. The accelerometers operated at a variable sampling rate of approximately 30 Hz and provided timestamped measurements along three spatial axes.

- **Video-recording:** Wall-mounted RGB cameras were installed in communal rooms of a fully furnished test-bed house, capturing participants' movements

in living room, dining room, kitchen, and hallway areas. Video was recorded at an average frame rate of 30 fps and a resolution of 640×480 pixels.

- **Skeleton pose data:** Two-dimensional (2D) and three-dimensional (3D) human pose skeletons were extracted from RGB video for sit-to-stand and turning episodes using OpenPose and HRNet-based models, respectively.

Sensor Type	Location
Accelerometer	Bilateral wrists
Camera	Wall-mounted (communal rooms)
Skeleton pose	Derived from RGB video

Table 4.22. Sensor Placement Summary

Tasks: Participants stayed for five consecutive days in a house configured as a home-like environment. Data were collected during both free-living (unscripted, unobserved) periods and structured clinical assessment sessions. Participants with PD were asked to withhold dopaminergic medication and, when applicable, deactivate deep brain stimulation (DBS) for a limited period to induce an "off" medication state. Clinical assessments included the MDS-UPDRS Part III, the Timed Up and Go (TUG) test, and repeated 20 m walking trials at normal, slow, and fast paces.

The dataset focuses on discrete mobility actions, primarily sit-to-stand (STS) transitions and turning in gait, together with a set of non-turning, non-STS activities (e.g., sitting, standing still, stair ascent and descent).

Task name	Acronym	Description
Sit-to-Stand	STS	Transition from seated to standing position
Turning in Gait	TURN	Turning episodes of 90°–360° during walking or standing
Non-STS/Non-Turn	NT	Activities such as sitting, standing, stair ascent/descent, and leaning

Table 4.23. List of task categories performed by participants.

The dataset consists of synchronized inertial, vision-based, and annotation data:

- **Accelerometer signals:** Tri-axial wrist accelerometry from two wearable devices.

- **Skeleton data:** Frame-level 2D and 3D joint coordinates for STS and turning episodes.

- **Annotations:** Human rater labels for each episode, including action type, duration, turning angle, number of turning steps, type of turn (pivot or step), and medication state ("on" or "off").

- **Metadata:** Participant group (PD or control), medication and DBS state, and clinical assessment indicators.

Signals were temporally aligned using timestamps and segmented into discrete action clips. Skeleton clips include additional frames before and after each action to allow for automatic detection and segmentation. Accelerometer clips were trimmed to contain only the corresponding action segment. Missing values are represented as `NaN`.

All data were preprocessed, quality-checked, and released in two forms: an open dataset containing coarsened skeleton pose data and cohort-level metadata, and a controlled-access dataset including full-resolution skeleton data and accelerometer signals. For the purposes of this study, we focused exclusively on the wearable accelerometer data and on sit-to-stand and turning annotations, omitting RGB video and raw skeleton pose data. This choice was made to ensure consistency with wearable-based mobility monitoring approaches and to facilitate reproducible benchmarking.

The REMAP dataset provides both free-living and clinically observed mobility data, enabling direct comparison between naturalistic and assessment-driven movement patterns. The dataset further captures motor performance under both "on" and "off" medication states, supporting the study of medication-related motor fluctuations.

4.2.13 HRMS

The HRMS dataset [109] includes wearable sensor data from 30 individuals with Parkinson's disease (20 male, 10 female; mean age 67.1 $\pm$ 10.2 years) with established motor fluctuations. Participants had a mean disease duration of 11.0 $\pm$ 5.1 years and Hoehn and Yahr stages ranging from 2 to 4. Inclusion criteria required a clinical diagnosis of Parkinson's disease and the presence of motor fluctuations, while exclusion criteria included severe cognitive impairment and inability to ambulate independently.

Sensors: Each participant was equipped with a single wearable inertial sensor, described below.

- **Wearable inertial sensor:** A wrist-worn Microsoft Band 2 device was placed on the most affected upper limb. The device recorded tri-axial acceleration with a sampling rate of approximately 62.5 Hz and a measurement range of ± 8 g. Gyroscope data were available but only accelerometer signals were used for motor state detection.

Sensor Type	Location
Accelerometer	Wrist (most affected side)

Table 4.24. Sensor Placement Summary

Tasks: Data were collected in a free-living, unscripted environment. Participants were continuously accompanied ("shadowed") by a movement disorder expert during daily activities in a hospital setting. Activities included sitting, standing, walking, lying, motor examination tasks, and other spontaneous movements. No predefined motor task protocol was imposed, allowing the capture of naturalistic movement patterns and motor fluctuations across the day.

The dataset consists of synchronized inertial and annotation data:

- **Accelerometer signals:** Tri-axial wrist acceleration.
- **Annotations:** Minute-level expert labels of motor state: OFF (bradykinetic), ON (normal motor function), and DYS (dyskinetic), based on MDS-UPDRS item 3.14 and AIMS item 5.
- **Metadata:** Demographic and clinical variables including UPDRS-III, AIMS, Hoehn and Yahr stage, and medication information.

Signals were filtered with a Butterworth band-pass filter (0.1–20 Hz), resampled to 60 Hz, and segmented into non-overlapping 1-minute windows (3600 samples per window). Data segments with negligible movement or missing annotations were discarded, resulting in a total of 8,661 minutes of usable data. Missing values are represented as `NaN`. Data augmentation using window shifting and rotational transformations increased the effective number of samples for model training.
All data were preprocessed, quality-checked, and curated by the original authors.

The HRMS dataset exhibits a moderately imbalanced class distribution, with approximately 27% OFF, 41% ON, and 32% DYS samples after augmentation. Motor state transitions occur frequently over short time scales, making this dataset particularly suitable for evaluating high-resolution motor state detection and temporal modeling approaches.

4.2.14 ADL

The ADL dataset [110] includes wearable inertial data collected from 72 individuals with Parkinson's disease (approximately 60% male; mean age 69.2 $\pm$ 10.2 years) during activities representative of daily living. Participants exhibited a wide range of motor impairment, with Hoehn and Yahr (H&Y) stages spanning mild to moderate/severe disease. Inclusion criteria required a clinical diagnosis of Parkinson's disease with motor symptoms and the ability to perform basic walking and turning tasks, while exclusion criteria included major comorbidities and cognitive or visual impairments preventing safe task execution.

Sensors: Each participant was equipped with a single wearable sensing device, described below.

- **Smartphone inertial sensors:** A Samsung S5 mini smartphone was secured at the lower back, approximately at the L3 vertebra, using an elastic belt. The embedded inertial sensors recorded tri-axial acceleration and tri-axial angular velocity at a sampling rate of 200 Hz.

Tasks: Data were collected during semi-supervised activities designed to reflect activities of daily living (ADL). Participants performed walking and mobility tasks that naturally occur during clinical visits or daily routines, including free walking, sit-to-stand and stand-to-sit transitions, standing, and turning. A particular focus was placed on 180° turns, which are known to be sensitive to postural instability and disease severity in Parkinson's disease.

Two participant subgroups were included:

- *G-PD1:* 59 patients recorded during routine outpatient clinical visits while performing UPDRS-related motor tasks.
- *G-PD2:* 13 patients recorded during a Six-Minute Walk Test (6MWT), walking back and forth along a 10 m corridor for six minutes.

Task name	Acronym	Description
Free Walking	FW	Unconstrained walking during clinical visit or ADL-like tasks
Turning	TURN	180° turns performed during walking or mobility tasks
Six-Minute Walk Test	6MWT	Continuous walking along a 10 m corridor for 6 min

Table 4.25. List of task categories performed by participants.

The dataset consists of synchronized inertial and annotation data:

- **Inertial signals:** Tri-axial acceleration and angular velocity from the smartphone IMU.
- **Annotations:** Manually annotated turning events, with precise onset and offset of 180° turns.
- **Clinical labels:** Hoehn and Yahr stage and UPDRS-Part III item scores, including gait and postural stability, available for the G-PD1 subgroup.
- **Metadata:** Demographic and clinical information for each participant.

Signals were reoriented to compensate for sensor placement variability, detrended, and filtered using a zero-phase second-order Butterworth low-pass filter with a cutoff frequency of 20 Hz. Only 180° turning segments were retained for analysis in order to reduce confounding factors and ensure consistency across participants. Missing values were not present in the segmented turn data.
All data were preprocessed and curated by the original authors. For the purposes of this study, we focused exclusively on the smartphone inertial data and on turning events extracted during activities of daily living, omitting straight walking segments, video recordings, and non-mobility-related clinical assessments. This choice was made to ensure consistency with unobtrusive, single-device monitoring approaches suitable for long-term ADL-based assessment.

The ADL dataset provides a large collection of turning events (339 total 180° turns) across a heterogeneous cohort of Parkinson's disease patients. Turning quality exhibits substantial inter-subject variability and correlates strongly with clinical measures of postural stability and disease severity, making this dataset particularly suitable for studying mobility impairment during daily living.

4.2.15 adPMD

The adPMD dataset [111] comprises wearable accelerometer data collected from individuals with advanced Parkinson's disease (PD) experiencing motor fluctuations. The dataset originates from the multicenter VALIDATE-PD study and includes data from 63 participants (46% male; median age 66 years, IQR 60–73) who provided sufficient high-quality recordings across two consecutive days. Participants exhibited advanced disease severity, with a median Hoehn and Yahr stage of 4 (IQR 3–5) and prominent motor complications including wearing-off and dyskinesia. Inclusion criteria required a clinical diagnosis of PD with documented motor fluctuations, while exclusion criteria included dementia (MoCA < 21), psychotic symptoms, and ongoing advanced therapies such as deep brain stimulation or infusion-based treatments. Written informed consent was obtained under Institutional Review Board–approved protocols at all participating centers in Germany and Sweden.

Sensors: Each participant was equipped with a single wearable inertial device, described below.

- **Wearable accelerometer:** A Parkinson's KinetiGraph (PKG®) device was worn on the wrist of the most affected upper limb. The device recorded continuous tri-axial acceleration using a proprietary sampling scheme that produces activity-related features every 2 min. Raw accelerometer data were internally processed to derive quantitative bradykinesia (BKS) and dyskinesia (DKS) scores.

Tasks: Data were collected during unscripted daily activities in a real-world clinical environment. Participants were continuously monitored between 08:00 and 18:00 over two consecutive days while performing habitual activities such as sitting, standing, walking, and clinical assessment maneuvers. Every 30 minutes, participants underwent a brief clinical examination including a modified 7 m Timed Up and Go Test (7m-TUGT), and motor state was independently rated by experienced movement disorder specialists.

Simultaneously, participants and clinicians completed Parkinson's disease home diary ratings, providing half-hour–level labels of motor state (Asleep, Off, On without dyskinesia, On with dyskinesia). These diary labels served as reference annotations for validation of the wearable-derived digital Parkinson's Motor Diary (adPMD).

The dataset consists of synchronized wearable and annotation data:

Motor state	Acronym	Description
Off state	OFF	Predominant bradykinetic motor state
On state	ON	Normal motor function without dyskinesia
Dyskinetic state	DYS	On state with involuntary dyskinetic movements

Table 4.26. Motor state categories in the adPMD dataset.

- **Accelerometer-derived features:** Quantitative bradykinesia (BKS) and dyskinesia (DKS) scores computed from continuous wrist accelerometry.
- **Motor diary labels:** Half-hour–resolution motor state labels provided by clinical observers and participants.
- **Clinical metadata:** Demographics, disease duration, MDS-UPDRS scores, Hoehn and Yahr stage, medication regimen, and motor complication history.

Wearable data were aggregated into non-overlapping 30 min windows to match the temporal resolution of the motor diary annotations. Each window was assigned a categorical motor state label based on thresholding of median BKS and DKS values. Multiple thresholding strategies were explored, including standard, calibrated, and individualized cut-offs. Periods with insufficient sensor wear or missing annotations were excluded. Missing values were represented as `NaN`.
All data were preprocessed and quality-controlled by the original authors. For the purposes of this study, we focused exclusively on the wrist-worn accelerometer–derived motor state labels (OFF, ON, DYS) and excluded sleep periods, raw clinical examination scores, and patient-reported diary data. This choice was made to ensure consistency with passive, single-sensor monitoring approaches and to enable reproducible benchmarking of motor state classification.

The adPMD dataset exhibits a pronounced class imbalance and substantial inter-subject variability in motor state transitions. While daily time spent in Off and Dyskinetic states shows moderate agreement with clinical ratings, temporal agreement at the half-hour level is limited, reflecting the challenges of high-resolution motor state detection in uncontrolled real-world environments.

4.2.16 Mobilise-D

The Mobilise-D dataset [112] was collected within the framework of the Mobilise-D project and comprises large-scale, multicentric recordings aimed at validating digital

mobility outcomes (DMOs) derived from wearable inertial sensors. The dataset includes participants from six cohorts: healthy older adults and individuals with Parkinson's disease (PD), chronic obstructive pulmonary disease, multiple sclerosis, proximal femoral fracture, and congestive heart failure.

For the purposes of this study, we focus exclusively on the Parkinson's disease cohort. Participants with PD were adults diagnosed according to Movement Disorder Society criteria and exhibited impaired mobility without severe cognitive deficits. Inclusion criteria required the ability to walk at least 4 m independently, with or without walking aids. Exclusion criteria included comorbid conditions significantly affecting mobility. Written informed consent was obtained under ethics approvals from all participating clinical sites across Europe.

Sensors: Each participant was equipped with a validated wearable inertial sensing system, described below.

- **Wearable inertial sensor:** A DynaPort MM+ (McRoberts BV, The Netherlands) inertial measurement unit was worn on the lower back using an elastic belt. The device recorded tri-axial acceleration and tri-axial angular velocity at a sampling rate of 100 Hz, with sensor ranges of ± 8 g (accelerometer) and ± 2000 dps (gyroscope).
- **Reference systems (validation phases only):** During laboratory-based assessments, additional reference systems were used, including stereophotogrammetry, pressure insoles, distance sensors, and foot-mounted IMUs (INDIP system). These reference modalities were used exclusively for validation and are not required for downstream analysis.

Sensor Type	Location
IMU (DynaPort MM+)	Lower back (L5, trunk)
IMU / insoles (reference)	Feet and lower back (lab only)
Camera / SP system	Laboratory reference only

Table 4.27. Sensor Placement Summary

Tasks: The Mobilise-D protocol includes both structured laboratory assessments and real-world mobility monitoring, enabling validation across contexts.

Laboratory-based assessments consisted of standardized mobility tasks, including straight walking at preferred, slow, and fast speeds, Timed Up and Go (TUG), L-Test, surface walking with turns, hallway walking with step negotiation, and simulated activities of daily living (e.g., sit-to-stand, object handling, room-to-room walking).

Real-world assessments comprised two phases: (i) a 2.5-hour semi-supervised observation period in habitual environments (home, work, or community), and (ii) a continuous 7-day unsupervised free-living monitoring period. During real-world recordings, participants performed daily activities without task constraints.

Context	**Acronym**	**Description**
Lab-based assessment	LAB	Structured walking and mobility tasks under controlled conditions
Real-world observation	RW-2.5h	Semi-supervised habitual activity monitoring
Free-living monitoring	RW-7d	Continuous unsupervised daily-life monitoring

Table 4.28. List of task categories performed by participants.

The dataset consists of synchronized inertial and annotation data:

- **IMU signals:** Tri-axial acceleration and angular velocity from the lower-back–mounted DynaPort MM+ device.
- **Digital Mobility Outcomes (DMOs):** Walking bout–level measures including walking speed, cadence, stride and step duration, stride length, gait variability, turning frequency and angle, and bout duration.
- **Contextual labels:** Indoor/outdoor classification, walking aid usage, and surface inclination (available for real-world recordings).
- **Clinical metadata:** Demographics, disease-specific clinical scores (e.g., MDS-UPDRS Part III), and cohort identifiers.

IMU signals were segmented into walking bouts, defined as sequences containing at least two consecutive strides. DMOs were computed at the walking-bout level rather than at fixed time windows. Missing data due to non-wear or signal loss were handled using a complete-case approach within each assessment phase.
All data were preprocessed and quality-controlled following a standardized multicenter protocol. At the time of publication, the Mobilise-D dataset is released in stages,

with public access to selected cohorts and assessment phases. For the purposes of this study, we focused exclusively on the lower-back IMU data and on gait- and turning-related digital mobility outcomes, omitting reference systems and non-PD cohorts. This choice was made to ensure consistency with single-sensor wearable monitoring and to facilitate reproducible analysis across datasets.

The Mobilise-D dataset provides an unprecedented combination of controlled laboratory data and ecologically valid real-world mobility recordings. While laboratory data offer high-fidelity reference measurements, real-world recordings capture substantial variability in walking behavior and context. As with other free-living datasets, the distribution of walking bouts is highly imbalanced, with short bouts dominating daily-life recordings.

The Mobilise-D dataset described in this thesis refers specifically to the Technical Validation Study dataset publicly released through Zenodo. The larger Clinical Validation Study dataset, which includes a substantially broader participant cohort, was not publicly available at the time of this analysis.

4.3 Cross-Dataset Analysis

Section 2.2 presented details of the individual datasets currently available for wearable-based research in Parkinson's disease. Building upon those descriptions, this subsection adopts a comparative perspective, examining how these resources differ and where common patterns emerge.

The analysis is organized around three main aspects. First, demographic and clinical characteristics of the cohorts are summarized. Second, attention is given to the motor symptoms primarily addressed by each dataset, distinguishing studies centered on freezing of gait from those targeting broader manifestations of motor impairment. Finally, the sensing technologies, body locations, and acquisition strategies are reviewed to highlight methodological variability across investigations.

4.3.1 Demographic and Clinical Information

Tables 4.29 and 4.30 summarize participants' demographic and clinical information in datasets focused on FoG or other symptoms, respectively. The sample size ranges from 7 (i.e., [100]) to 71 (i.e., [105]) subjects, with only three datasets comprising more than 50 patients (i.e., [97, 105, 110]). The small sample size limits the generalizability of the results to the entire PD population due to the high inter- and intra-subject

variability in symptom manifestations and movement patterns.

The demographic data indicate that most datasets involve middle-aged to elderly participants (mean age between 60 and 70), consistent with the typical onset and progression of the disease. Gender balance varies significantly, with some datasets under-representing female participants. On the one hand, this often reflects the unbalanced prevalence of PD in males and females. On the other hand, this can introduce bias and severely impact the generalizability of findings. Additionally, the reported years since diagnosis (YoD) span a wide range, highlighting the heterogeneity in disease progression among participants.

The performed tasks range from simple walking exercises to dual-task challenges and simulated daily activities. Gait tasks are commonly used in FoG datasets (Table 4.29), typically consisting of supervised walking (e.g., straight line, zigzag, random) and turning (e.g., figure of eight, 180-degree, 360-degree) activities, with or without stops, and with or without cognitive (e.g., counting backwards) or motor (e.g., negotiating obstacles, passage through doorways, carrying glass full of water) dual-tasks. Simulated activities of daily living (ADLs) are poorly represented, despite their importance for developing recognition systems that can work in real-life scenarios. Datasets focusing on other symptoms (Table 4.30) include additional activities such as standing, postural transitions, and balance tasks. Most importantly, data are also recorded during supervised or unsupervised free-living activities.

4.3.2 Target Symptoms

When examining PD datasets comprising wearable sensor data, they can mainly be divided into two main categories: (1) studies dedicated explicitly to FoG and (2) studies that encompass a broader range of motor symptoms, including, but not limited to, FoG. This distinction is vital, as FoG-targeted datasets often involve specific tasks and sensor placements tailored to capture this particular phenomenon. In contrast, broader datasets aim to provide insights into other motor impairments such as tremors, bradykinesia, and dyskinesia.

Tables 4.31 and 4.32 describe datasets specifically related to FoG or other symptoms, respectively, reporting information concerning target symptoms, total data collection time, medication status, activity labels, and clinical tests.

Table 4.31 provides valuable insights into the frequency, duration, and characteristics of FoG events. The datasets highlight considerable variability in the number of FoG events registered and their characteristics. For example, the duration of FoG

Dataset	# Patients (M/F)	Age	YoD	MoCA	MMSE	UPDRS	FoG-Q	H&Y	Tasks Performed
Daphnet [101]	10 (7/3)	66.5±4.8	13.7±9.67	n/a	n/a	n/a	n/a	2.6±0.65	1. Straight-line walking 2. Walking with pauses 3. Turning 4. Gait freezing events 5. Short resting intervals
IMU [102]	35 (19/16)	62.5±10.59	8.23±3.99	n/a	n/a	2: 8.54±3.78 3: 31.83±14.99	17.72±5.63	2.91±0.5	1. Walking 2. 360-degree turns 3. Resting periods 4. Straight-line and zigzag walking 5. Obstacle avoidance
Multimodal [98]	12 (6/6)	69.1±7.9	9.3±6.8	23.6±3.6	28.2±1.5	1: 10.4 ± 5.5 2: 16.3 ± 10.6 3: 45.0 ± 16.0 4: 2.2 ± 2.9	16.2±4.2	n/a	1. Walking with sensors 2. Performing figure eights 3. Walking at variable speeds 4. Ascending and descending stairs
Cupid Multimodal [99]	18 (n/a)	68.9±10.2	8.8±4.6	n/a	n/a	3: 41.7 ± 10.2	19.7 ± 7.1	2 to 4	1. Walking tasks 2. Obstacle avoidance 3. Dual-task activities 4. Gait freezing simulation
REMPARK [103]	21 (18/3)	69.29±9.72	10.3±3.96	n/a	27.75±1.87	3: 36.3± 14.02 (OFF) 3: 16.15± 9.4 (ON)	15.8±4.11	3.07±0.43	1. Simulated ADLs 2. Free-living activities 3. Walking during medication ON/OFF states 4. Task-specific monitoring
6MWT [113]	38 (28/10)	70±5.1	12±6	n/a	27±2	n/a	n/a	3	1. Six-minute walking test 2. Timed distance walking 3. Recording heart rate and stride length
FOGSTAR [94]	20 (12/8)	65±4	10±4.2	n/a	n/a	n/a	n/a	2.5	1. Virtual reality walking 2. Interactive gait tasks 3. Cognitive task challenges during walking
tDCS [105]	71 (57/14)	63±3.5	8±3.8	24±2.5	28±1	3:	18±6	2.7	1. Walking 2. Dual-task cognitive challenges 3. Continuous walking for fatigue assessment
Oday [100]	7 (4/3)	68±6	14±7.1	23±3	27±2	..	n/a	3.1	1. Gait analysis 2. Walking during ADLs 3. Step count and stride analysis 4. Motion symmetry testing
PERFORM [106]	24 (n/a)	69±7	15±8	22±4	26±3	..	n/a	2.9	1. Simulated daily living tasks 2. Turning 3. Step-to-step coordination 4. Gait rhythm analysis
wearGait [97]	61 (36/25)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	1. Motion analysis 2. Walking across flat surfaces 3. Walking in a crowded environment 4. Navigating narrow spaces

YoD: Years of Diagnosis; MoCA: Montreal Cognitive Assessment; MMSE: Mini-Mental State Examination; H&Y: Hoehn & Yahr Scale

Table 4.29. Demographic and Clinical information for datasets focusing on FoG.

Dataset	# Patients (M/F)	Age	YoD	MoCA	MMSE	UPDRS	FoG-Q	H&Y	Tasks Performed
Daneault [107]	28 (n/a)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	1. Standing 2. Straight-line walking 3. Balance tasks 4. Sitting-to-standing transitions 5. Gait monitoring during rest
REMAP [108]	12 (7/5)	61.25±8.5	8.2±6.5	n/a	n/a	n/a	n/a	2.3±0.8	1. Free-living monitoring 2. Walking tasks 3. Resting periods 4. Step length and speed measurement
HRMS [109]	30 (20/10)	67.1±10.2	11.0±5.1	25.7±2.8	n/a	3 (ON): 21.6±15.3	n/a	n/a	1. Free-living monitoring 2. Short walking bouts 3. Resting 4. Monitoring walking patterns during transitions 5. Gait analysis using wearable sensors
ADL [110]	59 (37/22)	69.2±10.2	6.7±5.3	n/a	n/a	n/a	n/a	2.14±0.8	1. ADL simulation 2. Gait and balance analysis 3. Postural control tests 4. Navigating everyday obstacles 5. Sit-to-stand movements
PERFORM [106]	12 (n/a)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	1. Extended free-living monitoring 2. Detailed walking analysis 3. Step frequency tracking 4. Task-related movement assessments 5. Long-term variability studies
adPMD [111]	15 (10/5)	64±4.5	9±4	26±3	29±1	n/a	n/a	2.8	1. Free-living monitoring 2. Sensor-based gait analysis 3. Step length and symmetry evaluation 4. Walking under variable conditions 5. Short bursts of high-intensity walking
Mobilise-D [112]	20 (16/4)	69.8 ± 7.1	n/a	24.6 ± 4.0	n/a	3: 28.4 ± 13.6	n/a	2.0 ± 0.7	1. Gait analysis in daily life 2. Walking with variable surfaces 3. Step consistency monitoring 4. Transitioning between walking speeds 5. Adaptive gait tasks

YoD: Years of Diagnosis; MoCA: Montreal Cognitive Assessment; MMSE: Mini-Mental State Examination; H&Y: Hoehn & Yahr Scale

Table 4.30. Demographic and Clinical information for datasets focused on motor symptoms in general.

events ranges widely across datasets, reflecting this phenomenon's complexity and individual variability. Additionally, some datasets provide rich contextual information, such as medication status and daily activity labels (e.g., CupidMultimodal [99]), while others (e.g., PERFORM-Long [106]) focus solely on raw data. This variability underlines the importance of careful dataset selection when designing studies or developing models for FoG analysis.

Table 4.32 summarizes datasets focusing on motor symptoms in general. These datasets aim to capture motor-related impairments, offering a broader perspective on PD symptomatology. However, many datasets lack sufficient annotation for specific symptoms, limiting their applicability for detailed analyses. Moreover, medication status and daily activity labels are inconsistently reported, which may hinder the generalizability of results across different clinical settings.

4.3.3 Sensors Type and on Body Position

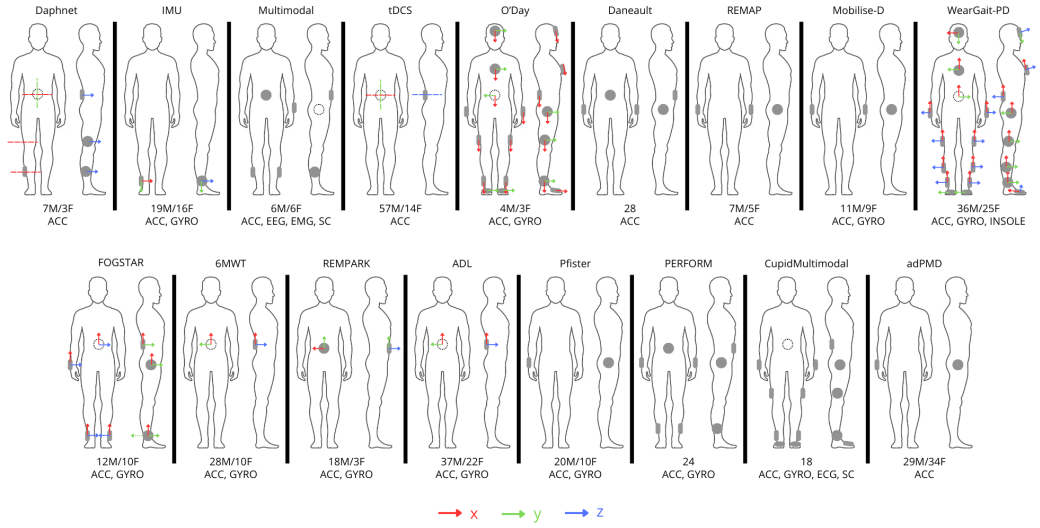


Figure 4.1. Sensor type, position, orientation in previously presented PD datasets. (→ (red) for the x-axis, → (green) for the y-axis, and → (blue) for the z-axis).

Figures 4.1 illustrates the sensor placements, orientation, and dataset characteristics for the publicly available and non-publicly available datasets, respectively.

The accelerometer is the most commonly used sensor type (100% of cases), followed by the gyroscope (65% of cases). Skin conductance sensors were used in two datasets, in combination with electroencephalography (EEG), electromyography (EMG), or electrocardiography (ECG) sensors. Finally, one dataset included smart pressure insoles. Overall, most datasets (82%) exclusively provided motion signals through inertial sensors. At the same time, multi-modal physical/physiological data

were available in three datasets (i.e., Multimodal [98], WearGait-PD [97], Cupid Multimodal [99]), all focused on FoG. While sensor positioning is described in all studies, the information on sensor orientation is totally (47% cases) or partially (11% cases) missing.

Figure 4.2 illustrates the distribution of sensor positions across datasets. The wrist and lower back are the most common locations, reflecting their clinical relevance for monitoring tremors, bradykinesia, and postural stability. The preference for wrist-mounted sensors aligns with their usability and convenience [100], while lower-back sensors effectively capture gait dynamics and falls [114].

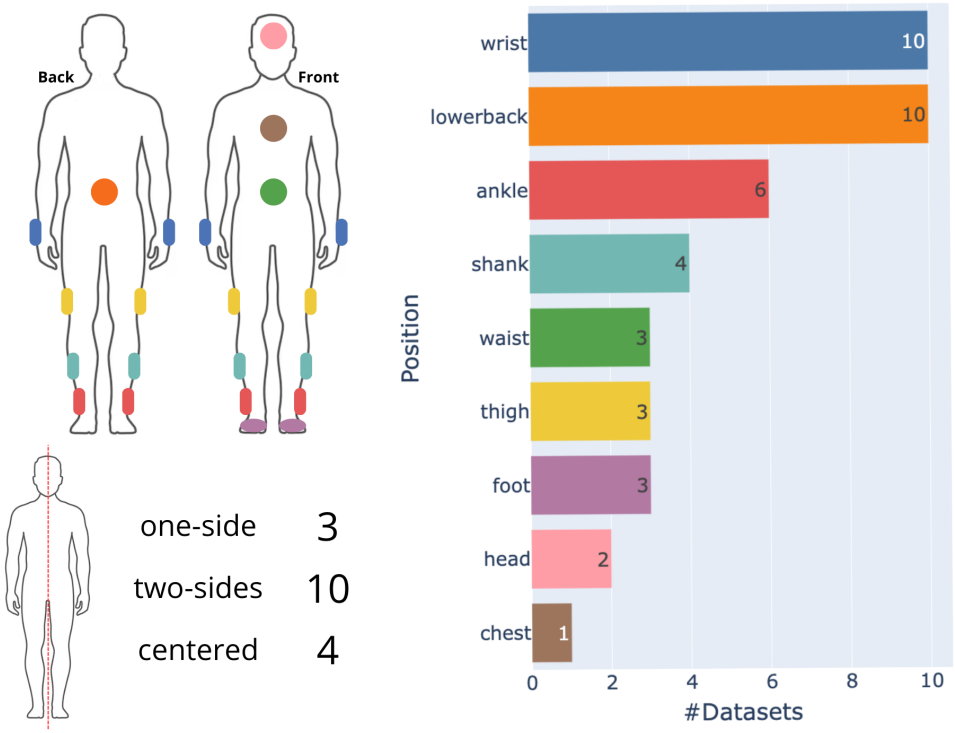


Figure 4.2. Distribution of sensor placements across datasets.

However, there is an under-representation of head-mounted sensors, which could provide valuable insights into head tremors and balance-related symptoms. Additionally, sensors on the chest and thigh are less frequently used despite their potential for monitoring vital signs (e.g., heart rate, respiration) and leg movements, respectively.

Figure 4.2 also highlights the distribution of sensors by body side. A few datasets adopt a one-sided sensor placement, which may lead to asymmetrical data and reduced accuracy for detecting bilateral symptoms. Two-sided placements are more common, offering improved balance and symmetrical monitoring. However, this configuration

is less practical for real-world use. Central placements, such as on the chest or lower back, provide robust data for gait and posture analysis but miss PD lateral asymmetries and specific movements of the limbs.

4.4 Limitations of Current Datasets

The comparative analysis presented in Subsection 2.3 highlights the remarkable progress achieved in the acquisition of wearable data for Parkinson's disease (PD). At the same time, it reveals persistent methodological constraints that limit reproducibility, cross-study comparability, and large-scale clinical translation.

A first critical issue concerns heterogeneity in sensing configurations. Although inertial measurement units are nearly ubiquitous, their number, placement, and orientation vary substantially across datasets. Wrist and lower-back positions dominate because of practicality and patient comfort, yet this prevalence leaves other potentially informative regions underexplored. As a result, specific manifestations such as head movements, trunk dynamics, or physiological correlates of stress and fatigue remain only partially captured. Furthermore, incomplete reporting of sensor orientation and calibration procedures complicates the replication of signal processing pipelines.

Another major limitation lies in the scope of targeted symptoms. Many datasets are designed with a strong focus on freezing of gait, often through protocols aimed at provoking episodes under controlled circumstances. While this strategy enables detailed annotation, it may reduce ecological validity and fails to represent the broader spectrum of motor and non-motor fluctuations experienced in daily life. Conversely, datasets emphasizing long-term monitoring frequently lack dense or event-level labeling, restricting their usefulness for supervised learning approaches.

Cohort characteristics introduce additional challenges. Sample sizes are often modest, and inclusion criteria may target narrow clinical phenotypes, such as patients with frequent FoG. Although appropriate for exploratory investigations, this restricts generalization to the heterogeneous PD population. Variability in demographic descriptors, disease duration, and availability of standardized clinical scales further complicates dataset integration and meta-analysis.

The analysis also shows that acquisition protocols tend to privilege structured tasks over realistic activities of daily living. Carefully scripted trajectories, repeated turns, and obstacle negotiations simplify annotation and experimental control, but they only approximate the environmental complexity encountered outside clinical settings.

Dataset	# Patients (M/F)	Dataset Duration (FoG seconds)	# FoG Events	FoG Stats [Min, Max, Mean, Std]	Medication Status	Activity Labels	PD Tests
Daphnet [101]	10 (7/3)	30000	273	[0.5, 40.5, 7.3, 6.7]	on-off	no	n/a
IMU [102]	35 (19/16)	n/a (1611)	n/a	n/a	on	no	UPDRS, NFOG-Q, MMSE, MOCA, FAB, HADS
Multimodal [98]	12 (6/6)	13320 (5280)	334	[1, 201, 8.2, n/a]	off	no	UPDRS, FOG-Q, MMSE, MOCA
CupidMultimodal [99]	18 (n/a)	n/a	184	[0.12, 98.88, 8.84, 14.87]	on	yes	UPDRS, FOG-Q
REMPARK [103]	21 (18/3)	n/a (5580)	1321	[3.2, 38, 5.4, n/a]	on-off	no	UPDRS, FOG-Q, MMSE
6MWT [113]	38 (28/10)	13140 (240)	33	[0, 21, n/a, n/a]	on	yes	UPDRS
FOGSTAR [94]	22 (12/10)	6474 (1074)	101	[0.6, 108, 9.2, 16.9]	off	yes	UPDRS, H&Y, FOG-Q, MOCA, FES-I, PDQ-8
tDCS [105]	71 (57/14)	55140 (17100)	100	[0.18, 687.62, 15.12, 51.90]	on-off	no	UPDRS, NFOG-Q
Oday [100]	7 (4/3)	5328 (1272)	211	[n/a, n/a, 7.8, n/a]	off	no	n/a
PERFORM-Short [106]	24 (n/a)	29500	824	[n/a, n/a, n/a, n/a]	on-off	no	UPDRS
PERFORM-Long [106]	24 (n/a)	1587994	13042	[n/a, n/a, n/a, n/a]	on-off	no	UPDRS
wear-GaitPD [97]	61 (36/25)	n/a	41	[n/a, n/a, n/a, n/a]	on-off	yes	UPDRS

NFOG-Q: New Freezing of Gait Questionnaire; MMSE: Mini-Mental State Exam; MOCA: Montreal Cognitive Assessment; FAB: Frontal Assessment Battery; HADS: Hospital Anxiety and Depression Scale; H&Y: Hoehn and Yahr; FES-I: Fall Efficacy Scale-International; PDQ-8: 8-item Parkinson's Disease Questionnaire.

Table 4.31. Summary of datasets focusing on FoG.

Dataset	Motor Symptoms	Dataset Duration	Symptoms Duration	# Symptom Events	Medication Status	Activity Labels	PD Tests
Daneault [107]	tremor, brady, dys	n/a	n/a	n/a	on-off	yes	UPDRS
REMAP [108]	n/a	n/a	n/a	n/a	on-off	yes	UPDRS, PIGD, NMSS, PDSS, PDQ, TUG, RSBDQ
HRMS* [109]	dys, brady	519660	n/a	n/a	on-off	no	UPDRS, MOCA
ADL [110]	n/a	n/a	n/a	n/a	on	yes	UPDRS
PERFORM-Short [106]	tremor, brady, LID, FoG	119518	n/a	n/a	on-off	no	UPDRS
PERFORM-Long [106]	tremor, brady, LID, FoG	6404861	n/a	n/a	on-off	no	UPDRS
adPMD [111]	n/a	43200	n/a	n/a	on-off	yes	UPDRS
Mobilise-D [112]	n/a	n/a	n/a	n/a	on-off	no	UPDRS

dys: dyskinesia; brady: bradykinesia; LID: Levodopa-induced dyskinesia; UPDRS: Unified Parkinson's Disease Rating Scale; PIGD: Postural instability and gait disorders; NMSS: Non-Motor Symptoms Scale; PDSS: Parkinson's Sleep Scale; PDQ: Parkinson's Disease Questionnaire; TUG: Timed Up-and-Go; RSBDQ: REM Sleep Behaviour Disorder Questionnaire; MOCA: Montreal Cognitive Assessment

Table 4.32. Overview of datasets related to motor symptoms in PD that do not specifically target FoG.

As a consequence, algorithms trained exclusively on these data may experience performance degradation when deployed in naturalistic contexts.

Finally, fragmentation in data formats, annotation schemas, and metadata reporting reduces interoperability. Even when datasets are publicly available, substantial harmonization efforts are required before cross-dataset experimentation becomes feasible. This situation slows methodological progress and limits opportunities for benchmarking.

Taken together, these elements indicate that, despite the richness of existing initiatives, important gaps remain in the balance between experimental control, ecological realism, annotation reliability, and scalability. Addressing these aspects is essential for advancing wearable-based assessment from proof-of-concept studies toward robust and clinically meaningful solutions.

These considerations naturally outline several directions for future dataset development. Greater consistency in sensing configurations, improved interoperability across resources, richer contextual information, and experimental paradigms capable of balancing symptom elicitation with ecological validity appear particularly important. Some of these aspects are already being progressively incorporated in recent initiatives, reflecting a shift toward more reproducible and deployment-oriented data collection strategies.

The following section introduces the FOGSTAR dataset, which was conceived with some of these challenges in mind.

4.5 The FOGSTAR Dataset

The analysis of the current dataset landscape reveals important trade-offs between experimental control, ecological realism, sensing portability, and annotation reliability. While existing resources have enabled significant methodological advances, the diversity of acquisition strategies and labeling schemes continues to limit interoperability and systematic evaluation across studies. In light of these considerations, the FOGSTAR dataset was developed to provide a coherent experimental framework in which wearable sensing, controlled elicitation of freezing phenomena, and rigorous temporal alignment are jointly addressed. The following sections describe the methodological design of the acquisition protocol and the characteristics of the resulting data resource.

4.5.1 Methods

4.5.1.1 Experimental design.

Participants. Individuals with idiopathic PD, diagnosed according to the international Movement Disorder Society (MDS) clinical diagnostic criteria, were recruited from two Italian university movement disorders clinics: the Department of Neurosciences and Mental Health, University Hospital Trust of Torino (Turin), and the Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona (Verona). The inclusion criteria are summarized in Table 4.33, while exclusion criteria included dementia and musculoskeletal, cardiovascular, psychiatric, or other neurological conditions that could substantially affect gait.

Criterion	Description
Diagnosis	Idiopathic PD according to the MDS clinical diagnostic criteria [**?**]
Disease stage	H&Y score between 2 and 4
FoG	Daily FoG episodes in the past month, with NFoG-Q score of 1 on Question 1 and ≥2 on Question 2 [38]
Treatment stability	Stable PD medication for more than one month prior to enrollment
Walking ability	Able to walk continuously and independently for at least 15 minutes (use of cane or rollator permitted)

Table 4.33. Inclusion criteria for participant recruitment. MDS: Movement Disorder Society; H&Y: Hoehn and Yahr stage; NFoG-Q: New FoG Questionnaire.

The study enrolled 22 people with PD, whose demographic and clinical characteristics are summarized in Table 4.34. To minimize medication-related effects and increase the probability of observing FoG, participants were instructed not to take their last dose of levodopa the day before testing, ensuring evaluation after a 12-hour drug washout. Clinical evaluations were performed in the morning and were followed immediately by sensor-based data collection. All participants provided their written informed consent prior to enrollment. The study was carried out following the ethical principles outlined in the Declaration of Helsinki.

Age (years)	Gender	Disease duration (years)	H&Y	UPDRS-III	FOG-Q	MOCA	FES-I	PDQ-8
70.9 ± 9.3	13M, 9F	10.1 ± 6.8	2.7 ± 0.8	38.9 ± 14.6	18.3 ± 5.4	21.3 ± 3.9	36.9 ± 26.9	14.4 ± 6.9

Table 4.34. Variables for demographic and clinical-related data. H&Y: Hoehn and Yahr stage; UPDRS: unified Parkinson's disease rating scale; FOG-Q: FoG questionnaire; MOCA: Montral cognitive assessment; FES-I: Fall efficacy scale international; PDQ-8: Parkinson's disease questionnaire.

Wireless Body Area Network (WBAN). Participants were instrumented with a wireless body area network (WBAN) [115] consisting of four IMUs (Nordic Thingy:52, Nordic Semiconductor). Each device is compact (6×6 cm, 47 g) and integrates a 9-axis motion-sensing module (3-axis accelerometer, 3-axis gyroscope, and compass) along with Bluetooth Low Energy (BLE) communication and an internal power supply. Although the hardware includes additional environmental sensors (microphone, speaker, humidity, temperature, pressure, gas, color, and light), only the accelerometer and gyroscope signals were used in this study. Although the Nordic Thingy:52 is primarily a prototyping platform, the embedded inertial sensor is the MPU-9250 (TDK InvenSense), a widely adopted 9-axis IMU used in research-grade and commercial motion-analysis systems; in our recordings, we did not observe abnormal drift or excessive noise beyond expected behavior. Technical specifications are summarized in Table 4.35.

IMU placement is illustrated in Figure 4.3: two units were positioned on the outer lower legs just above the ankles (Fig. 4.3c), one on the lower back at the L3–L5 level (Fig. 4.3d), and one on the wrist of the most affected side (Fig. 4.3b). This configuration captures both lower-limb and trunk dynamics and preserves an upper-limb reference. Data were streamed via BLE to a smartphone (Realme Pro 7) that served as gateway; the same app simultaneously recorded video, enabling annotation with a shared time base. Participants expressed high satisfaction with the wearable technology, as reflected by a Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST) [116] device score of 38.2 ± 1.9 (range: 34–40; maximum 40). Each of eight device-related aspects (comfort, weight, durability, adjustability, ease of use, dimensions, effectiveness, safety) was rated on a 0–5 scale.

Feature	Specification
Device model	Nordic Thingy:52
Dimensions	6×6 cm
Weight	47 g
Motion sensors	MPU-9250 (TDK InvenSense): 3-axis accelerometer, 3-axis gyroscope, compass
Other sensors (not used)	Microphone, speaker, humidity, temperature, pressure, gas, color, light
Sampling rate	60 Hz (accelerometer and gyroscope)
Accelerometer range	± 2 g to ± 16 g (sensitivity: 4800 LSB/g)
Gyroscope range	± 250 dps to ± 2000 dps
Data transmission	Bluetooth Low Energy
Data collection	Smartphone gateway (Realme Pro 7)
Video recording	30 fps, synchronized with IMU data for annotation

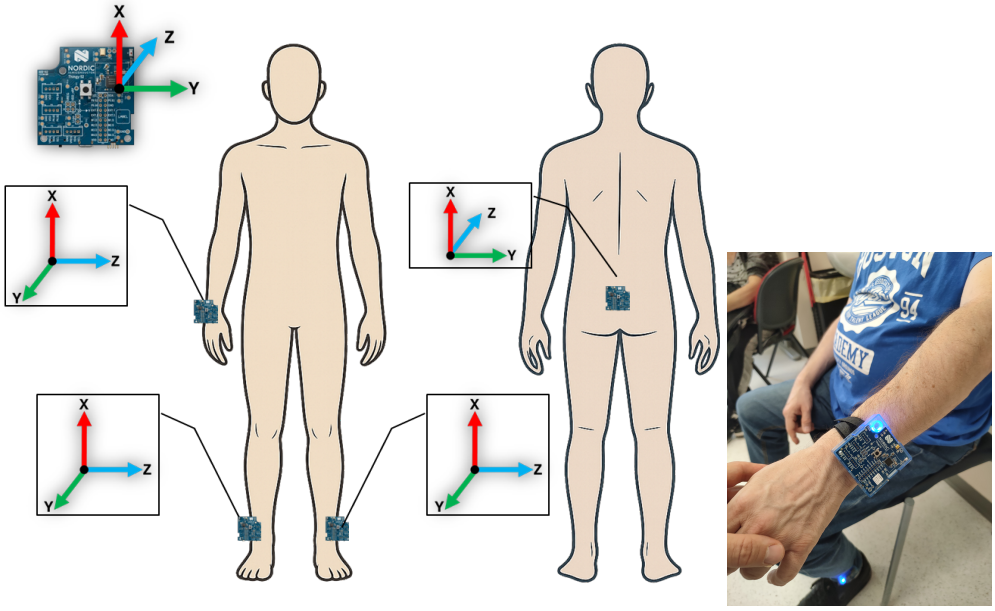
Table 4.35. Technical specifications of the IMUs used in the WBAN.

4.5.1.2 Data collection.

Participants performed a set of standardized motor tasks with self-paced rest intervals between trials. Clear instructions were provided before each task, and participants were free to withdraw at any time. The protocol included seven tasks covering daily activities, gait under environmental constraints, dual-task conditions, and turning maneuvers (Table 4.36). All participants performed the experiments in the "Off" condition, after a 12-hour washout period, to increase the likelihood of FoG episodes.

4.5.1.3 Data processing.

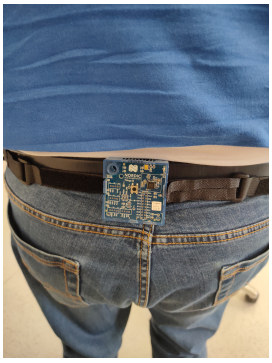
Video annotation. Two neurologists, both specialists in PD and movement disorders, independently annotated activities and FoG episodes by reviewing synchronized video recordings, following a standardized assessment protocol. The definition of



(a) Schematic representation of sensor placement and orientation axes.

(b) Sensor on the (most affected side) wrist.

(c) Sensors on the ankles.



(d) Sensor on the lower back.

Figure 4.3. Sensor placement configuration for the dataset acquisition. (a) Schematic rendering with orientation of the local coordinate systems for each sensor. (b) Real-world sensor attachment examples: wrist, (c) ankles, and (d) lower back.

FoG (inability to produce effective steps) and its subtypes (start hesitation, turn hesitation, FoG during turning, hesitation in narrow spaces, and destination/open-space hesitation) were adopted from prior studies [30, 117]. FoG onset was defined as the frame corresponding to the last effective step before freezing, and offset as the first effective step afterward. The severity of each FoG episode was rated on a three-level scale (Table 4.37); when different severities occurred sequentially within the same episode, adjacent segments could receive different labels.

Task ID	Activity
1	Timed up and go test
2	Keep a static upright position with eyes open for 1 minute
3	Walk back and forth for 10 meters
4	Walk back and forth for 10 meters with passage through a doorway
5	Walking back and forth for 10 meters carrying a glass full of water
6	Walking back and forth for 10 meters counting backward from 100 to 0 with steps of 7
7	360 degree turn

Table 4.36. Tasks included in the experimental protocol.

Severity	**Manifestation**
1	Shuffling forward with small steps
2	Trembling in place with alternating rapid knee movements
3	Complete akinesia, no limb or trunk movement

Table 4.37. Severity scale for FoG episodes.

To simplify annotation, video recordings were downsampled to 10 fps, ensuring a temporal resolution of 100 ms for both activity and FoG identification. Annotations were performed using the Python Video Annotator software [118] and exported in CSV format. The intra-class correlation coefficient (ICC) between the two raters was 0.80 for the number of FoG episodes and 0.88 for the percent time spent with FoG, indicating good-to-strong inter-rater agreement [119, 120]. Discrepancies were resolved through consensus; a conflict was flagged when (i) one rater marked an episode not identified by the other, (ii) overlap between annotated episodes was below 80%, or (iii) onset times differed by more than 0.5 s. For episodes with high inter-rater agreement, final onset and offset were computed as the average of both annotations.

Sensor synchronization. In the WBAN, individual nodes did not communicate with each other and no native synchronization with the smartphone aggregator

was available; synchronization of the IMUs was therefore performed offline using an empirical procedure [115]. For each node, we estimated the delay between the smartphone issuing the "connect"/"start acquisition" commands and the effective start of sensing at the node. Since samples were timestamped upon reception by the smartphone, these delays were used to back-calculate the approximate generation time of each sample at the node. A synchronization offset Δt was computed for each session and subtracted from the smartphone timestamps to realign samples across nodes. Because BLE transmission latency exhibits limited variability relative to the 16.7 ms sampling period at 60 Hz, a single session-level correction was sufficient in our recordings. The workflow is illustrated in Figure 4.4. As both video frames and BLE packets are handled by the same smartphone app [115], video and IMU streams share a common time base, enabling frame-accurate localization of FoG onset/offset without external hardware.

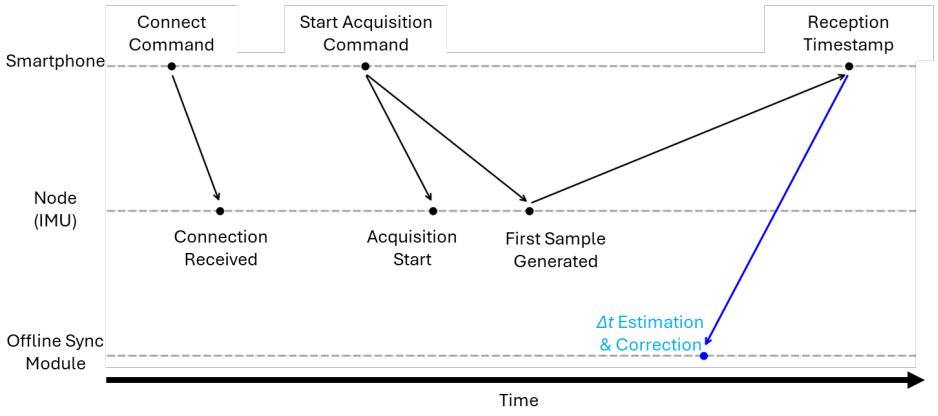


Figure 4.4. Overview of the synchronization workflow. Commands are issued by the smartphone, processed by the IMU nodes with intrinsic delays, and aligned offline by estimating a time offset (Δt) to reconstruct the approximate sample generation time.

4.5.2 Data

4.5.2.1 Data records.

The dataset is deposited and available at https://zenodo.org/records/17838806 [121]. To ease dataset exploration, visualization, and analysis, data were organized in two *CSV* files: *Sensor Data* and *Clinical Data*.

Sensor data. The file *sensor_data.csv* includes the sensor readings from all experiments. The file has $m = 329{,}027$ rows/samples and $n = 31$ columns/variables.

Variables include the *timestamp* ($F_s = 60$ Hz); 24 *sensor readings* variables, made up of 6 channels (3-axis accelerometer and 3-axis gyroscope) recorded by each of the 4 IMUs; the *activity* label indicating the activity class for each sample; the *FoG* label indicating the presence of FoG in each sample; and *FoG severity*, describing the severity of FoG in each sample. It is worth noting that activity and FoG labels are stored as independent categorical/binary variables and can overlap in time. Three additional columns report the ID of the *subject*, *session*, and *task* for each sample. A detailed description of the file is provided in Table 4.38. Each sample associated with a FoG event also has a corresponding activity class, providing context for the FoG manifestation.

Clinical data. A second file, named *clinical_data.csv*, contains the demographic and clinical information of each participant. The file is organized into $m = 22$ rows (one per subject) and $n = 10$ columns/variables. The *subjectID* variable matches the *subjectID* in *sensor_data.csv*, enabling direct linkage. Variables include demographic descriptors (*age*, *gender*) and clinical measures capturing disease duration and progression (*disease_duration*, *h_y*), motor performance (*updrs_iii*), FoG severity (*fog_q*), cognition (*moca*), fear of falling (*fes_i*), and quality of life (*pdq_8*). A detailed description is provided in Table 4.39.

4.5.2.2 Data overview.

The dataset offers a detailed representation of motor activities and FoG episodes in individuals with PD. To support dataset comprehension, we report representative traces and distributions describing recording times and FoG severity. Figure 4.5 shows an example segment with synchronized labels (task, activity, FoG, severity) overlaid on lower-back acceleration.

Figure 4.6a illustrates recording time distributions across (i) subjects, reflecting inter-individual variability in total duration (ranging from about 2 to 10 minutes per subject); (ii) tasks, showing that most time was spent in walking-related tasks and the 360° turn; and (iii) activity classes, highlighting the prevalence of walking, turning, and standing compared to postural transitions such as sit-to-stand.

Figure 4.6b focuses on FoG severity annotations. Episodes are categorized into three severity levels—shuffling, trembling, and akinesia—and are presented as total annotated samples, cumulative duration in seconds, and equivalent duration in minutes. Table 4.40 summarizes event count and duration statistics stratified by severity type.

Column	Variable	Description
1	*Timestamp*	Float number corresponding to the estimated (as shown in Figure 4.4) timestamp in milliseconds. The sampling rate is equal to 60 Hz.
2–25	*Sensor readings*	Float numbers representing inertial signals, stored in the following format: [*sensor position*]_[sensor type]_[direction], where the sensor position includes left (ankleL) and right ankle (ankleR), back, and wrist; sensor type includes accelerometer (acc) and gyroscope (gyro); and the direction comprises x, y, and z. The unit of measurement for the accelerometer and gyroscope readings is g-force and degree/s, respectively.
26	*activity*	Integer number that identify the specific activity performed. It includes 1: walking, 2: sit, 3: stand, 4: sit-to-stand, 5: stand-to-sit, 6: turn-right, and 7: turn-left.
27	*fog*	Binary value that identify the presence of FoG. It can be either 0 (non-FoG) or 1 (FoG).
28	*fog_-severity*	Integer number indicating the severity of FoG, including 1: shuffling forward with small steps, 2: trembling in place with alternating rapid knee movements, and 3: complete akinesia without limbs or trunk.
29	*subjectID*	Integer number in the range from 1 to 22 that identify the specific subject. This number corresponds to that in the related demographic and clinical information.
30	*sessionID*	Integer number referring to the data acquisition session. In most cases, a single session was recorded for each subject. In some particular cases (e.g., issues in sensor synchronization), additional session(s) were registered.
31	*taskID*	Integer number corresponding to the specific motor task. It includes 1: timed up-and-go, 2: stand for 1 minute, 3: walk back and forth, 4: walk back and forth with passage through a doorway, 5: walk back and forth carrying a glass full of water, 6: walk back and forth counting backward, and 7: 360 degree turn. Not all tasks are available for each subject.

Table 4.38. Variables for sensor-related data.

Column	Variable	Description
1	*subjectID*	Integer number in the range from 1 to 22 that identify the specific subject. This number corresponds to that in the related sensor data.
2	*age*	Integer number referring the participant's age.
3	*gender*	Categorical value (M/F) corresponding to the participant's gender.
4	*disease_-duration*	Integer number reporting the number of years from the clinical diagnosis.
5	*h_y*	H&Y stage. Float number between 0 and 5, proportional to the disease progression.
6	*updrs_iii*	Integer number from 0 to 76, corresponding to the sum of all items from the MDS-UPDRS part III. Higher scores correspond to more severe motor impairment.
7	*fog_q*	Integer number from 0 to 24, calculated as the total score from the FOG-Q. Higher scores correspond to more severe FOG.
8	*moca*	Integer number from 0 to 30, calculated as the total MOCA score. Lower scores correspond to more severe cognitive decline.
9	*fes_i*	Integer number between 16 and 64, obtained as the total score from the FES-I. Higher scores correspond to more severe concern about falling.
10	*pdq_8*	Integer number between 0 and 32, calculated as the sum of all items from the PDQ-8. Higher scores signify poorer quality of life.

Table 4.39. Variables for demographic and clinical-related data.

The total number of severity-specific events exceeds 101 (the number of FoG episodes) because six episodes contain adjacent segments with different severity labels (e.g., transitions between shuffling, trembling, and akinesia).

When integrating this dataset with other PD cohorts, it's recommended normalizing sampling frequencies and time axes, and verifying annotation formats for consistency, as done in a previous cross-dataset study [122]. It is also worth noting

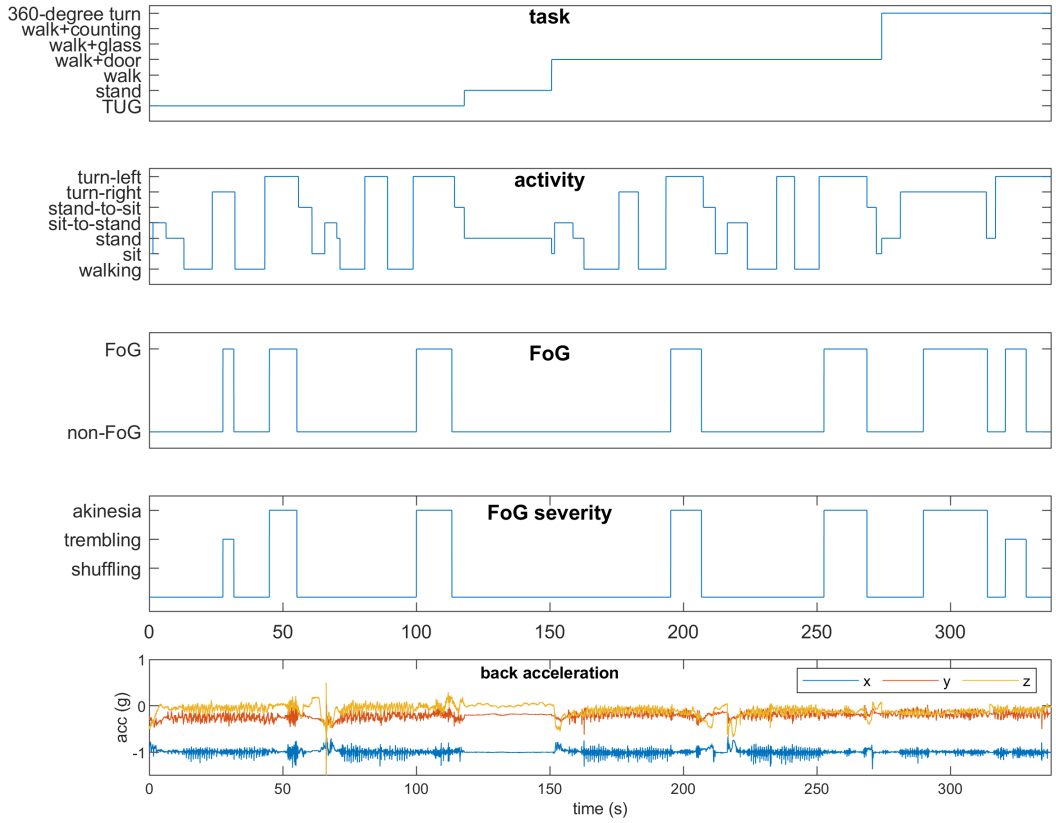
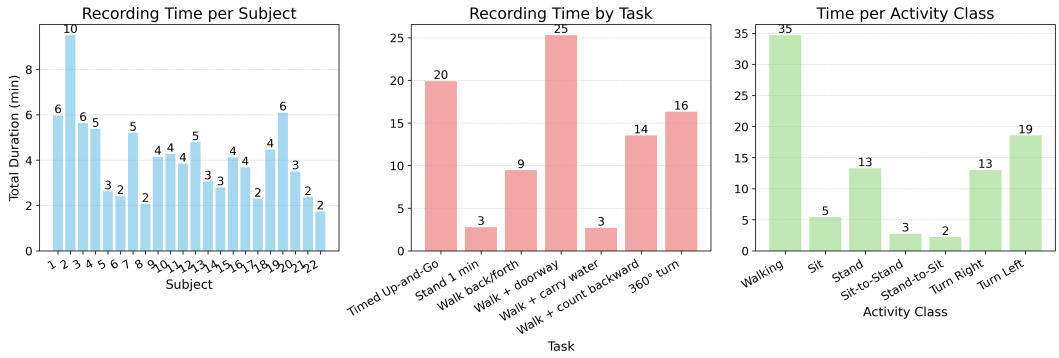


Figure 4.5. Representative synchronized view of the dataset. From top to bottom: task, activity, binary FoG label, FoG severity (1–3)/manifestation(shuffling, trembling, akinesia), and raw acceleration from the lower-back IMU. This figure illustrates how annotations align with sensor signals and can be reproduced with the provided scripts.

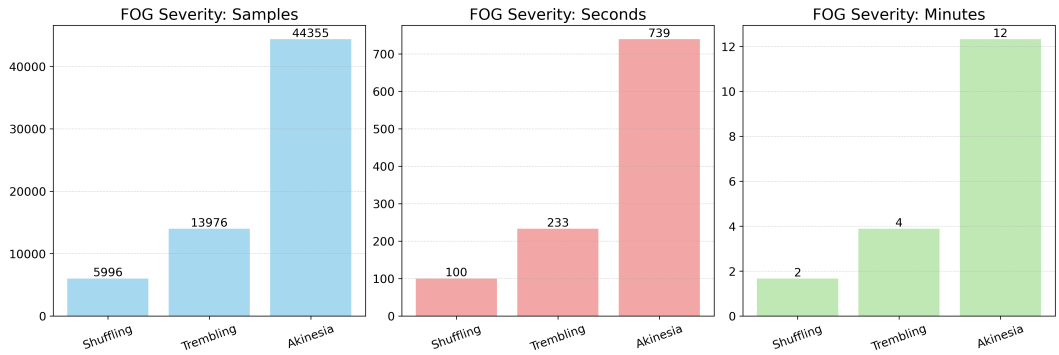
Type	Events	Mean (seconds)	Median (seconds)	Min (seconds)	Max (seconds)	95^{th} perc. (seconds)
Shuffling	33	3.02	2.40	0.6	9.9	7.04
Trembling	40	5.82	4.90	0.2	24.1	12.40
Akinesia	37	19.98	12.40	1.8	108.68	71.65

Table 4.40. Descriptive statistics of FoG episodes by severity type.

that data were recorded from patients in the "Off" condition to control for motor fluctuations and increase the probability of FoG manifestation. The standardized CSV structure was designed to facilitate integration and ensure compatibility across studies. There are no restrictions on the use of this dataset beyond the accompanying license. All data have been fully anonymized, and no personally identifiable information is



(a) Recording time by (i) Subject, (ii) Task, and (iii) Activity.



(b) Distribution by FoG Severity in terms of (i) Samples, (ii) Seconds, and (iii) Minutes.

Figure 4.6. Dataset overview. (a) Recording time distributions across subjects, tasks, and annotated activity classes; each bar indicates the total duration of recordings, highlighting variability across participants and protocol tasks. (b) Distribution of FoG severity levels—shuffling, trembling, and akinesia—quantified in annotated samples, cumulative seconds, and minutes. Together, the panels illustrate the dataset composition in terms of both motor activity coverage and FoG manifestations.

included; users are nevertheless encouraged to follow best practices for secure data handling, particularly when combining this dataset with other clinical resources.

In summary, FOGSTAR complements the dataset landscape reviewed in this chapter by providing a wearable-only resource with standardized sensing, explicit activity context, and fine-grained FoG annotations (including severity), supported by rigorous clinical characterization and a reproducible synchronization and labeling workflow. Together with the datasets surveyed in Section 4.2, it highlights how different design choices—ranging from highly controlled protocols to more ecological recordings—shape the type of analyses and the level of clinical interpretability that can be achieved. The following chapters build on this landscape to investigate and compare methodological approaches on these resources, leveraging both established datasets and FOGSTAR to support robust development and evaluation of wearable-based

models for PD motor assessment.

4.6 Summary and Implications for Methodological Development

The dataset landscape reviewed in this chapter highlights both the rapid progress and the persistent fragmentation of wearable-based research in Parkinson's disease. Over the past decade, numerous datasets have been developed to capture motor behavior using wearable sensors, ranging from tightly controlled laboratory experiments to long-term monitoring in real-world environments. Together, these resources provide a valuable foundation for studying gait impairments, postural transitions, and freezing of gait.

At the same time, the analysis reveals substantial variability across datasets in terms of cohort characteristics, experimental protocols, sensing modalities, and annotation strategies. Some datasets prioritize precise event-level annotation under controlled conditions, while others emphasize ecological validity through recordings of daily-life activities. Sensor configurations also vary widely, from single-sensor setups aimed at practical deployment to complex multimodal systems capturing detailed biomechanical or physiological signals.

These differences have important methodological implications. On one hand, richly annotated laboratory datasets enable the development and validation of supervised learning models targeting specific motor events such as freezing episodes. On the other hand, datasets collected in more naturalistic conditions are essential for evaluating the robustness of these methods in realistic scenarios. As a consequence, no single dataset fully captures the complexity of Parkinsonian motor behavior, and complementary datasets must often be combined or analyzed jointly.

The FOGSTAR dataset introduced in this chapter contributes to this landscape by providing a wearable-only resource with standardized sensing, explicit activity context, and fine-grained annotation of freezing events and their severity. Positioned alongside existing datasets, it enables controlled experimentation while remaining compatible with common wearable sensing configurations.

The dataset analysis presented in this chapter not only characterizes the current landscape of wearable-based FoG research but also directly motivates the methodological developments described in Chapter 5. The substantial variability observed in sensor configurations, body locations, and acquisition protocols motivated the

development of harmonization procedures and the Mixture-of-Experts frameworks designed to explicitly address sensing heterogeneity. The scarcity and imbalance of annotated FoG events motivated the investigation of transfer learning and fine-tuning strategies aimed at improving data efficiency and personalization. Similarly, the diversity of annotation procedures and experimental protocols reinforced the need for patient-wise validation, cross-dataset evaluation, and the adoption of performance metrics suitable for imbalanced classification problems. Therefore, the methodological choices explored in this thesis were not developed independently of the dataset landscape but emerged directly from the limitations and challenges identified through the systematic analysis of currently available datasets.

The following chapter builds upon these points. In particular, it introduces the signal processing pipelines, modeling strategies, and evaluation procedures used to analyze wearable motion signals and to develop computational approaches for the automatic detection of motor symptoms in Parkinson's disease.

Chapter 5

Methods

5.1 HAR for FoG-linked activities in Parkinson

This section describes the methodological pipeline used to investigate HAR for motor activities associated with FoG. The analysis is conducted using the WearGait-PD dataset, which was introduced in Chapter 4. This dataset provides synchronized recordings from multiple wearable inertial measurement units (IMUs) together with detailed annotations of locomotor tasks and postural transitions.

WearGait-PD was selected for this stage of the study for several reasons. First, it includes a rich set of annotated motor activities—such as walking, turning, and postural transitions—that represent the behavioral contexts in which FoG episodes are commonly observed. Second, the dataset provides a relatively large number of recordings compared with many FoG-focused datasets, enabling the training and evaluation of activity recognition models under subject-disjoint conditions. Finally, the use of multiple body-mounted sensors allows the analysis of motion dynamics from different anatomical perspectives, which is valuable for learning robust representations of Parkinsonian motor behavior.

Building reliable HAR models on this dataset therefore provides a structured way to characterize the motor context surrounding FoG and to learn representations of movement primitives that can support subsequent FoG detection models.

5.1.1 Pre-processing

Under-sampling: The raw data collected from the 13 IMUs is sampled at 100 Hz. To improve computational efficiency while preserving relevant dynamics of human motion, uniform temporal downsampling was applied to reduce the sampling

frequency to 50 Hz. This is done independently for each trial (i.e., each CSV file) by retaining every second row using index-based slicing.

Remove irrelevant events: Following synchronization and merging of IMU data streams, segments associated with irrelevant general events and standing balance trials were removed (e.g., `unlabeled`, `EO_FeetShoWidth`). This step reduces noise and ensures that only structured, interpretable locomotor tasks are preserved in the dataset.

Remove irrelevant features: To refine the dataset further, all features (columns) that exhibit zero variance across samples were eliminated. Also, any feature containing `NaN` values across the dataset was discarded to prevent training biases and inconsistencies during model fitting. This step ensures that all retained features are informative and reliable.

Segmentation: A sliding window segmentation strategy with a fixed window size of 1-second (i.e., 50 samples) was applied. Each segment represents a multivariate time-series snapshot of sensor readings. For each segment, the raw sensor values is retained along with corresponding metadata (e.g., `Task`, `Subject`, `SourceFile`). Segment labels are assigned using a majority voting approach, taking the mode value of the ground-truth task label within the window.

Training-test split: To ensure fair evaluation and prevent data leakage, we perform a subject-disjoint train/test split. Specifically, we use a 75%/25% hold-out strategy at the subject level, where all windows associated with selected training subjects are included in the training set and the remaining in the test set. This guarantees that no overlapping samples or sessions appear in both splits.

5.1.2 Machine Learning models

Three classic ML classifiers were implemented, trained, and optimized. These models were selected due to their superior performance in HAR tasks [123, 124].

RF is an ensemble of decision trees, robust to overfitting and effective in handling high-dimensional data. The number of trees, maximum depth, and feature subset size parameters were tuned.

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
```

```
3
4 param_grid = {
5 'n_estimators': [50, 100, 150, 200],
6 'max_depth': [None, 10, 20],
7 'min_samples_split': [2, 5, 10]
8 }
9 model = GridSearchCV(RandomForestClassifier(random_state=42),
    param_grid, cv=3)
10 model.fit(X_train, y_train)
```

Listing 5.1: Random Forest Classifier with Grid Search

SVM is effective in high-dimensional spaces and works well for non-linear boundaries when combined with a Radial Basis Function (RBF) kernel. A grid search over the regularization term (C) and kernel coefficient (gamma) was performed.

```
1 from sklearn.model_selection import GridSearchCV
2 from sklearn.svm import SVC
3
4 param_grid = {
5 'C': [1, 10],
6 'gamma': [0.01, 0.001],
7 'kernel': [linear, poly, rbf]
8 }
9 model = GridSearchCV(SVC(probability=True), param_grid, cv=3)
10 model.fit(X_train, y_train)
```

Listing 5.2: Support Vector Machine with Grid Search

Multi-layer Perceptron (MLP) learns non-linear patterns using multiple fully connected layers. The architecture and learning rate were tuned using grid search.

```
1 from sklearn.model_selection import GridSearchCV
2 from sklearn.neural_network import MLPClassifier
3
4 param_grid = {
5 'hidden_layer_sizes': [(128, 64), (256, 128)],
6 'learning_rate_init': [0.001, 0.0005],
7 'max_iter': [300]
8 }
9 model = GridSearchCV(MLPClassifier(random_state=42), param_grid
    , cv=3)
10 model.fit(X_train, y_train)
```

Listing 5.3: MLP Classifier with Grid Search

Hyperparameter tuning was performed using a grid-search approach over a three-fold cross-validation to identify the optimal configuration of parameters. Table 5.1 reports the tuned parameters for each model, along with the search space.

Model	Parameter	Search space
RF	number of estimators, max depth, min samples split	(50, 100, 150, 200), (10, 20), (2, 5, 10)
SVM	cost, gamma, kernel	(1, 10), (0.01, 0.001), (linear, poly, rbf)
MLP	number of hidden layers, number of neurons, learning rate, maximum number of iterations	(2), (128, 64), (256, 128), (0.0005, 0.01), (50, 300)

Table 5.1. Tuned hyper-parameters and search space for machine learning models.

5.1.3 Deep learning models

Different end-to-end DL models were implemented and optimized. Deep learning models have demonstrated strong performance in time-series analysis and human activity recognition tasks [125–127], also due to their ability to handle directly raw data for providing the final output.

CNN exploits local spatial patterns in time-series segments via one-dimensional (1D) convolutional layers. It is well-suited for capturing motion patterns and transitions in short time frames [**?**]. A simple 1D-CNN was designed, with a single convolutional layer (64 filters, kernel size of 3) followed by a pooling layer (pool size of 2). The latter is connected to a dense layer with 64 neurons, and a final output layer. Rectified Linear Unit (ReLU) activation function was employed in all layers except for the final output, where a softmax activation function ensured that the output is expressed as a probability of belonging to each class.

```
from tensorflow.keras import models, layers
model = models.Sequential([
layers.Conv1D(64, 3, activation='relu', input_shape=input_shape
    ),
layers.MaxPooling1D(2),
layers.Flatten(),
```

```
layers.Dense(64, activation='relu'),
layers.Dense(num_classes, activation='softmax')
])
model.compile(optimizer='adam', loss='
    sparse_categorical_crossentropy', metrics=['accuracy'])
```

Listing 5.4: Convolutional Neural Network

Recurrent Neural Networks (RNNs) are specifically designed to handle temporal dependencies, thus they well fit time-series classification tasks [127]. Different RNNs were implemented and optimized, including LSTM, Bidirectional LSTM (Bi-LSTM), Bi-LSTM with Attention, and Gated Recurrent Unit (GRU). Bidirectional variants of LSTM process the input signals in both forward and backward directions, while attention-enhanced models learn to focus on salient time steps within each segment. The LSTM network has 2 LSTM layers with 128 and 64 neurons, and a dense layer with 64 neurons.

```
model = models.Sequential([
layers.LSTM(128, return_sequences=True, input_shape=input_shape
    ),
layers.LSTM(64),
layers.Dense(64, activation='relu'),
layers.Dense(num_classes, activation='softmax')
])
model.compile(optimizer='adam', loss='
    sparse_categorical_crossentropy', metrics=['accuracy'])
```

Listing 5.5: LSTM Network

Bi-LSTM networks are used to capture temporal dependencies in both forward and backward directions. This helps the model learn richer temporal patterns. The Bi-LSTM has 2 LSTM layers with 64 and 32 units, and a dense layer with 64 units.

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Bidirectional, LSTM, Dense

model = Sequential([
Bidirectional(LSTM(64, return_sequences=True), input_shape=(
    timesteps, features)),
Bidirectional(LSTM(32)),
Dense(64, activation='relu'),
Dense(num_classes, activation='softmax')
])
```

```
model.compile(optimizer='adam', loss='
    sparse_categorical_crossentropy', metrics=['accuracy'])
```

Listing 5.6: Bi-LSTM model using Keras

An attention mechanism can be applied on top of a Bi-LSTM network to help the model focus on the most relevant parts of the input sequence [128]. The LSTM with attention mechanism has a single LSTM layer with 64 units, a Global Average Pooling (GAP) layer with attention, and a dense layer with 64 units.

```
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, Dense, LSTM,
    Bidirectional, Attention, GlobalAveragePooling1D

inputs = Input(shape=(timesteps, features))
x = Bidirectional(LSTM(64, return_sequences=True))(inputs)
attention = Attention()([x, x])
x = GlobalAveragePooling1D()(attention)
x = Dense(64, activation='relu')(x)
outputs = Dense(num_classes, activation='softmax')(x)
model = Model(inputs, outputs)
model.compile(optimizer='adam', loss='
    sparse_categorical_crossentropy', metrics=['accuracy'])
```

Listing 5.7: Bi-LSTM with Attention model

GRUs are a lightweight alternative to LSTMs with fewer parameters and comparable performance. The GRU network has 2 GRU layers with 64 and 32 units, respectively, and a dense layer with 64 units [129].

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import GRU, Dense

model = Sequential([
GRU(64, return_sequences=True, input_shape=(timesteps, features
    )),
GRU(32),
Dense(64, activation='relu'),
Dense(num_classes, activation='softmax')
])
model.compile(optimizer='adam', loss='
    sparse_categorical_crossentropy', metrics=['accuracy'])
```

Listing 5.8: GRU model using Keras

Transformer-based Models, include temporal attention architectures (e.g., Temporal Convolutional Networks–TCN and Transformers) that handle sequence modeling efficiently through parallel self-attention mechanisms and long-range temporal dependency modeling [130]. A multi-head self-attention mechanism with 4 heads and a key dimension of 64 captures contextual relationships within the sequence. The attention output is then reduced via GAP, summarizing the sequence into a fixed-length vector. This representation is passed through a dense layer with 128 units and ReLU activation for feature transformation, and finally mapped to class probabilities through a softmax-activated output layer.

```
inputs = layers.Input(shape=input_shape)
x = layers.LayerNormalization()(inputs)
x = layers.MultiHeadAttention(num_heads=4, key_dim=64)(x, x)
x = layers.GlobalAveragePooling1D()(x)
x = layers.Dense(128, activation='relu')(x)
outputs = layers.Dense(num_classes, activation='softmax')(x)
model = models.Model(inputs, outputs)
model.compile(optimizer='adam', loss='
    sparse_categorical_crossentropy', metrics=['accuracy'])
```

Listing 5.9: Transformer-based Model

All models were trained with an Adaptive Moment (Adam) optimizer, sparse categorical cross-entropy loss function, and accuracy evaluation metric. Training was performed with a batch size of 32 and a learning rate of 0.001 , and maximum number of iterations of 200.

5.1.4 Model evaluation

Performance is evaluated on the test set using standard classification metrics.

Accuracy (Eq. 5.1) measures the overall correctness of the model. Recall (or sensitivity) (Eq. 5.2) measures the ability of the model to detect all actual positives. Precision(Eq. 5.3) indicates the proportion of true positive predictions among all predicted positives and reflects how reliable positive predictions are. F1-score (Eq. 5.4) is the harmonic mean of precision and recall, providing a balanced metric in classification tasks with imbalanced distribution of instances in the different classes. TP and TN represent True Positives and True Nagatives instances respectively, while FP and FN refer to False Positive and False Nagative instances.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$recall = \frac{TP}{TP + FN} \tag{5.2}$$

$$precision = \frac{TP}{TP + FP} \tag{5.3}$$

$$F1 - score = \frac{2 \cdot recall \cdot precision}{recall + precision} \tag{5.4}$$

The confusion matrix (Tab. 5.2) is a $C \times C$ matrix where each element M_{ij} represents the number of instances from class i predicted as class j. It is essential to provide a full picture of the distributions of errors across the different classes.

Actual \Predicted	**Class 1**	**Class 2**	…	**Class C**
Class 1	$TP_{1,1}$	$FP_{1,2}$	…	$FP_{1,C}$
Class 2	$FN_{2,1}$	$TP_{2,2}$	…	$FP_{2,C}$
⋮	⋮	⋮	⋱	⋮
Class C	$FN_{C,1}$	$FN_{C,2}$	…	$TP_{C,C}$

Table 5.2. Example structure of a confusion matrix for C classes.

5.1.5 Experimental Results

This section presents the experimental results obtained using the described pipeline and models. The models were evaluated using a subject-disjoint hold-out split, and performance was measured using standard metrics including accuracy, precision, recall, and F1-score.

5.1.5.1 Performance Evaluation

Table 5.3 summarizes the classification performance of the evaluated models on the WearGait-PD dataset using a 1-second sliding window. The RF classifier achieved the highest accuracy, precision, recall, and F1-score, all at 92%. Other ML models, such as SVM, demonstrated significantly lower performance. Shallow neural networks (ANNs) cannot adequately model complex signal patterns, as evidenced by classification metrics of less than 80%. Among the DL models, LSTM-based models achieved the best results, performing significantly better than CNN, RNN, GRU, and Transformer approaches. These results are expected, as LSTM models are specifically designed and optimized to process time series by capturing temporal dependencies.

The CNN showed the worst results, possibly due to a small feature extraction module, in addition to the fact that CNN is better able to model spatial relationships than temporal patterns. Interestingly, attention mechanisms did not improve performance in LSTM while increasing computational complexity. Similarly, Bi-LSTM performed similarly to LSTM while requiring past and future data to compute predictions.

		Macro		
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RF	92%	92%	92%	92%
SVM	85%	84%	85%	84%
MLP	78%	79%	78%	78%
CNN	83%	83%	83%	82%
LSTM	91%	91%	91%	91%
RNN	76%	77%	76%	76%
Bi-LSTM	91%	91%	91%	91%
Bi-LSTM + Attention	90%	91%	90%	90%
GRU	88%	88%	88%	88%
Transformer	85%	85%	85%	85%

Table 5.3. Performance metrics of various models evaluated on the WearGait-PD dataset.

The confusion matrices for the best ML (RF) and DL (LSTM) models are reported in Figure 5.1, providing deeper insights into per-class classification performance.

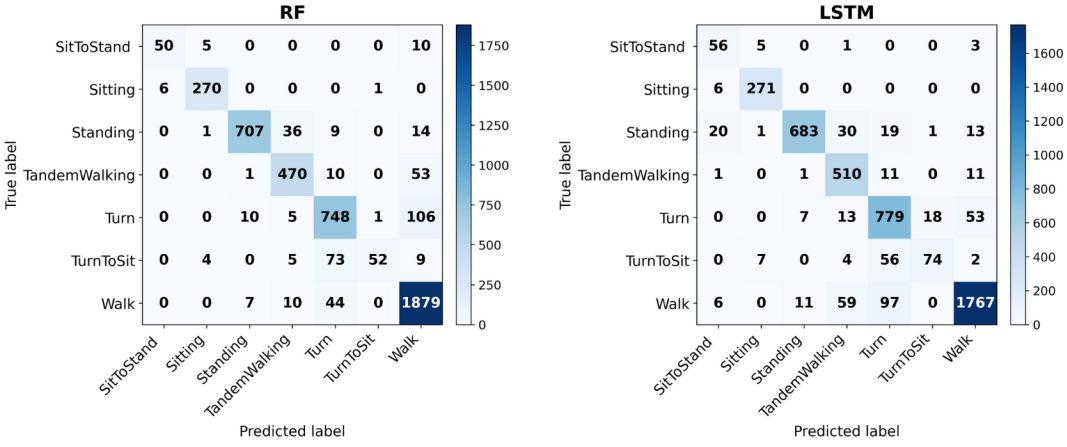


Figure 5.1. Confusion matrices for Random Forest (RF) and Long Short Term Memory (LSTM). Results refer to the test set.

The results are consistent for the two models and show that a subset of samples related to `SitToStand` activity is confused with either `Sitting` or `Walking`. This

is expected, since short portions of `Sitting` may be included in `Sit2Stand`, due to the non-perfect temporal resolution in data labelling. In addition, `Sit2Stand` is followed by `Walking` in the list of scripted tasks. `Sitting` is classified with very high accuracy, while a small fraction of `Standing` is confused with `Walking` or `TandemWalk`. `TandemWalk` is confused with `Walking` only in a small portion of cases (around 10%), and this proves that models properly learned specific signal patterns (e.g., short vs normal steps). Around 10% of `Turning` is confused with `Walking`, demonstrating the difficulty in discriminating these two activities with very high precision. Finally, `Turn2Sit` represents the most challenging activity, confused with `Turning` in more than half cases, while `Walking` is classified with an accuracy over 91%.

Analysing the differences between the two models, it emerges that the LSTM model better recognize `Sit2Stand`, `TandemWalk`, `Turning` and `Turn2Sit`. On the other hand, the RF model gives the best results in classifying `Standing` and `Walking`, with the latter representing the most numerous class.

5.1.5.2 Evaluation of sensor locations

To investigate the importance of each sensor in recognizing human activities, the performance as calculated for each specific sensor position, as previously done in similar studies [100, 131]. The analyses were conducted by training and evaluating the RF model as previously described, and changing the input data at each iteration. The results are reported in Fig. 5.2, expressed in terms of macro F-score (i.e., arithmetic mean of the F-score for each class). This metric assign to all classes the same weight, while the weighted F-score weights per-class performance by their related support.

Sensors positioned on the thighs performed best, with an F1-score of 80%-81%. These were followed by sensors located on the lower back (70%), sternum/xiphoid (67%), and head (66%). Sensors on shanks and ankles resulted in significantly lower performance, with an F1-score ranging from 58% to 60%. Sensors positioned on the extremities (wrists and feet) provided the worst performance, with an F1-score in the range of 48-55%. Overall, the gyroscope provided significantly better results than the accelerometer. This is particularly evident for the devices on the shanks (+15% in F1-score), sternum (+12%), and feet (+9%). The only exception is represented by the sensor on the head, where the accelerometer performed better (+9%) than the gyroscope. Combining both sensors is beneficial, as it provides an average increase in the F1-score of 4.9% (SD: 3.7%, Range: 0–11%) over the best sensor only.

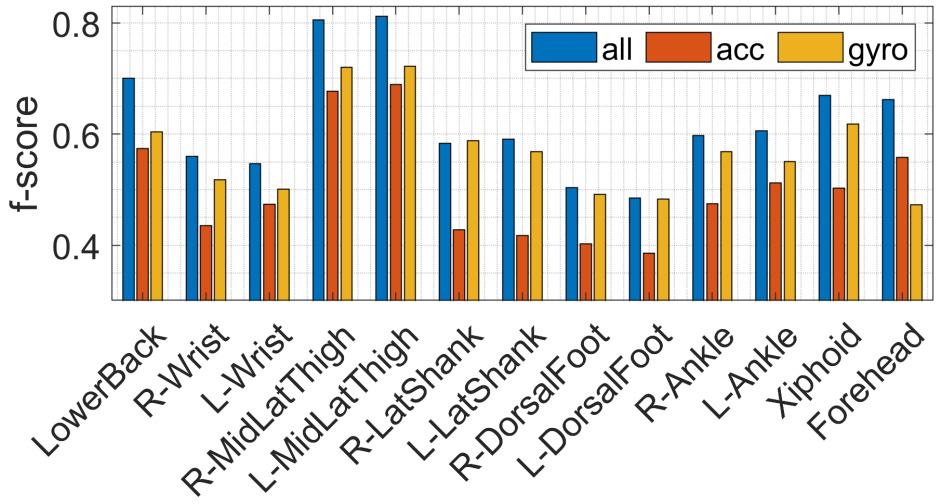


Figure 5.2. Comparison of performance for each specific sensor position. Results refer to test set and the random forest algorithm. Acc: accelerometer; gyro: gyroscope; all: combination of accelerometer and gyroscope.

5.2 Fine-tuning for optimized FoG detection

This section outlines the pipeline adopted in this study, encompassing data preprocessing, feature extraction and selection, model training, and fine-tuning. This pipeline transforms raw wearable sensor signals into a structured representation suitable for FoG detection. A schematic overview is provided in Fig. 5.3, and each component is described in the following subsections.

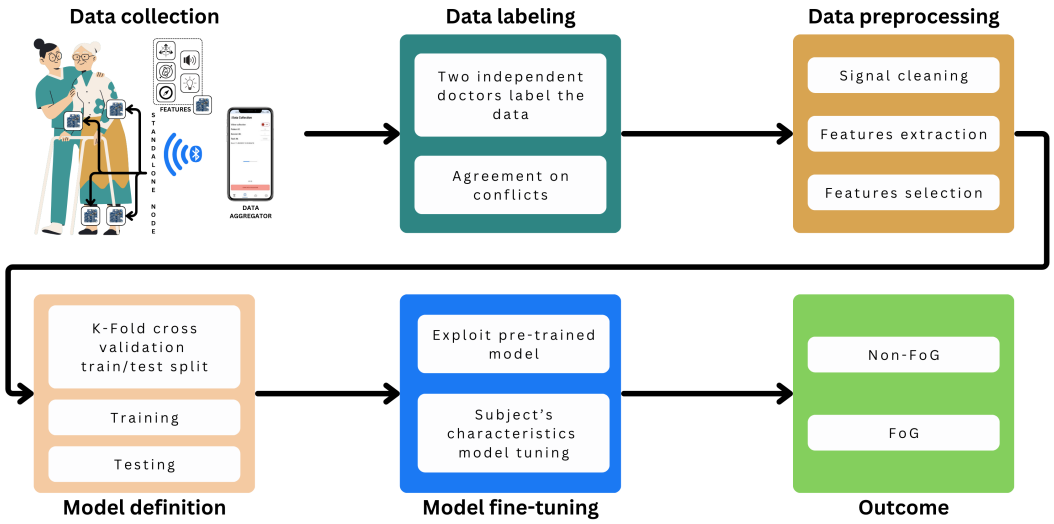


Figure 5.3. Schematic overview of the protocol used in this study for fine-tuning for FoG detection

5.2.1 Dataset

The experiments described in this section were conducted using the FoG-STAR dataset [94], a multi-level annotated wearable dataset designed for the study of gait freezing in Parkinson's disease, which was presented in Chapter 4.

The dataset includes synchronized inertial signals acquired from multiple body locations, together with detailed annotations of FoG episodes, motor activities, and symptom manifestations. Its structured annotation scheme and multi-sensor configuration make it suitable for both activity recognition and FoG detection tasks, as well as for investigating subject-specific adaptation through fine-tuning.

5.2.2 Data preprocessing

The phase of Data preprocessing was guided by the open-source Baseline HAR (B-HAR) framework [132, 133], which serves as the container of each of the following steps.

5.2.2.1 Signal cleaning:

The signal processing step to enhance data quality comprises the utilization of a fourth-order Butterworth filter in bandpass mode to mitigate the impact of signal noise on the analysis [134]. A check for NAs in sensors' signals revealed 24 data collections with at least one column in the signal section affected by missing values. Since these NAs were present throughout the entire duration of the tasks, interpolation was not achievable, so these data collections had to be excluded from further analysis. Since the sampling frequency of the sensors is 60 Hz, the Nyquist frequency is 30 Hz. Thus, the filter removes noise at frequencies below 0.5 Hz and above 20 Hz, keeping only the frequencies related to human motion. This procedure helped reduce potential biases and errors in the dataset and ensured that the subsequent stages relied on data of the highest possible quality.

5.2.2.2 Feature extraction:

The Time Series Feature Extraction Library (TSFEL) [135] was used to perform feature extraction. Features were extracted across all three domains: statistical, temporal, and spectral, obtaining 205 features for each segment/window of data.

These features capture a broad range of characteristics necessary to enhance the classification capabilities of ML models. Temporal windows for feature extraction were strategically selected to cover a wide range of the signal's dynamics over time: the segment/window sizes selected are 2, 3, and 4 seconds long, with varying degrees of overlap: 0%, 25%, 50%, and 75%. To determine the label of each window the *majority* approach was used. This means that a segment is labeled as the majority of the labels into it. For example, if there is a segment of 2 seconds (i.e., 120 samples of data) where 100 samples are labeled as FoG and the remaining 20 samples as Non-FoG, the segment label will be FoG. This method ensures that the assigned label reflects the most frequent activity within the window, aiming for a precise representation of the window's primary activity.

5.2.2.3 Feature selection:

The last phase of the data preprocessing pipeline focuses on identifying the most informative features from the extensive set extracted previously. This selection process consists of three progressive steps to refine the feature set, guaranteeing the retention of the most relevant features. Firstly, features showing minimal variability were filtered out, as these are unlikely to contribute meaningful information for distinguishing between different states or conditions. Specifically, any feature with a variance below a ± 0.001 threshold was removed. After removing low-variance features, the focus shifted to multicollinearity, which can compromise the interpretability of ML models and increase the importance of correlated features. To address this, features exhibiting a high correlation with others were eliminated, setting the threshold at a Pearson correlation coefficient of 0.95 [136]. The final step in the feature selection process involved leveraging a tree-based method to further discern the most impactful features. To do so, 50 randomized decision trees were fitted, a strategy that allowed measurement of the relative importance of each feature based on how frequently it was used to split decision nodes across the ensemble of trees.

5.2.3 Model definition

In this subsection, the definition of the models is described, focusing on two key aspects: the division of the dataset into training and testing sets and the subsequent training of the models.

5.2.3.1 Train-Test split

A k-fold cross-validation approach ($k = 7$) at the subject level was employed to appropriately prepare the data for model training. This involved using five sets for training (70% of the subjects), one for validation (15% of the subjects), and one for testing (15% of the subjects). This process was repeated k times across the folds, allowing independent training, validation, and testing of the models over all subjects. This procedure is performed for each time window length and overlap mentioned in Section 5.2.2.2. This training, validation, and testing procedure aims to maximize the generalization capabilities of the proposed FoG recognition models. It is important to note that at this point, patients with no FoGs at all were left out from further steps, because i) they are not informative for what this study was aimed to and ii) possible formation of test partitions with few or no FoGs was an issue in the step of model evaluation.

5.2.3.2 Training

The capabilities of several ML and DL models were explored to provide a comprehensive overview of the studied problem. In particular, among the ML models, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Random Forest (RF) were selected. Additionally, the following DL models were defined: Convolutional Neural Network (CNN), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM). During training of the ML models, the grid search hyperparameter tuning phase was integrated into the workflow to optimize the performance of the LDA, QDA, and RF models. This systematic approach ensures identification of the most effective parameter combinations for these models. The DL models share some common characteristics: they are trained for 200 epochs, using a batch size of 32, *softmax* as output activation function, and the *Adam* optimizer with a learning rate set at 0.0001.

1 The models' training features early stopping, and the best models are saved at each epoch if they achieve the minimum loss on the validation set.

5.2.4 Fine-tuning

The complex characteristics of PD and their manifestations through motor symptoms, such as FoG, introduce significant variability in PwPD data. Factors such as the severity of symptoms, the stage of PD, and individual patient characteristics can significantly influence the movement data collected. Hence, deploying a model trained on data from a specific group of patients and expecting it to perform accurately on a different set of patients or under varied conditions often leads to suboptimal results. Furthermore, training a model from the ground up for each new patient or set of conditions is not feasible due to the extensive time and data requirements.

To address these challenges, fine-tuning [137, 138] is employed to adapt models more efficiently to new patient data with reduced additional training. Fine-tuning involves adjusting the weights of an already trained model to suit new data better. This adjustment can be applied selectively to specific layers of a neural network. In this application to PD and FoG detection, the model was fine-tuned by freezing the initial convolutional layers, which are generally responsible for capturing basic motion features typical across different patients. These features include general movement patterns and other non-specific motor characteristics not uniquely affected by PD. The deeper layers of the model, which are more specialized in detecting specific patterns specific to PD symptoms like FoG, are then fine-tuned with new patient data. This approach allows the model to leverage its existing knowledge base while adapting to the unique characteristics of new patients or conditions, thus enhancing its accuracy.

5.2.5 Experimental results

5.2.5.1 FoG classification results

To measure and compare the performances of the FoG recognition DL/ML models mentioned above, the training and testing approach introduced in Section 5.2.3 was used. In particular, the FOGSTAR dataset was used to train and test three DL models (i.e., CNN, LSTM, MLP) and three classic ML models (i.e., RF, LDA, QDA). Table 5.4 presents the results obtained during both the training and testing phases of all models, which were employed conventionally without undergoing fine-tuning. The results show good accuracy on the train set, but all the proposed models show a noticeable drop in the test set metrics. The precision, recall, F-1 score are calculated for each label, and then their unweighted mean is calculated. The CNN model achieved the best performances overall in the 2-second time window with 75% overlap (i.e.,

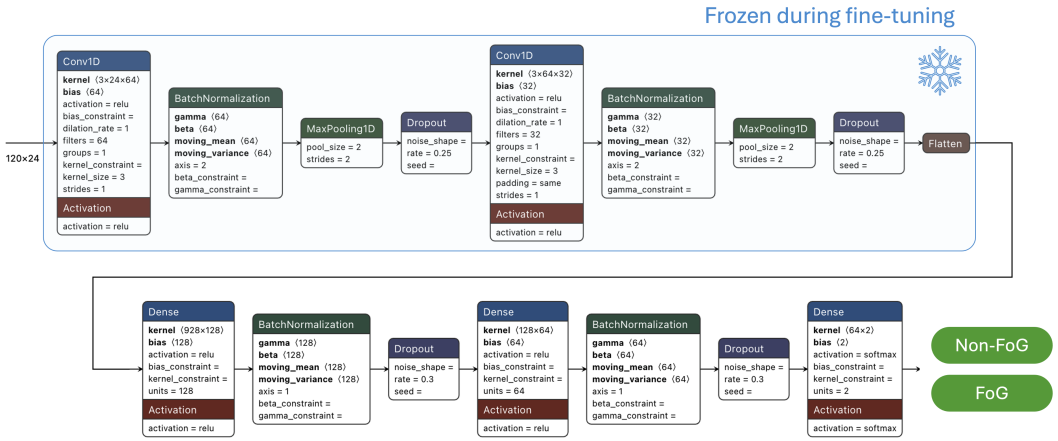


Figure 5.4. CNN best model architecture

red row in Table 5.4).

The CNN model's architecture, depicted in Figure 5.4, comprises a convolutional layer with 64 filters, utilizing a Rectified Linear Unit (ReLU) activation function. Like subsequent ones, this layer incorporates L2 regularization to restrain overfitting by penalizing large weights. The architecture progresses through stages of batch normalization to stabilize learning, max pooling to reduce feature map dimensions, and dropout to diminish the risk of overfitting by randomly omitting units during training. A second convolutional layer maintains the feature extraction process with 32 filters. The model employs two densely connected layers with neurons of 128 and 64. These layers serve as the classification portion of the network, with dropout and batch normalization interspersed to enhance model generalizability and mitigate overfitting. The output of the network is a dense layer that applies a softmax activation function, offering probability distributions over the two classes (Non-Fog, FoG).

The conventionally used CNN model achieves an average precision, recall, and F-1 score of 70.12 ± 8.65, 78.3 ± 3.97, and 66.56 ± 13.76, respectively, on the train set, while 84.24 ± 9.99, 82.68 ± 17.2 76.68 ± 20.2 respectively on the test set (i.e., red row in Table 5.4).

Generally, a slight reduction in testing performances is expected; however, in this case, the best model score in the testing phase is slightly higher than the training. On the other hand, the overall performance of the model could be more satisfactory. This is due to the fact that PD affects people differently and with different levels of severity, it is hard to create a model that is a *panacea* for everyone.

TW	O	Model	Train set			Test set			O	Train set			Test set		
			precision	recall	F1-score	precision	recall	F1-score		precision	recall	F1-score	precision	recall	F1-score
2 sec	0%	**CNN**	62.6± 0.4	69.3± 2.8	62.3± 0.2	61.3± 7.5	67.6± 10.1	60.9± 7.2	25%	67.0± 4.5	77.9± 5.5	65.6± 7.9	66.6± 8.9	79.5± 2.5	63.9± 10.1
		LSTM	62.0 ± 6.6	72.7 ± 10.1	66.7 ± 6.8	39.5 ± 14.0	51.3 ± 11.0	42.1 ± 7.1		60.3 ± 9.4	70.5 ± 4.6	64.7 ± 7.0	41.2 ± 16.5	51.9 ± 6.9	44.1 ± 11.9
		MLP	87.2 ± 4.7	93.8 ± 2.6	90.3 ± 2.3	45.5 ± 17.1	65.7 ± 15.3	49.6 ± 11.3		77.8 ± 16.5	94.1 ± 2.7	84.2 ± 11.3	50.6 ± 24.0	70.3 ± 18.3	52.0 ± 15.2
		RF	92.6 ± 3.7	99.9 ± 0.1	96.1 ± 2.0	61.3 ± 21.3	67.2 ± 11.2	61.1 ± 11.4		91.3 ± 3.6	99.9 ± 0.2	95.4 ± 1.9	61.8 ± 20.4	66.9 ± 14.7	60.6 ± 11.8
		LDA	97.9 ± 0.9	98.3 ± 0.8	98.1 ± 0.7	43.4 ± 17.8	64.9 ± 11.5	48.3 ± 10.1		94.3 ± 0.4	95.5 ± 0.5	94.9 ± 0.4	52.5 ± 18.9	64.3 ± 11.2	54.0 ± 8.1
		QDA	99.9 ± 0.2	100 ± 0.0	99.9 ± 0.1	52.2 ± 41.1	11.5 ± 13.1	17.0 ± 18.7		99.1 ± 1.1	99.4 ± 0.6	99.3 ± 0.8	37.8 ± 35.9	4.3 ± 4.1	6.1 ± 6.3
	50%	**CNN**	77.7± 10.1	85.0± 8.8	78.8± 11.3	66.5± 3.6	87.0± 12.7	69.8± 8.6	75%	70.1± 8.7	78.3± 4.0	66.6± 13.8	84.2± 10.0	82.7± 17.2	76.7± 20.2
		LSTM	64.3 ± 5.1	72.9 ± 8.0	68.0 ± 3.5	40.4 ± 15.4	52.3 ± 10.6	43.3 ± 8.8		70.3 ± 8.4	73.4 ± 7.3	71.5 ± 6.4	51.7 ± 18.2	57.7 ± 6.9	52.7 ± 10.5
		MLP	85.1 ± 3.0	95.5 ± 1.6	90.0 ± 2.4	49.7 ± 20.8	69.8 ± 13.5	53.2 ± 12.3		87.8 ± 2.3	94.9 ± 2.2	91.2 ± 1.1	54.8 ± 19.8	72.5 ± 12.0	58.2 ± 11.3
		RF	93.3 ± 5.4	99.9 ± 0.2	96.4 ± 2.9	61.8 ± 20.5	63.6 ± 12.1	60.2 ± 12.1		91.4 ± 5.8	99.9 ± 0.2	95.4 ± 3.2	61.4 ± 21.4	66.6 ± 11.3	61.3 ± 11.8
		LDA	93.1 ± 0.9	93.8 ± 1.0	93.4 ± 0.6	49.8 ± 17.2	64.9 ± 12.7	52.8 ± 8.6		91.5 ± 0.8	92.0 ± 1.6	91.7 ± 1.1	52.9 ± 20.1	59.7 ± 13.0	51.7 ± 9.8
		QDA	96.1 ± 3.4	98.6 ± 1.2	97.3 ± 2.2	17.2 ± 15.5	12.6 ± 14.5	12.4 ± 12.8		91.0 ± 4.2	97.1 ± 0.9	93.9 ± 2.2	43.4 ± 26.0	28.5 ± 21.3	27.3 ± 15.1
3 sec	0%	**CNN**	64.7± 1.6	70.6± 4.0	65.5± 1.0	58.0± 9.1	57.6± 10.2	56.7± 10.4	25%	62.7± 0.3	70.0± 0.8	62.4± 1.0	49.8± 1.3	56.9± 11.9	46.5± 2.6
		LSTM	61.9 ± 7.7	57.5 ± 16.9	57.5 ± 9.5	38.3 ± 23.3	35.0 ± 17.9	35.8 ± 19.3		59.0 ± 4.8	77.6 ± 4.0	66.8 ± 3.2	41.3 ± 17.8	53.3 ± 16.5	45.4 ± 16.0
		MLP	84.9 ± 4.9	94.5 ± 2.0	89.3 ± 3.0	50.3 ± 15.8	69.7 ± 11.0	57.0 ± 12.7		75.4 ± 23.9	95.4 ± 2.0	81.4 ± 19.0	48.5 ± 16.2	72.8 ± 11.4	55.3 ± 9.8
		RF	98.9 ± 2.2	100 ± 0.0	99.4 ± 1.1	61.6 ± 21.5	51.8 ± 12.1	53.9 ± 12.2		95.2 ± 3.9	99.9 ± 0.3	97.4 ± 2.0	63.8 ± 19.8	61.9 ± 14.0	59.9 ± 11.3
		LDA	100 ± 0.0	100 ± 0.0	100 ± 0.0	27.0 ± 15.5	39.3 ± 16.1	29.1 ± 13.3		99.1 ± 1.3	99.7 ± 0.3	99.4 ± 0.7	35.9 ± 18.7	57.8 ± 21.9	42.3 ± 17.4
		QDA	92.5 ± 11.3	100 ± 0.0	95.7 ± 6.6	33.0 ± 17.7	39.6 ± 20.7	35.8 ± 19.0		99.9 ± 0.3	100 ± 0.0	99.9 ± 0.1	21.6 ± 23.4	9.3 ± 9.6	13.0 ± 13.6

Table 5.4. Performance comparison on the six different models, using three time-windows (TW) size and four overlap (O). The models have been trained on a set of subjects and tested on a different set of subjects. Results are reported in % and averaged over the 7 folds (Part I).

TW	O	Model	Train set			Test set			O	Train set			Test set		
			precision	recall	F1-score	precision	recall	F1-score		precision	recall	F1-score	precision	recall	F1-score
3 sec	50%	CNN	62.6± 4.0	72.8± 0.8	60.5± 7.6	61.4± 12.1	63.6± 16.5	60.5± 14.2	75%	65.2± 4.1	75.7± 1.1	61.4± 3.0	56.3± 5.6	83.5± 2.6	53.4± 10.3
		LSTM	69.5 ± 8.0	73.3 ± 5.3	71.2 ± 6.2	46.9 ± 15.2	46.4 ± 10.1	45.5 ± 11.2		72.5 ± 5.1	83.3 ± 6.0	77.5 ± 4.8	42.2 ± 19.7	51.6 ± 5.4	43.7 ± 13.4
		MLP	85.3 ± 3.6	95.7 ± 1.2	90.2 ± 2.3	47.2 ± 18.8	66.3 ± 13.0	51.2 ± 12.4		87.4 ± 4.3	95.1 ± 3.4	91.0 ± 1.4	56.5 ± 16.3	73.8 ± 11.3	61.4 ± 10.6
		RF	96.6 ± 4.2	99.9 ± 0.2	98.2 ± 2.2	62.3 ± 19.9	56.3 ± 15.2	58.1 ± 15.3		95.4 ± 5.2	100 ± 0.0	97.6 ± 2.7	62.1 ± 21.2	54.6 ± 16.6	55.7 ± 12.5
		LDA	96.9 ± 0.7	96.6 ± 0.6	96.8 ± 0.4	42.8 ± 16.8	53.4 ± 5.9	45.5 ± 10.5		95.0 ± 0.2	95.5 ± 1.0	95.3 ± 0.6	51.9 ± 19.3	50.2 ± 13.2	47.9 ± 8.9
		QDA	99.7 ± 0.4	99.7 ± 0.3	99.7 ± 0.3	25.5 ± 38.0	1.3 ± 2.1	1.6 ± 2.4		98.0 ± 1.6	98.3 ± 1.4	98.2 ± 1.5	12.8 ± 18.9	7.7 ± 6.3	7.9 ± 8.2
4 sec	0%	CNN	59.3± 0.6	60.2± 0.4	59.7± 0.5	55.2± 5.8	53.0± 5.5	53.8± 5.3	25%	65.4± 3.6	69.9± 3.1	66.7± 3.9	59.1± 8.1	60.5± 9.1	59.1± 8.6
		LSTM	66.0 ± 11.5	65.0 ± 13.9	63.9 ± 6.8	34.8 ± 18.7	28.6 ± 18.0	30.0 ± 16.5		57.5 ± 12.1	70.3 ± 11.6	63.2 ± 11.9	35.1 ± 26.0	36.3 ± 16.8	33.1 ± 18.7
		MLP	89.4 ± 4.7	94.8 ± 4.1	91.9 ± 3.1	61.6 ± 14.8	68.5 ± 5.2	63.4 ± 7.2		81.8 ± 10.8	96.7 ± 0.7	88.3 ± 6.4	43.9 ± 17.9	75.6 ± 9.2	53.1 ± 15.4
		RF	99.2 ± 1.6	100 ± 0.0	99.6 ± 0.8	63.7 ± 19.2	52.9 ± 15.1	56.4 ± 14.3		98.8 ± 1.9	100 ± 0.0	99.4 ± 1.0	65.8 ± 19.8	53.4 ± 16.9	57.9 ± 15.9
		LDA	100 ± 0.0	100 ± 0.0	100 ± 0.0	36.3 ± 17.9	60.9 ± 9.9	42.5 ± 14.9		100 ± 0.0	100 ± 0.0	100 ± 0.0	31.4 ± 12.9	63.1 ± 10.9	39.2 ± 10.0
		QDA	84.5 ± 9.6	100 ± 0.0	91.3 ± 5.7	35.7 ± 14.3	63.6 ± 12.9	44.0 ± 14.2		93.3 ± 8.1	100 ± 0.0	96.3 ± 4.5	35.9 ± 17.7	54.6 ± 13.4	41.6 ± 16.4
	50%	CNN	66.1± 3.1	72.6± 3.5	67.4± 2.8	63.4± 2.9	67.5± 0.9	63.9± 2.4	75%	69.3± 5.0	80.3± 4.5	69.1± 5.3	70.3± 3.5	82.3± 2.3	70.6± 1.6
		LSTM	64.7 ± 8.1	67.2 ± 15.3	65.0 ± 9.4	36.2 ± 21.7	32.9 ± 19.3	33.4 ± 19.2		71.6 ± 8.8	79.6 ± 6.9	75.1 ± 6.8	46.6 ± 21.1	49.3 ± 13.9	45.7 ± 14.6
		MLP	83.1 ± 8.0	94.5 ± 3.2	88.2 ± 4.9	58.9 ± 17.8	70.8 ± 11.2	61.1 ± 9.2		91.2 ± 2.6	94.6 ± 2.8	92.8 ± 1.6	64.4 ± 14.7	67.4 ± 8.8	64.4 ± 8.2
		RF	95.9 ± 3.7	99.8 ± 0.2	97.8 ± 1.9	62.2 ± 21.2	56.2 ± 20.2	57.9 ± 18.2		94.2 ± 3.2	99.9 ± 0.1	96.9 ± 1.7	63.2 ± 20.1	64.3 ± 10.0	61.0 ± 8.7
		LDA	99.8 ± 0.3	99.7 ± 0.3	99.7 ± 0.2	35.7 ± 20.3	45.7 ± 13.7	37.8 ± 15.6		98.0 ± 0.4	97.4 ± 0.5	97.7 ± 0.2	51.7 ± 22.6	54.3 ± 3.8	49.9 ± 11.1
		QDA	100 ± 0.0	100 ± 0.0	100 ± 0.0	42.1 ± 42.3	1.8 ± 1.9	3.3 ± 3.6		99.8 ± 0.2	99.7 ± 0.6	99.7 ± 0.4	16.0 ± 32.0	1.0 ± 2.0	1.9 ± 3.7

Table 5.5. Performance comparison on the six different models, using three time-windows (TW) size and four overlap (O). The models have been trained on a set of subjects and tested on a different set of subjects. Results are reported in % and averaged over the 7 folds (Part II).

5.2.5.2 Enhancing FoG classification via model fine-tuning

To address this challenge and tailor the model to subject-specific characteristics, the fine-tuning method is exploited. Based on the results reported in 5.4, the CNN with the 2-second time window (75% overlap) is the most effective model and structurally suitable for applying the fine-tuning approach mentioned in 5.2.4. Thus, to analyze the proposed approach performances, the fine-tuning approach was applied as follows.

Starting from the CNN models trained using the k-fold technique, discussed in 5.2.3, the performance of such models was tested on each corresponding test subject. Subsequently, for each test subject, fine-tuning was performed on the model, thus tailoring the model for each subject. During the fine-tuning process, the convolutional layers, as depicted in 5.4, were frozen to retain the generic patterns learned by these layers and specialize the final part of the network. To do such a process, 70% of the subject's data was employed for tuning the model and the remaining 30% for testing the performance of the fine-tuned model.

As presented in 5.6 (i.e., Fine-tuned block), the results indicate that for 11 out of 13 subjects, the fine-tuned model significantly improved precision, recall, and f1-score compared to the not fine-tuned model. These results emphasize the value of fine-tuning as a technique to enhance the performance of machine learning models. Firstly, it enables a pre-trained model on an existing dataset to be fine-tuned on newly collected data from a different subject, eliminating the need to gather an entirely new dataset from multiple subjects. Secondly, as demonstrated in this specific investigation, it results in an average precision increase of 14.4%, recall increase of 17.4%, and a 20.9% improvement in F1-score. The high variability in the pre-fine-tuning performances suggests that the model struggles to generalize, although this is expected. Since FoG episodes are complex and multi-factorial, it is acknowledged that the dataset may not capture all the facets of the FoG phenomenon. In future developments, expanding the data with more participants and time collections can be evaluated to achieve a more generalized, comprehensive model.

5.3 Positional Mixture-of-Experts for FoG detection

5.3.1 Datasets Included in the Study

Seven publicly available datasets containing inertial recordings of Parkinson's disease patients with annotated Freezing of Gait (FOG) episodes were included in this

Subject	Pre-fine-tuning			Fine-tuned		
	precision	recall	F1-score	precision	recall	F1-score
S1	80.3	73.4	73.6	87.1	79.9	80.5
S2	63.3	51.2	38.8	92.6	90.9	91.3
S3	63.3	66.9	55.2	78.0	86.1	79.7
S4	51.7	81.5	42.0	62.5	98.0	69.0
S5	83.3	99.2	89.6	100.0	100.0	100.0
S6	100.0	100.0	100.0	100.0	100.0	100.0
S7	98.1	99.4	98.7	100.0	100.0	100.0
S8	47.5	29.5	36.4	52.6	64.9	50.3
S9	45.0	43.8	44.1	94.1	86.0	89.3
S10	58.0	50.8	32.0	92.6	93.9	92.9
S11	55.1	69.9	54.3	54.9	74.6	50.4
S12	57.1	78.6	48.9	75.0	96.4	81.5
S13	100.0	100.0	100.0	100.0	100.0	100.0
Mean	69.4	72.6	62.6	83.8	90.1	83.5
Std	19.5	22.7	25.1	16.9	10.8	16.9

Table 5.6. Results of the CNN model on the test subjects, before and after fine-tuning.

study: *Daphnet*, *IMU*, *FOGSTAR*, *TDCS*, *Multimodal*, *Oday*, and *Rempark*. A detailed description of these datasets was provided in Chapter 4.

These datasets differ substantially in several aspects, including sensor placement, number and type of wearable devices, experimental protocols, and data annotation prodecures. Such variability reflects the diversity of real-world sensing configurations used in Parkinson's disease studies.

By integrating datasets with different sensor configurations and recording conditions, the proposed framework aims to assess the ability of the model to generalize beyond the characteristics of a single dataset. This is particularly relevant for Freezing of Gait detection, where models trained on a single dataset often exhibit limited transferability due to differences in sensor modalities, patient populations, and experimental protocols.

For this reason, all datasets were harmonized through a common preprocessing pipeline and used within a multi-dataset learning framework combined with a Leave-

One-Dataset-Out evaluation protocol, allowing systematic assessment of cross-dataset generalization.

5.3.2 Preprocessing

The datasets considered in this study were collected using heterogeneous acquisition protocols, sensor configurations, and sampling frequencies. Consequently, a preprocessing pipeline was implemented to harmonize the signals and ensure consistency across datasets prior to model training.

5.3.2.1 Datasets Harmonization

First, the raw datasets were restructured into a unified tabular format. Each recording was organized according to the hierarchical identifiers *Patient*, *Session*, and *Task*, allowing the preservation of experimental context and preventing temporal mixing between different activities or acquisition sessions. Dataset-specific metadata and inconsistent column naming conventions were standardized to ensure a common representation of sensor modalities.

Second, inertial signals were cleaned to remove invalid or missing values. Non-numeric entries and corrupted samples were filtered, and columns corresponding to unavailable sensors were handled consistently across datasets. To avoid collisions between subjects originating from different datasets, patient identifiers were remapped to globally unique integer IDs obtained by combining the dataset identifier with the original subject identifier.

Third, signals were resampled to a common sampling frequency of 40 Hz. Because the datasets were originally recorded at different rates, resampling was necessary to ensure temporal consistency across all recordings and to enable the use of a uniform segmentation strategy during model training. Linear interpolation was used to resample the inertial time series while preserving signal continuity [139].

Following this, each dataset was represented using the same structural format and sampling rate, enabling integration within the subsequent multi-dataset learning framework.

5.3.2.2 Vertical Component Extraction and Horizontal Magnitude Representation

Because the datasets included in this study were collected using heterogeneous sensor configurations and did not always provide explicit information about sensor orientation, a preprocessing step was introduced to derive a consistent representation of the inertial signals across datasets.

The vertical axis of each sensor was estimated automatically by exploiting the gravitational component present in accelerometer measurements. Two complementary strategies were employed depending on the availability of gyroscope data.

When gyroscope signals were available, quasi-static windows were identified by scanning the recordings for short temporal segments characterized by low angular velocity and an acceleration magnitude close to gravitational acceleration. These conditions indicate periods in which the sensor is approximately stationary and the measured acceleration is dominated by gravity. Within such windows, the mean acceleration along the three axes was computed, and the axis exhibiting the largest absolute mean value was selected as the vertical axis [140, 141].

When gyroscope signals were not available or quasi-static windows could not be reliably identified, the gravitational component was estimated using a low-pass filtering approach applied to the accelerometer signals. The filtered signal provided an estimate of the gravity vector, and the axis most consistently aligned with this component over time was selected as the vertical direction [140].

Once the vertical axis had been identified, the original triaxial inertial signals were transformed into a representation that is less sensitive to sensor orientation differences. Specifically, each triaxial signal was collapsed into two derived components. The vertical component was obtained by taking the absolute value of the acceleration or angular velocity measured along the identified vertical axis. The horizontal motion was instead summarized by computing the Euclidean magnitude of the two remaining orthogonal axes:

$$x_{\text{mag}}(t) = \sqrt{x_{h1}(t)^2 + x_{h2}(t)^2}.$$

This transformation was applied consistently to both accelerometer and gyroscope signals whenever the three axes were available. As a result, each triaxial inertial signal was reduced from three channels to two features: a rectified vertical component and a horizontal magnitude.

The resulting representation offers two main advantages in the context of multi-

dataset learning. First, it reduces the sensitivity of the input features to dataset-specific sensor orientations and mounting conventions, which is critical when combining heterogeneous experiments. Second, it reduces the dimensionality of the input space while preserving biomechanically meaningful information related to gait dynamics. In particular, the vertical component of acceleration reflects step impacts and rhythmic locomotor patterns, which are known to be informative for gait analysis and Freezing of Gait detection [49, 57, 142]. The derived signals were subsequently used as input to the learning models described in the following sections.

5.3.2.3 Bilaterality Management and Sensor Position Assimilation

The datasets included in this study exhibit substantial heterogeneity in terms of sensor placement and body coverage. Some datasets provide bilateral measurements (e.g., left and right ankle sensors), whereas others include sensors positioned on different anatomical locations. To ensure consistency across datasets and reduce redundancy, a bilaterality management and sensor position assimilation strategy was applied.

When bilateral sensors were available, a side-selection procedure was applied to retain a single representative signal for each bilateral sensor position. Human gait exhibits a largely symmetric and periodic structure during normal locomotion, resulting in highly correlated motion patterns between the left and right sides of the body [143]. This bilateral symmetry is commonly observed in wearable sensor recordings, where signals from corresponding segments (e.g., left and right limbs) often contain redundant information [144]. Consequently, retaining both sides would increase the dimensionality of the input space without providing substantial additional discriminative information, while also increasing computational cost and model complexity. Reducing redundant features is a common strategy in machine learning to improve efficiency and generalization [123]. Rather than selecting a side arbitrarily, the choice was determined through a data-driven comparison between the left and right signals within each dataset.

For each bilateral sensor stem, inertial signals were first segmented into fixed-length windows of 80 samples, corresponding to a temporal duration of 2 s at the common sampling frequency of 40 Hz. A sliding-window strategy with 50% overlap was used, and windowing was performed independently within each *Patient–Session–Task* segment to avoid temporal mixing between different recording contexts. Each window was assigned a label based on the majority Freezing of Gait (FoG) annotation within the window, with ties resolved in favor of the non-FoG class.

For both the left and right sides, a compact feature representation was then extracted from the transformed inertial signals. Specifically, simple descriptive statistics were computed from the vertical and horizontal-magnitude components of the accelerometer signals and, when available, the gyroscope signals. The resulting feature vector for each window included the mean and standard deviation of these signals:

$$[\mu(a_v), \sigma(a_v), \mu(a_{\text{mag}}), \sigma(a_{\text{mag}}), \mu(g_v), \sigma(g_v), \mu(g_{\text{mag}}), \sigma(g_{\text{mag}})].$$

A linear classifier was trained independently for the left and right sides and evaluated using stratified 5-fold cross-validation with class-balanced sample weighting. The primary performance metric used to compare the two sides was the area under the precision–recall curve (AUPRC), which is more informative in the presence of class imbalance typical of FoG detection tasks [145]. For each bilateral sensor position, the side achieving the highest cross-validated performance was retained, provided that a sufficient number of valid windows were available. In cases where both sides achieved comparable performance, the side containing a larger number of usable windows was preferred; if the tie persisted, the left side was selected by convention.

Regression-based approaches have previously been used in gait analysis to evaluate the contribution of individual sensor signals and identify redundant measurements, enabling the definition of minimal yet informative sensor configurations [146]. Consistent with this principle, the linear model employed here provides a lightweight and interpretable framework for comparing the predictive contribution of bilateral signals. Although more sophisticated feature selection techniques exist, the objective of this step is to identify redundant measurements rather than to construct the final predictive model. A simple linear approach therefore provides a computationally efficient and sufficiently informative solution for side selection.

This side-selection strategy ensures that redundant bilateral signals are removed while preserving the side that provides the most informative representation for FoG discrimination within each dataset, while simultaneously reducing the dimensionality of the input space and the computational cost of subsequent experiments.

In addition to bilaterality management, a sensor position assimilation step was introduced to reconcile differences in sensor placement across datasets. Sensors positioned on slightly different anatomical locations can capture similar biomechanical characteristics of gait. For example, sensors placed on the shank and ankle both measure distal lower-limb motion, while sensors located on the arm and wrist capture upper-limb dynamics.

To harmonize these configurations, sensor positions were mapped to a common set of canonical locations based on their biomechanical similarity. The mapping adopted in this study is summarized in Table 5.7.

Original Position	**Assimilated Position**
Shank	Ankle
Arm	Wrist
Waist	Lower back

Table 5.7. Sensor position assimilation used to harmonize heterogeneous datasets based on biomechanical similarity.

By combining bilaterality management with sensor position assimilation, the preprocessing pipeline reduces redundancy and harmonizes heterogeneous sensor configurations. This step facilitates the integration of multiple datasets within the proposed multi-dataset learning framework while maintaining a coherent interpretation of the underlying gait dynamics.

5.3.2.4 Building Samples Around FoG Dynamics

Several methodological choices adopted throughout this thesis were motivated by both signal-processing considerations and the temporal characteristics of Freezing of Gait. Window-based segmentation was selected because FoG manifests as a transient phenomenon that unfolds over short temporal intervals, requiring sufficient contextual information to capture both normal gait dynamics and freezing-related alterations. The use of overlapping windows was intended to reduce boundary effects and improve the representation of episodes occurring near window transitions, thereby increasing temporal continuity in the extracted features. Resampling and sensor harmonization procedures were introduced to address the substantial heterogeneity observed across publicly available datasets. Differences in sampling frequencies, sensor hardware, and acquisition protocols can introduce dataset-specific biases that hinder cross-dataset learning. Standardizing these characteristics allows the models to focus on movement-related patterns rather than acquisition-related variability. Similarly, the harmonization of sensor locations was motivated by the biomechanical nature of FoG. Although freezing episodes involve whole-body motor control, the manifestation of FoG-related kinematic signatures differs across anatomical locations. The adopted harmonization strategy aimed to preserve the most relevant locomotor information while enabling comparisons across datasets with different

sensor configurations. Finally, thresholding and decision criteria were chosen to balance sensitivity and specificity in a highly imbalanced detection problem. Since FoG episodes are relatively infrequent compared with normal gait, threshold selection directly influences the trade-off between missed events and false alarms. Consequently, these methodological decisions should be viewed not merely as technical preprocessing steps but as mechanisms for adapting the analysis pipeline to the physiological, temporal, and biomechanical characteristics of Freezing of Gait.

5.3.3 Mixture-of-Expert Framework

5.3.3.1 Multi-dataset split (Leave-One-Dataset-Out protocol)

The Mixture-of-Expert framework was implemented within a multi-dataset evaluation setting using a Leave-One-Dataset-Out (LODO) protocol. Given the set of available datasets, one dataset was iteratively designated as the target evaluation dataset, while the remaining datasets were combined to form the training pool. This strategy ensures that the model is trained exclusively on data originating from different experimental studies than the one used for testing, thereby providing a rigorous assessment of cross-dataset generalization. During each iteration of the protocol, the datasets included in the training pool were merged into a single training corpus while preserving metadata describing the originating dataset, subject identifier, and acquisition context. The left-out dataset was kept completely isolated during model training and used only during the final evaluation stage.

5.3.3.2 Window segmentation

Continuous inertial recordings were converted into fixed-length temporal windows in order to obtain a representation suitable for supervised learning. Segmentation was performed using a sliding-window strategy applied to the resampled inertial signals at the common sampling frequency of 40 Hz. Each window contained $L = 80$ samples, corresponding to a temporal duration of 2 s, and consecutive windows were generated with 50% overlap, resulting in a stride of 40 samples. To preserve the experimental structure of the recordings, segmentation was performed independently within each *Patient–Session–Task* segment, ensuring that windows never crossed boundaries between different subjects, sessions, or motor tasks.

5.4 Multiple-Expert framework for cross-dataset FoG detection

For each valid segment of length T_g, the number of windows was determined by the standard sliding-window relation

$$N_g = 1 + \left\lfloor \frac{T_g - L}{S} \right\rfloor,$$

where $S = L(1 - \text{overlap})$ denotes the stride. Segments shorter than the required window length were discarded. Each window was assigned a binary label using majority voting over the frame-level Freezing of Gait (FoG) annotations contained within the window. Let $y_t \in \{0, 1\}$ denote the sample-level FoG annotation. The window label was defined as

$$y = \mathbb{I}\left(\frac{1}{L} \sum_{t=1}^{L} y_t \geq 0.5 \right),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Consequently, a window was labeled as FoG whenever at least half of its samples corresponded to a FoG state. Exact ties were resolved in favor of the positive class in order to preserve windows containing transitional FoG events. This procedure yielded a collection of labeled signal windows together with their associated subject identifiers, dataset identifiers, and contextual metadata, which were subsequently used for expert construction and model training.

5.4.1 Expert definition from sensor configurations

Following segmentation, a set of experts was defined to explicitly model the heterogeneous sensor configurations present across datasets. Each expert corresponds to a specific subset of sensors associated with a given body location and signal modality, and is therefore responsible for processing a particular view of the inertial signals. Let $\mathcal{E} = \{e_1, \ldots, e_E\}$ denote the set of experts defined in the framework. Each expert e is characterized by the subset of input channels required for its operation, typically consisting of the vertical component and horizontal magnitude derived from the accelerometer signals and, when available, the corresponding components of the gyroscope signals for that sensor location.

For every segmented window, the preprocessed signals were reorganized according to the sensor subsets associated with the experts. This resulted in an expert-oriented representation in which the input data are grouped by sensor

configuration rather than by their original column ordering. The resulting tensor representation can be expressed as

$$X \in \mathbb{R}^{N \times E \times T \times C},$$

where N denotes the number of segmented windows, E the number of experts, T the temporal length of each window, and C the number of signal channels associated with the sensor subset processed by each expert. Each expert therefore receives a structured view of the input window restricted to the channels corresponding to its sensor configuration. When a dataset does not provide the channels required by a particular expert, the corresponding entries are temporarily filled with zeros and later excluded through dataset-dependent masking during the expert activation stage. This expert-oriented representation allows heterogeneous sensor configurations to be integrated within a single model while preserving the temporal dynamics of the inertial signals.

5.4.2 Expert availability modeling

Because the datasets included in this study provide different combinations of sensors, not all experts can operate on every dataset. To explicitly encode this constraint, an expert availability model was introduced to determine which experts are compatible with the signals provided by each dataset. Let D denote the number of datasets and E the number of defined experts. For each dataset $d \in \{1, \dots, D\}$ and expert $e \in \{1, \dots, E\}$, a binary compatibility indicator $A_{d,e} \in \{0, 1\}$ was defined, where $A_{d,e} = 1$ if dataset d provides all sensor channels required by expert e, and $A_{d,e} = 0$ otherwise. The resulting matrix

$$A \in \{0, 1\}^{D \times E}$$

summarizes the compatibility between the datasets and the available experts. This representation captures the heterogeneous sensing configurations of the datasets and determines which experts can meaningfully process the signals originating from each dataset.

5.4.3 MoE input tensor construction

Following the definition of the expert set and their availability across datasets, the segmented windows were reorganized into a unified tensor representation compatible

with the Mixture-of-Expert architecture. For each window n, the signals corresponding to the sensor subset required by each expert e were extracted and arranged in a consistent temporal structure. The resulting tensor representation can be expressed as

$$X \in \mathbb{R}^{N \times E \times T \times C},$$

where N denotes the number of segmented windows, E the number of experts, T the temporal length of each window, and C the number of channels associated with the sensor configuration processed by each expert. When a dataset does not provide the channels required by a particular expert, the corresponding tensor entries are temporarily filled with zeros. This unified tensor construction enables the model to process heterogeneous sensor configurations within a common representation while preserving the temporal dynamics of the inertial signals.

5.4.4 Dataset-dependent expert masking

To prevent experts from processing signals originating from unsupported sensor configurations, a dataset-dependent masking mechanism was introduced. The dataset-level availability matrix A was propagated to the window level to construct an activation mask

$$M \in \{0, 1\}^{N \times E},$$

where each element $M_{n,e} = A_{d(n),e}$ indicates whether expert e is valid for window n, and $d(n)$ denotes the dataset associated with that window. During model training and inference, this mask is applied to the expert outputs to ensure that only experts supported by the sensor configuration of the originating dataset contribute to the prediction. Experts corresponding to unavailable sensors are therefore automatically deactivated. This masking strategy guarantees that the mixture-of-experts model remains consistent with the sensing constraints imposed by each dataset while allowing a unified architecture to operate across heterogeneous experimental setups.

5.4.5 Expert activity feature computation

In addition to the raw expert inputs, auxiliary activity statistics were computed to characterize the overall signal intensity associated with each expert for a given input window. These statistics were derived directly from the expert-oriented input tensor

$$X_{\text{moe}} \in \mathbb{R}^{N \times E \times T \times 2},$$

where the last dimension corresponds to the vertical component and the horizontal magnitude representation of the inertial signals associated with the sensor subset processed by each expert.

For each window n and expert e, the signals were first transformed into their absolute values and flattened across the temporal and channel dimensions. Let $x_{n,e,t,c}$ denote the value of the signal at time index t and channel c, with $c \in \{v, \text{mag}\}$. The activity statistics were then computed as the mean and standard deviation of the absolute signal values over the entire window:

$$\mu_{n,e} = \frac{1}{2T} \sum_{t=1}^{T} \sum_{c} |x_{n,e,t,c}|, \qquad \sigma_{n,e} = \sqrt{\frac{1}{2T} \sum_{t=1}^{T} \sum_{c} \left(|x_{n,e,t,c}| - \mu_{n,e}\right)^2}.$$

The resulting activity representation can therefore be written as

$$A^{\text{act}} \in \mathbb{R}^{N \times E \times 2},$$

where the last dimension contains the mean absolute signal value and its standard deviation for each expert. These features provide a compact summary of the magnitude and variability of the signals available to each expert and are later used to modulate expert contributions during the gating stage. By incorporating these activity statistics, the framework can adapt the influence of each expert according to the amount of motion information present in its input signals.

5.4.6 Gate activation

Expert activation was determined using a gating mechanism that combines dataset-dependent availability constraints with prior weights associated with each expert. For each input window n, the binary mask $M_{n,e}$ defined in the previous step specifies whether expert e is compatible with the dataset from which the window originates. Experts for which $M_{n,e} = 0$ are automatically deactivated and excluded from the mixture. For the remaining valid experts, prior weights π_e were introduced to reflect the relative prevalence of each expert configuration within the training datasets. These priors were estimated from the distribution of available windows across datasets and incorporated in logarithmic form to stabilize numerical computations. The final gate activation therefore depends on the joint effect of the availability mask, the expert priors, and the activity statistics computed from the expert inputs. This mechanism determines the subset of experts contributing to the prediction for each window while ensuring that only experts supported by the corresponding sensor configuration are considered.

5.4.7 Mixture-of-Experts architecture

The proposed framework is implemented as a neural Mixture-of-Experts (MoE) model composed of three main components: a set of expert networks, a gating network, and a prediction head. Each expert is implemented as a lightweight one-dimensional convolutional neural network that processes the temporal signals associated with its corresponding sensor configuration 5.8. Specifically, each expert applies a sequence of convolutional layers with decreasing kernel sizes to capture local temporal patterns at multiple scales, followed by a global average pooling operation that aggregates the temporal information into a fixed-length embedding. A dropout layer is applied to the pooled representation to reduce overfitting, and a final dense projection maps the extracted features into a shared latent space of dimension d_{embed}. This design enables each expert to learn a compact representation of the motion dynamics captured by its associated sensor subset.

Layer	**Type**	**Configuration**	**Output shape**
Input	–	$(T, 2)$	$(T, 2)$
1	Conv1D	filters=64, kernel=7, ReLU	$(T, 64)$
2	Conv1D	filters=64, kernel=5, ReLU	$(T, 64)$
3	Conv1D	filters=64, kernel=3, ReLU	$(T, 64)$
4	GlobalAvgPool	–	(64)
5	Dropout	$p = 0.1$	(64)
6	Dense	units=d_{embed}, ReLU	(d_{embed})

Table 5.8. Architecture of the ExpertCNN module.

The gating network is responsible for dynamically weighting the contribution of each expert for a given input window. It operates on a set of expert-level features that include the expert availability mask, the logarithm of prior expert weights, the activity statistics derived from the input signals, and, optionally, a learned embedding representing the dataset of origin 5.9. These features are concatenated and processed through a couple of dense layers to produce a set of logits over the experts. A masked softmax operation is then applied to obtain normalized gating coefficients, ensuring that only experts compatible with the input dataset contribute to the mixture. This mechanism allows the model to adaptively combine expert outputs based on both signal characteristics and dataset-specific context.

The final prediction head operates on the aggregated expert representation, computed as the weighted sum of the expert embeddings according to the gating

Component	Type	Configuration	Output shape
Inputs	–	log priors, mask, activity, dataset ID	(E, F)
1	Dataset Embedding	$d_{dataset} = 16$	$(E, 16)$
2	Concatenation	all features	(E, F')
3	Dense	units=64, ReLU	$(E, 64)$
4	Dense	units=64, ReLU	$(E, 64)$
5	Dense	units=1	$(E, 1)$
6	Masking	invalid experts $\rightarrow -\infty$	(E)
7	Softmax	across experts	(E)

Table 5.9. Architecture of the gating network (GateNet).

coefficients 5.10. This combined representation is passed through a small fully connected network with dropout regularization to produce the final output. The inclusion of dropout in both the expert networks and the prediction head helps mitigate overfitting, particularly given the variability across datasets and subjects. Overall, this architecture enables the integration of heterogeneous sensor configurations within a unified framework, while maintaining flexibility through adaptive expert selection and robust feature learning.

Layer	Type	Configuration	Output shape
Input	–	aggregated embedding $\sum_e \alpha_e z_e$	(d_{embed})
1	Dropout	$p = 0.1$	(d_{embed})
2	Dense	units=64, ReLU	(64)
3	Dropout	$p = 0.1$	(64)
4	Dense	units=1	(1)

Table 5.10. Architecture of the prediction head.

5.4.8 Patient-wise cross-validation

To evaluate model performance while preventing information leakage between training and validation data, subject-independent cross-validation folds were generated within the combined training datasets. Let S denote the set of subjects present in the development pool. The segmentation windows were grouped according to their

corresponding subject identifiers, and a stratified K-fold procedure was applied at the subject level so that all windows originating from the same subject were assigned to the same fold. This strategy ensures that no subject contributes data simultaneously to both training and validation subsets. During model development, each fold was used in turn as the validation set while the remaining folds formed the training subset. This patient-wise partitioning provides a realistic estimate of the model's ability to generalize to unseen individuals while preserving the class distribution across folds.

5.4.9 MoE training

The Mixture-of-Expert model was trained on the combined development pool obtained at each Leave-One-Dataset-Out iteration. The architecture consists of a set of expert-specific temporal encoders and a gating module that computes expert weights for each input window. Concretely, each expert receives an input tensor of shape $T \times 2$, corresponding to the vertical and horizontal-magnitude signals over the temporal window, and processes it through a dedicated one-dimensional convolutional encoder. Each expert encoder is implemented as a small temporal convolutional network composed of three successive 1-D convolutional layers with kernel sizes 7, 5, and 3, respectively, followed by global average pooling, dropout, and a dense projection layer producing a latent embedding of dimension d_{embed}. Let $z_{n,e} \in \mathbb{R}^{d_{\text{embed}}}$ denote the embedding generated by expert e for window n. Stacking the outputs of all experts yields a latent representation

$$Z \in \mathbb{R}^{N \times E \times d_{\text{embed}}}.$$

The gate receives, for each expert, the dataset-dependent availability mask, the expert prior weight in logarithmic form, the activity statistics described above, and optionally a learned embedding of the dataset identity. These quantities are concatenated and processed through a small multilayer perceptron composed of two fully connected layers with ReLU activation, followed by a linear output layer producing one scalar gate logit for each expert. Experts unavailable for a given window are forcibly deactivated by assigning a large negative value to their logits before normalization. The final expert weights are then obtained by applying a softmax over the valid experts, yielding coefficients

$$\alpha_{n,e} = \frac{\exp(g_{n,e})}{\sum_{e' \in \mathcal{E}_n} \exp(g_{n,e'})},$$

where $g_{n,e}$ denotes the gate logit for expert e and $\mathcal{E}_n$ is the set of experts active for window n. The mixture representation is computed as the weighted sum of expert embeddings,

$$m_n = \sum_{e=1}^{E} \alpha_{n,e} z_{n,e},$$

and is subsequently processed by a prediction head composed of dropout, a dense hidden layer with 64 units and ReLU activation, a second dropout layer, and a final linear output neuron producing the FoG logit.

Training was performed independently on each patient-wise fold of the combined development pool. The model was optimized using the Adam optimizer with gradient clipping (clipnorm $= 1.0$) and binary cross-entropy loss applied to the raw logits. To mitigate class imbalance, class-balanced sample weighting was used, with weights inversely proportional to the frequency of the FoG and non-FoG classes in the training partition. Model selection was based on the area under the precision–recall curve (AUPRC) computed on the validation fold, which was preferred to AUROC because of the highly imbalanced nature of FoG detection. Early stopping was implemented by monitoring validation AUPRC and terminating training when no improvement was observed for a fixed number of epochs. For each fold, the model parameters corresponding to the best validation AUPRC were retained, and the resulting trained model, training history, and validation metrics were saved to disk.

5.4.10 Test dataset preprocessing

Once training on the combined train datasets was completed, the held-out dataset was processed using exactly the same pipeline applied to the training data. First, the continuous inertial recordings of the left-out dataset were segmented into fixed-length windows of 80 samples with 50% overlap, preserving the *Patient–Session–Task* structure and assigning window labels through majority voting on the frame-level FoG annotations. The segmented windows were then reorganized into the same expert-oriented tensor representation used during training, ensuring that the test inputs had shape

$$X_{\text{test}} \in \mathbb{R}^{N_{\text{test}} \times E \times T \times 2}.$$

This step is critical because the model expects the same expert-wise arrangement of vertical and horizontal-magnitude signals used during training.

In parallel, the dataset-dependent expert availability mask for the test windows was constructed by indexing the dataset–expert compatibility matrix with the identifiers

of the windows belonging to the held-out dataset. This yielded a binary test mask

$$M_{\text{test}} \in \{0,1\}^{N_{\text{test}} \times E},$$

which specifies which experts are allowed to contribute to each test window. Activity statistics were also computed for the test expert inputs using the same procedure adopted during training, namely by taking the mean absolute value and standard deviation of the expert signals over the temporal and channel dimensions. As a result, the held-out dataset was transformed into a complete MoE-compatible representation composed of expert inputs, activity features, dataset identifiers, and expert-availability masks, all generated through the same preprocessing steps used for the source datasets.

5.4.11 MoE inference on unseen dataset

Inference on the unseen dataset was performed using the fold-specific models obtained during training. For each saved fold model, the MoE architecture was reloaded and applied to the expert-formatted test windows together with the corresponding test masks, activity statistics, dataset identifiers, and expert prior weights. The forward pass followed the same procedure used during training: each expert independently produced an embedding from its assigned input channels, the gate computed expert weights conditioned on the valid experts and auxiliary gating features, and the weighted mixture of expert embeddings was fed to the prediction head to produce a scalar logit for each test window. The final FoG probability was then obtained by applying the sigmoid function to the output logit.

To convert probabilities into binary predictions, the decision threshold was not fixed a priori to 0.5, but instead inherited from the corresponding validation fold. More specifically, for each trained fold model, the optimal threshold was selected on the validation partition as the threshold maximizing the F_1-score on the precision–recall curve. During inference, this fold-specific threshold was applied to the probabilities predicted on the held-out dataset:

$$\hat{y}_n = \begin{cases} 1, & \text{if } p_n > \tau_k, \\ 0, & \text{otherwise,} \end{cases}$$

where p_n denotes the predicted probability for test window n and τ_k is the threshold selected from validation fold k. This design preserves consistency between model selection and deployment, and avoids imposing a threshold that may be suboptimal under severe class imbalance.

5.4.12 Cross-dataset evaluation

The final evaluation was performed on the left-out dataset and therefore quantified the ability of the proposed framework to generalize across heterogeneous experimental domains. Since a separate model was trained for each patient-wise fold of the combined source datasets, the held-out dataset was evaluated independently with each fold-specific model. For each model, standard binary classification metrics were computed on the window-level predictions, including the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), the F_1-score, balanced accuracy, and Matthews correlation coefficient (MCC). In addition, the number of test windows and the proportion of FoG windows in the held-out dataset were recorded in order to contextualize performance with respect to class imbalance.

The per-fold test results were stored separately, yielding a distribution of performance estimates for each left-out dataset rather than a single point estimate. This choice provides a more robust view of generalization performance because it reflects the variability induced by the patient-wise validation splits used during source-domain training. Repeating the entire procedure for each dataset in the Leave-One-Dataset-Out protocol produced a complete cross-dataset evaluation of the Multiple-Expert framework, allowing the assessment of how well the learned expert representations and gating mechanism transfer to unseen sensor configurations and patient cohorts.

5.4.13 Experimental results

5.4.13.1 Effect of resampling

To ensure temporal consistency across datasets, all inertial recordings were resampled to a common frequency. Figure 5.5 shows a representative example from the FOGSTAR dataset (patient 107, session 1, task 40), comparing the original and resampled signals. The resampled signal closely follows the original trajectory, preserving both amplitude and temporal dynamics while reducing the number of samples. This qualitative observation indicates that the interpolation procedure does not introduce visible distortions in the signal structure.

To quantitatively evaluate the impact of resampling, statistical similarity metrics were computed between the original and resampled signals across all datasets (Table 5.11). Specifically, the Mean Absolute Error (MAE), the Root Mean Squared

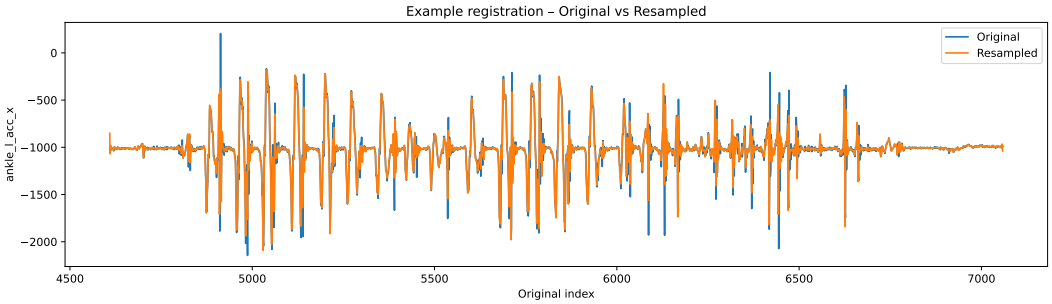


Figure 5.5. Example of inertial signal before and after resampling (FOGSTAR dataset, patient 107). The resampled signal preserves the temporal structure and amplitude of the original recording.

Error (RMSE), and the Pearson correlation coefficient are reported. MAE and RMSE quantify the absolute deviation between the original and resampled signals, with RMSE placing greater emphasis on larger deviations [123], while the Pearson correlation coefficient measures the linear relationship between the two signals and is particularly informative for assessing the preservation of temporal structure and signal morphology [147].

Dataset	MAE	RMSE	Pearson
FOGSTAR	0.93	4.05	0.985
IMU	2.55	11.60	0.947
Multimodal	75.14	351.99	0.996
Rempark	0.21	0.48	0.987
Daphnet	35.95	131.09	0.941

Table 5.11. Statistical comparison between original and resampled signals across datasets.

Across all datasets, the Pearson correlation coefficient remains consistently high, ranging from 0.941 (Daphnet) to 0.996 (Multimodal). This indicates that the resampling procedure preserves the temporal structure and overall signal morphology with high fidelity, which is particularly important for downstream time-series modeling tasks. In contrast, the MAE and RMSE values exhibit larger variability across datasets. This behaviour is expected, as these metrics are sensitive to the amplitude and scaling of the signals, which differ substantially between datasets due to variations in sensor specifications, preprocessing pipelines, and measurement units.

For instance, datasets such as Rempark and FOGSTAR, which present lower signal magnitudes, exhibit correspondingly low MAE and RMSE values, whereas datasets such as Multimodal and Daphnet show larger absolute errors due to higher signal

amplitudes. Importantly, these higher error values do not indicate a degradation of signal quality, as confirmed by the consistently high correlation coefficients. Instead, they reflect scale-dependent differences that do not affect the relative temporal patterns captured by the signals.

Overall, the combined analysis of error-based and correlation-based metrics provides a comprehensive assessment of the resampling procedure. While MAE and RMSE quantify absolute deviations, the high Pearson correlation demonstrates that the essential temporal dynamics are preserved. This confirms that resampling effectively harmonizes heterogeneous datasets without compromising the information content relevant for subsequent segmentation and modeling stages.

5.4.13.2 FoG segmentation and class distribution

Following the segmentation procedure, all time-series recordings were partitioned into fixed-length windows and labeled according to the presence or absence of FoG events. Figure 5.6 provides an overview of the number of FoG and non-FoG windows across the considered datasets, highlighting the class distribution obtained after preprocessing.

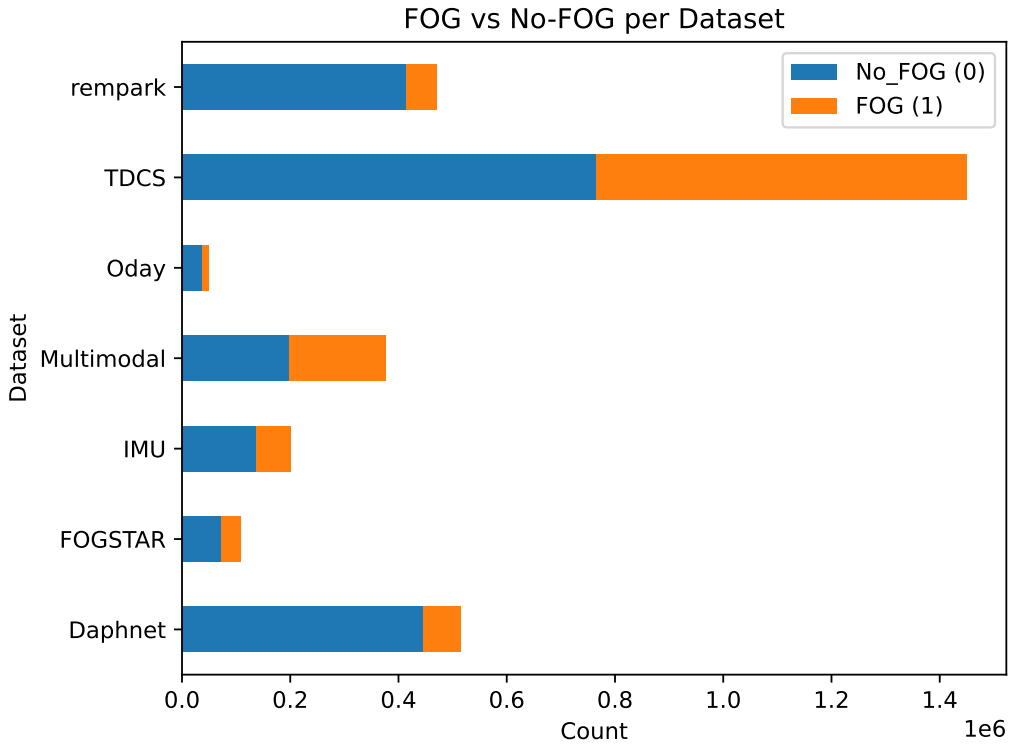


Figure 5.6. Distribution of FoG and non-FoG windows across datasets after segmentation.

To complement this visualization, Table 5.12 reports the absolute number of segmented windows and the corresponding class proportions for each dataset. The results confirm that class imbalance is a pervasive characteristic of wearable-based FoG detection, although its severity varies substantially across datasets.

Dataset	**FoG windows**	**Non-FoG windows**	**Total**	**FoG (%)**	**Non-FoG (%)**
Daphnet	69,229	446,195	515,424	13.43	86.57
FOGSTAR	35,302	72,889	108,191	32.63	67.37
IMU	64,476	137,124	201,600	31.98	68.02
Multimodal	177,760	199,074	376,834	47.17	52.83
Oday	11,614	37,984	49,598	23.42	76.58
TDCS	684,827	765,712	1,450,539	47.21	52.79
Rempark	56,542	414,090	470,632	12.01	87.99

Table 5.12. Number and proportion of FoG and non-FoG windows after segmentation.

A marked variability in class proportions can be observed across datasets. At one end of the spectrum, Rempark and Daphnet exhibit pronounced imbalance, with FoG windows accounting for only 12.01% and 13.43% of the total, respectively. In these datasets, the large dominance of non-FoG windows reflects the episodic nature of freezing events and the strong prevalence of normal gait or non-freezing behavior during acquisition. At the other end, Multimodal and TDCS display a substantially more balanced distribution, with FoG windows representing approximately 47% of all segments. FOGSTAR and IMU occupy an intermediate position, with FoG proportions around one third of the total windows, whereas Oday shows a lower but still substantial FoG prevalence (23.42%).

These differences are not merely statistical, but also reflect the underlying acquisition protocols and the clinical characteristics of the cohorts. Datasets collected under highly controlled, symptom-eliciting conditions tend to contain a larger proportion of FoG events, whereas datasets acquired in more naturalistic or broader behavioral settings are dominated by non-FoG segments. Consequently, the segmentation results provide an indirect but informative picture of the extent to which each dataset is enriched for freezing episodes.

From a modeling perspective, this distribution has important implications. First, the imbalance confirms that standard accuracy alone would be insufficient to characterize model performance, since a classifier biased toward the majority class

could still achieve deceptively high scores. This observation supports the use of metrics that are more informative under imbalanced conditions, such as precision, recall, F1-score, and the area under the precision–recall curve (AUPRC). Second, the strong variability in FoG prevalence across datasets indicates that cross-dataset learning is inherently challenging: models developed on more balanced datasets may behave differently when transferred to datasets in which FoG is considerably rarer.

Overall, the segmentation analysis confirms that FoG detection is intrinsically an imbalanced and dataset-dependent problem. This reinforces the need for evaluation strategies and modeling approaches that explicitly account for heterogeneous class distributions in subsequent experiments.

5.4.13.3 MoE performance under LODO evaluation

The performance of the Mixture-of-Experts (MoE) model was evaluated using a Leave-One-Dataset-Out (LODO) protocol, in which each dataset was iteratively excluded from training and used exclusively for testing. The results are reported in Table 5.13.

Left-out dataset	**F1**	**AUROC**	**AUPRC**	**Bal. Acc.**	**MCC**	**Test windows**
Multimodal	0.669	0.500	0.503	0.500	0.000	9386
TDCS	0.650	0.652	0.642	0.500	0.000	35645
FOGSTAR	0.506	0.500	0.338	0.500	0.000	2636
IMU	0.491	0.436	0.291	0.510	0.069	4998
Oday	0.446	0.605	0.320	0.550	0.148	1223
Daphnet	0.241	0.523	0.141	0.507	0.045	12872
rempark	0.202	0.526	0.118	0.519	0.027	11715

Table 5.13. Performance of the MoE model under Leave-One-Dataset-Out (LODO) evaluation.

Performance varies substantially across datasets, with a clear spread between higher-performing and lower-performing cases. The highest F1-scores are observed for *Multimodal* (0.669) and *TDCS* (0.650), which also correspond to the highest AUPRC values (0.503 and 0.642). These datasets exhibit both relatively balanced precision–recall trade-offs and more stable ranking performance. Intermediate performance is obtained for *FOGSTAR* (F1 = 0.506), *IMU* (F1 = 0.491), and *Oday* (F1 = 0.446), where AUPRC values range between 0.291 and 0.338, indicating partial but less consistent discrimination of FoG events. The lowest performance is observed

for *Daphnet* (F1 = 0.241) and *rempark* (F1 = 0.202), where AUPRC values fall below 0.15, reflecting limited precision in identifying positive instances.

A consistent gap can be observed between AUROC and AUPRC across all datasets. While AUROC values remain within a relatively narrow range (0.436 to 0.652), AUPRC exhibits a much wider variability, suggesting that the ranking capability of the model is more stable than its precision under class imbalance. This discrepancy is particularly evident in datasets such as *Multimodal*, where AUROC remains at chance level (0.500) despite a higher F1-score, and in *TDCS*, where both AUROC and AUPRC reach their highest values.

Balanced accuracy values are generally close to 0.5, indicating that the model often operates near a neutral decision boundary when evaluated across unseen datasets. Slight deviations are observed for *Oday* (0.550) and *IMU* (0.510), suggesting a modest improvement in class-wise balance for these datasets. Similarly, MCC values remain low overall, with only *Oday* (0.148) and *IMU* (0.069) showing non-negligible correlation between predictions and ground truth, while other datasets remain close to zero.

The number of test windows varies considerably across datasets, ranging from 1,223 samples for *Oday* to 35,645 samples for *TDCS*. Despite this variability, no direct relationship between dataset size and performance is observed, as both high- and low-performing results are present across small and large datasets.

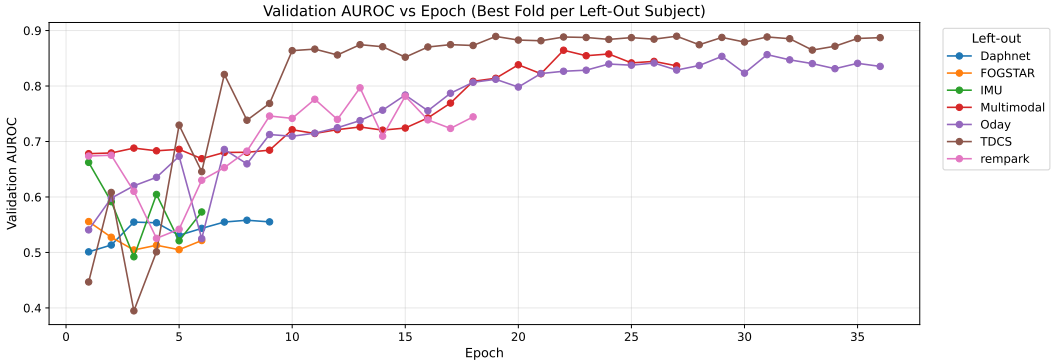


Figure 5.7. Validation AUROC as a function of training epochs for the best-performing fold of each left-out dataset.

The training dynamics are shown in Fig. 5.8 and Fig. 5.7. The validation AUPRC curves (Fig. 5.8) exhibit a progressive increase during training for most datasets, although the rate of convergence and the final plateau differ substantially. In particular, *TDCS* shows rapid improvement and reaches values above 0.85, while datasets such as *FOGSTAR*, *IMU*, and *Daphnet* stabilize at lower levels, often below 0.6. Some

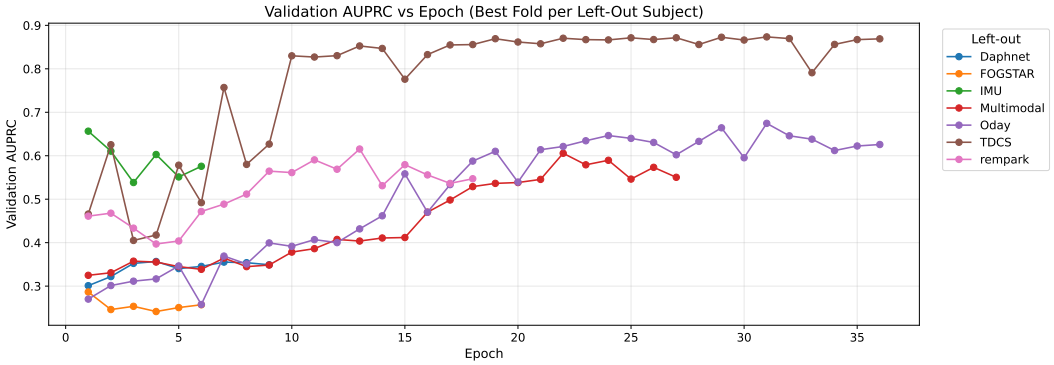


Figure 5.8. Validation AUPRC as a function of training epochs for the best-performing fold of each left-out dataset.

datasets also display higher variability during early epochs, followed by gradual stabilization.

The validation AUROC curves (Fig. 5.7) show a more uniform behavior across datasets, with smoother trajectories and earlier convergence. Datasets such as *TDCS* and *Oday* reach stable values above 0.8, while others, including *IMU* and *Daphnet*, remain closer to lower performance levels throughout training. Compared to AUPRC, AUROC curves exhibit reduced variability between datasets, confirming the different sensitivity of these metrics to class imbalance and prediction confidence.

5.5 Code and implementation availability

The code developed for the experiments presented in this thesis is publicly available at: `https://github.com/MikiTebaldi/Michele-Tebaldi-Ph.D-Thesis`. The repository includes the implementations of the Human Activity Recognition models, the fine-tuning framework for subject-specific FoG detection, and the Mixture-of-Experts architecture described in this work.

Chapter 6

Discussion

This work investigated the problem of wearable-based analysis of Parkinsonian motor behavior from four complementary perspectives: the characteristics and limitations of currently available datasets, the recognition of FoG-related motor activities through Human Activity Recognition (HAR), the use of fine-tuning strategies to improve subject-specific freezing-of-gait (FoG) detection, and the exploration of Mixture-of-Experts (MoE) architectures for position-aware modeling of wearable motion signals. Rather than representing independent contributions, these components should be interpreted as interconnected elements of a broader methodological challenge. In particular, the results and methodological analyses presented throughout this thesis highlight that the performance and reliability of AI-based approaches in Parkinson's disease (PD) are strongly shaped by the structure, quality, and variability of the data on which they are developed, as well as by the extent to which models can exploit contextual, individual, and positional information.

6.1 Dataset heterogeneity and methodological implications

The analysis of the dataset landscape reveals a field that has progressed significantly in recent years, yet remains fragmented. A wide range of datasets has been proposed to monitor PD motor symptoms through wearable sensing, each reflecting different design choices in terms of sensor configuration, acquisition protocols, and clinical focus. This diversity is valuable, as it captures different aspects of Parkinsonian motor behavior, but it also introduces substantial variability that complicates comparison, reproducibility, and generalization.

Sensor placement emerges as one of the most influential factors. Configurations centered on the lower back are widely adopted due to their ability to capture global body dynamics, including gait and posture, with relatively stable signals and limited artifacts [148, 149]. However, this dominance may implicitly bias the development of models toward a specific biomechanical perspective. When transferred to alternative configurations, such models may fail to capture relevant motion patterns, particularly in distal segments. This limitation is further exacerbated by the asymmetric nature of PD symptoms, where motor impairment often differs between the left and right sides of the body. Datasets relying on unilateral or reduced sensor setups may therefore overlook clinically relevant information about disease progression and motor compensation strategies [150].

Beyond placement, inconsistencies in sensor orientation and acquisition settings represent an additional barrier. Differences in coordinate systems, attachment procedures, and sampling frequencies introduce variability at the signal level that is rarely standardized or fully documented. As a result, integrating datasets or transferring models across studies becomes non-trivial and may lead to unreliable performance. This issue is particularly critical when scaling AI solutions beyond controlled experimental environments.

Further reinforcing these observations, Tables 4.32, 4.29, and 4.30 highlight important gaps in dataset documentation. Missing information on symptom characteristics, medication state, activity labels, and clinical assessments reduces interpretability and limits the possibility of meaningful cross-dataset comparisons. In addition, variability in disease stage representation complicates the construction of balanced cohorts, as many datasets focus primarily on moderate stages of PD. This restricts the ability to study early manifestations or longitudinal progression of symptoms. Addressing these issues requires not only technical improvements, but also coordinated efforts toward standardized reporting and data-sharing practices. The recommendations summarized in Table 6.1 reflect this need and outline possible directions for improving dataset design, accessibility, and integration.

Table 6.1. Key Findings, Recommendations, and Future Directions.

Finding	Recommendations and Future Directions
Data Accessibility and Standardization	
Limited accessibility to high-quality datasets hinders reproducibility and external validation.	Establish open repositories for PD datasets and encourage public sharing while ensuring compliance with ethical guidelines (e.g., GDPR).
Variability in data collection protocols, sensor types, and task designs complicates cross-dataset learning.	Develop standardized protocols for sensor placements, orientation, and task designs to improve comparability across datasets.
Datasets lack open-source availability, reducing reproducibility.	Prefer open-source or publicly available datasets to foster collaboration and reproducibility in the research community.
Dataset Size, Diversity, and Integration	
Small sample sizes limit the generalizability of findings to the broader PD population.	Select datasets with large sample sizes to ensure statistical power and robustness. Prioritize datasets with diverse demographics (e.g., age, gender, ethnicity) and clinical characteristics (e.g., disease severity, comorbidities).
Focus is often on moderate stages of PD, neglecting early-stage data critical for early diagnosis.	Focus on datasets where disease duration is minimal to improve early-stage PD detection and diagnosis.
Heterogeneity in datasets for disease severity classification is limited.	Promote datasets with variability in clinical information and disease duration to capture the spectrum of disease progression.
Limited dataset integration affects model robustness.	Combine multiple datasets, when feasible, to enhance model robustness and enable cross-dataset generalization.
Task and Activity Variety	
Focus on controlled tasks over real-world, free-living activities reduces ecological validity.	Opt for datasets that include a wide range of activities or tasks, especially those reflecting free-living or naturalistic settings to better represent real-world behavior.

Continued on next page

Finding	Recommendations and Future Directions
Simulated activities of daily living (ADLs) are underrepresented.	Include tasks representative of ADLs in datasets to improve the ecological validity of models.
Validation and Testing	
Lack of rigorous validation protocols affects model generalizability.	Employ rigorous validation schemes (e.g., leave-one-subject-out cross-validation) and validate algorithms on independent datasets to ensure robustness across populations and sensor setups.
Sensor Placement and Configuration	
Dominance of wrist and lower-back sensor placements, with limited use of head-mounted or chest-mounted sensors.	Diversify sensor placements to include head-mounted sensors for head tremors and chest-mounted sensors for vital sign monitoring.
One-sided sensor placements may lead to asymmetrical data collection, reducing accuracy for detecting bilateral symptoms.	Opt for two-sided or central placements to improve balance and enhance detection of lateralized symptoms.
Differences in sensor settings (e.g., orientation, sampling rates) affect performance.	Address differences in sensor settings across datasets to enable cross-dataset generalization and performance in heterogeneous environments.
Contextual and Longitudinal Data	
Datasets often lack contextual metadata to interpret sensor readings.	Select datasets that include contextual metadata (e.g., activity logs) to improve model interpretability.
Limited availability of longitudinal datasets for disease progression monitoring.	Use datasets that include longitudinal data to capture disease progression and enable long-term outcome predictions.
AI and Computational Techniques	
Limited use of advanced AI techniques for symptom detection and dataset integration.	Employ multimodal AI, transfer learning, and federated learning to enhance symptom detection and integrate datasets efficiently.

Continued on next page

Finding	Recommendations and Future Directions
Lack of real-time applicability in current models due to hardware limitations.	Develop AI models optimized for resource-constrained edge devices to enable real-time symptom monitoring.
Clinical Relevance	
Datasets often lack clinical validation and fail to capture heterogeneous patient demographics.	Collaborate with clinicians during dataset design and ensure inclusion of diverse demographics to improve clinical relevance.
Focus on motor symptoms over non-motor symptoms limits comprehensive PD management strategies.	Align dataset development with clinical priorities by incorporating both motor and non-motor symptoms into data collection protocols.

6.2 HAR as a contextual layer for FoG-related activities

Within this fragmented landscape, the HAR analysis provides an important intermediate perspective. The results demonstrate that, despite the variability inherent to Parkinsonian movement, motor activities closely linked to FoG can be recognized with high accuracy using wearable inertial data. Compared with previous studies, the present work benefits from a larger cohort, a richer set of annotated activities, and a dense multi-sensor configuration. This combination enables a more detailed characterization of movement dynamics and supports robust subject-independent evaluation, which is essential for assessing real-world applicability.

To contextualize these findings, Table 6.2 presents a comparison with representative studies in the literature. A clear pattern emerges: previous works often rely on limited sample sizes, simplified activity sets, or reduced sensing configurations. For instance, early studies such as [78] and [80] focus on a small number of basic activities (e.g., sitting, standing, walking) and employ minimal sensor setups, resulting in relatively modest performance (72–75% accuracy). While these approaches demonstrate feasibility, they do not fully capture the complexity of Parkinsonian motor behavior.

More recent work, such as [81], reports higher performance using deep learning and a more diverse activity set. However, these studies still rely on single-sensor

configurations and often adopt evaluation strategies that are not fully transparent or may not reflect subject-independent generalization. In contrast, studies that incorporate more complex sensor configurations, such as [79], tend to suffer from reduced generalizability, as indicated by the drop in F1-score when models are evaluated across different subject groups.

In this context, the present study distinguishes itself by combining three key aspects: a relatively large cohort of 60 PwPD, a comprehensive full-body IMU configuration with 13 sensors, and a strict subject-wise hold-out validation strategy. This combination allows not only for high performance (91–92% accuracy and F1-score), but also for a more reliable estimation of generalization to unseen individuals. Importantly, the use of a subject-disjoint split prevents information leakage and avoids the overly optimistic performance often observed in sample-wise validation schemes.

The achieved performance, with accuracy and F1-scores in the range of 91–92%, indicates that the selected activities—such as walking, turning, and postural transitions—have sufficiently distinct motion signatures to be reliably identified. At the same time, the analysis of misclassifications provides valuable insight into the limits of current approaches. Errors are predominantly concentrated in transitional activities, particularly between `TurnToSit` and `Turning`. This reflects the intrinsic overlap between these movements and suggests that the main challenge is not distinguishing clearly separated motor states, but rather modeling continuous transitions and composite actions.

This observation is consistent with the nature of real-world motor behavior, where activities do not occur as isolated events but as fluid sequences. In this sense, HAR models that rely on discrete labeling may inherently struggle to capture transitional dynamics, especially when class imbalance favors more frequent activities.

From a sensing perspective, the results confirm that both sensor modality and placement significantly influence recognition performance. Gyroscope signals provide richer information than accelerometers alone, particularly for capturing rotational dynamics during turning and transitions. The combination of both modalities yields the best results, highlighting the importance of multimodal sensing. In terms of body location, proximal segments such as the thighs, lower back, and sternum emerge as the most informative, likely due to their central role in locomotion and posture control. These findings have practical implications for wearable system design, suggesting that carefully selected sensor placements may achieve strong performance even with a reduced number of devices, thereby improving usability and patient compliance.

Study	Subjects	Sensors	Activities	Validation	Results
[78]	18 HS, 8 PD	1 acc (pocket)	4 (sitting, standing, holding the phone, walking)	LOSO	Acc = 75.1%
[80]	19 HS, 14 PD	2 acc (wrist, ankle)	3 (standing, sitting, walking)	Cross-test	Acc = 72%
[81]	18 PD	1 IMU (back)	7 (walking, turning, sitting, standing, sit-to-stand, lying, get up)	Hold-out (unknown)	Acc = 91.1% F1-score = 90.7%
[79]	8 HS, 8 PD	3 IMUs (chest, forearm, shank)	21 exercises	Cross-test	F1-score = 50%
Present	60 PD	13 IMUs (head, sternum, back, wrists, thighs, shins, ankles, feet)	6 (walk, turn, sit-to-stand, stand, turn-to-sit)	Hold-out (test = 12)	Acc = 91–92% F1-score = 91–92%

Table 6.2. Comparison of methods and results of the present study with those of similar works. PD: Parkinson's Disease; HS: Healthy Subjects; acc: accelerometer; LOSO: Leave-One-Subject-Out.

Although classical models such as Random Forest show strong performance, deep learning approaches offer additional advantages. Their ability to learn hierarchical spatiotemporal representations directly from raw signals makes them more adaptable to variability and noise. Recurrent architectures such as LSTM, in particular, provide a good balance between modeling temporal dependencies and maintaining computational efficiency, making them suitable candidates for real-time wearable applications.

However, the results of the HAR analysis also highlight an important limitation: high activity recognition performance does not automatically translate into reliable FoG detection. This gap becomes evident in the subsequent analysis of FoG recognition, where models trained on population-level data exhibit reduced performance when evaluated on unseen subjects. This confirms that FoG is not only a rare and imbalanced event, but also a highly individualized phenomenon, with substantial variability in its manifestation across patients. Consequently, while HAR provides a robust contextual layer for understanding motor behavior, additional methodological strategies—such as personalization through fine-tuning—are required to effectively address the complexity of FoG detection.

6.3 Subject-specific adaptation through fine-tuning

In this context, the fine-tuning experiments clarify a point that is often suggested in the literature but less frequently demonstrated under stringent evaluation settings: the main limitation of wearable-based FoG detection is not simply the lack of expressive models, but the difficulty of transferring a global representation of freezing to new individuals. The baseline experiments already show that even when a model learns meaningful freezing-related patterns, these patterns do not generalize uniformly across subjects. This is consistent with the clinical nature of FoG, whose manifestation varies substantially in duration, motor expression, and triggering context from one patient to another.

From this perspective, the value of fine-tuning is not that it improves performance in a generic sense, but that it reduces the mismatch between population-level learning and subject-specific motor expression. The adaptation step appears to work precisely because the pretrained model has already learned useful generic representations of gait interruption and abnormal movement dynamics, while the final layers remain flexible enough to specialize these representations to the target subject. In other words, the global model captures the broad structure of the problem, whereas fine-tuning

calibrates the decision boundary to the individual phenotype.

This interpretation helps position the present findings relative to previous studies without simply repeating them. Prior work has already indicated that transfer learning and fine-tuning can be beneficial for FoG recognition, but these studies were generally conducted on small cohorts and often under validation settings that make performance difficult to compare directly. In particular, the literature tends to rely on limited datasets, frequently derived from Daphnet or similarly small public resources, and often adopts random or partially overlapping splits that risk inflating performance by exposing the model to subject-specific characteristics during development. As a result, those studies demonstrate the promise of personalization, but they do not fully resolve whether fine-tuning remains effective once the initial model has been trained and tested under a stricter subject-independent framework.

The results presented in Section 5 are valuable precisely because they address that point. Here, fine-tuning is not applied to an already favorable experimental setup, but to a baseline model whose limits have already emerged under cross-subject evaluation. This changes the interpretation of adaptation. Rather than acting as a minor optimization step, fine-tuning becomes a mechanism for recovering subject-specific information that the global model cannot represent adequately on its own. The fact that adaptation is most beneficial for subjects with weaker baseline performance is particularly meaningful: it suggests that personalization is most useful where inter-subject variability is greatest, namely in those cases where a generic detector is least reliable.

Another important implication concerns how FoG detection systems should be conceived in practice. The results suggest that the field may benefit less from pursuing a single universal detector and more from designing hybrid pipelines in which a general model is first trained on a structured cohort and then adapted with a small amount of target-user data. This is a more realistic scenario for clinical deployment. In real-world use, it is unlikely that every new patient will resemble the training population closely enough for a fixed model to remain robust; at the same time, collecting a fully new labeled dataset for each subject is clearly impractical. Fine-tuning occupies the middle ground between these extremes, making personalization feasible without abandoning the advantages of shared pretraining.

The subject-level variability observed after adaptation also offers an important nuance. Fine-tuning substantially improves performance in most cases, but not uniformly. This indicates that personalization is necessary, yet not sufficient on its own. Some subjects remain difficult even after adaptation, which likely reflects

deeper sources of variability such as limited within-subject data, atypical freezing signatures, class imbalance, or annotation ambiguity. This is an important point because it prevents overinterpreting fine-tuning as a complete solution. Instead, it should be viewed as one essential component of a broader strategy that may also require improved data diversity, richer contextual information, and more temporally aware models.

The methodological choices of the present study also reinforce this interpretation. The use of a CNN as the base architecture suggests that the learned convolutional filters capture transferable low-level motion patterns, while the fine-tuned dense layers adapt these features to the subject-specific structure of the data. This supports a layered view of FoG modeling: some aspects of freezing-related movement are shared across patients and can be learned globally, whereas others remain individual and must be calibrated locally. In this sense, the results do not argue against general models; rather, they redefine their role as the foundation for adaptive systems rather than as complete standalone solutions.

More broadly, these findings refine how the literature on transfer learning in FoG should be interpreted. Previous studies were correct in identifying personalization as a promising direction, but the present work suggests that its importance becomes even clearer when evaluated under more realistic constraints. Under permissive validation settings, fine-tuning may appear as an optional enhancement. Under subject-independent evaluation, it instead emerges as a practical necessity for bridging the gap between cohort-level modeling and patient-level deployment.

Taken together, the fine-tuning experiments support a shift in perspective. The central challenge is not only to detect FoG accurately in aggregate, but to build systems that remain reliable once confronted with the heterogeneity of individual patients. From that standpoint, personalization is not an accessory added after model development, but a central design principle for clinically meaningful wearable FoG monitoring.

6.4 Cross-dataset modeling through Mixture-of-Experts

The Mixture-of-Experts (MoE) experiment provides a complementary perspective on one of the central challenges that emerged throughout this thesis: the difficulty of generalizing FoG detection models across datasets collected under different sensing

configurations and experimental conditions. While the previous analyses showed that both activity recognition and subject-specific fine-tuning can be effective under more controlled or personalized settings, the MoE framework was introduced to address a different level of variability, namely the structural heterogeneity of wearable data across studies. In this sense, the MoE results should not be interpreted simply in terms of absolute classification performance, but as an evaluation of whether position-aware expert specialization can mitigate cross-dataset mismatch.

This question is particularly relevant because, as highlighted in the literature review, no prior work has explicitly applied MoE architectures to wearable-based FoG detection. Existing MoE applications in inertial sensing and biomedical AI suggest that expert specialization can be advantageous when different subspaces of the input correspond to distinct sensing regimes, movement patterns, or data domains. However, these studies do not address the specific problem considered here, where the heterogeneity is not only statistical but also anatomical: different datasets provide different subsets of body locations, with distinct combinations of accelerometer and gyroscope signals. The proposed framework was designed precisely to exploit this structure by assigning experts to sensor-specific views of motion and constraining their activation according to dataset compatibility.

From this perspective, the results highlight both the potential and the limitations of the proposed strategy. The model does not achieve robust cross-dataset generalization under the stringent Leave-One-Dataset-Out (LODO) protocol. While *TDCS* and *Multimodal* represent the most favorable cases, performance remains modest overall, and several metrics stay close to chance level. This indicates that the information learned from the heterogeneous source datasets is only weakly transferable to unseen target domains. Importantly, this should be interpreted in the context of the difficulty of the LODO setting, where the model is exposed not only to new subjects but also to entirely new data domains, including different sensor placements, acquisition protocols, preprocessing pipelines, and class distributions. In this respect, even limited transfer suggests that expert specialization is able to capture some sensor-location-dependent regularities that remain partially informative across datasets.

At the same time, the results clearly show that this mechanism is not sufficient to overcome domain shift. Performance drops substantially for datasets such as *Daphnet* and *rempark*, and remains only moderate for *FOGSTAR*, *IMU*, and *Oday*. This variability indicates that the transferability of expert representations depends strongly on the degree of compatibility between source and target sensing configurations, as well as on broader dataset characteristics. In other words, while the MoE framework

provides a structured way to organize heterogeneous inputs, it does not eliminate the underlying discrepancies between datasets. The architecture can guide the model in selecting relevant experts, but the usefulness of those experts ultimately remains constrained by the quality and representativeness of the source data.

A particularly informative aspect of the results is the difference between AUROC and AUPRC. AUROC remains relatively constrained across datasets, whereas AUPRC varies much more substantially. This pattern is consistent with the segmentation analysis presented earlier, which showed strong class imbalance and considerable differences in FoG prevalence across datasets. In this context, AUPRC offers a more faithful picture of practical FoG detection performance because it reflects the model's ability to identify the positive class under skewed class distributions. The fact that AUPRC remains comparatively high only for a subset of datasets indicates that expert specialization alone does not guarantee robust positive-class detection under severe distribution shift. Put differently, the MoE framework improves structural flexibility, but the imbalance and rarity of FoG events remain fundamental obstacles.

The training curves further support this interpretation. For datasets such as *TDCS*, validation AUPRC and AUROC increase steadily and stabilize at relatively high values, suggesting that the model can exploit the available experts in a consistent and informative way. In contrast, datasets such as *Daphnet* and *IMU* show lower and less favorable trajectories, indicating that the learned expert representations are less transferable to those domains. This difference is methodologically important because it suggests that the MoE architecture does not fail uniformly: rather, its effectiveness depends on how well the latent expert structure learned from the training pool aligns with the target dataset. In some cases, the target domain appears sufficiently covered by the available experts; in others, the mismatch remains too strong.

An additional factor that may have contributed to the limited performance is the decision to remove the dataset embedding from the gating network after the resampling stage. This choice was motivated by the assumption that, once resampled to a common frequency and preprocessed under a unified pipeline, the datasets could be treated as a single coherent data distribution. However, the results obtained under the LODO protocol suggest that this assumption may be overly restrictive. While resampling ensures temporal consistency, it does not eliminate deeper sources of heterogeneity, such as differences in sensor placement, acquisition protocols, annotation strategies, and population characteristics. By removing the dataset-specific embedding, the gating mechanism is prevented from explicitly modeling these domain-level differences, and is therefore forced to rely solely on signal-derived

features and expert availability.

In this context, the absence of dataset-level information may limit the ability of the MoE framework to adapt its expert weighting to the specific characteristics of each domain. This suggests that explicitly encoding dataset identity, or more generally domain-related information, could be beneficial for improving cross-dataset generalization in future extensions of the model.

A further additional aspect that may have contributed to the limited performance is the simplicity of the expert architecture itself. Each expert is implemented as a relatively shallow one-dimensional CNN, designed to remain lightweight and computationally efficient. While this choice is consistent with the goal of enabling deployment on resource-constrained wearable devices, it may limit the ability to capture the complexity of FoG-related motion patterns, especially in a cross-dataset setting. In particular, shallow convolutional architectures may struggle to model long-range temporal dependencies and the subtle temporal organization that characterizes freezing episodes. This limitation becomes more pronounced when the model is required to operate across datasets with different sensor configurations, noise characteristics, and acquisition protocols.

Moreover, within the MoE framework, the role of the experts is not only to extract features, but to provide meaningful alternatives that the gating mechanism can combine. If the representations produced by individual experts lack sufficient discriminative power, the gating network is constrained to reweight features that are already limited in expressiveness. In this sense, the performance observed in the LODO experiments suggests that the bottleneck may arise from the interaction between domain shift and the limited capacity of the expert representations, rather than from the gating strategy alone.

These findings help refine the role of MoE within the broader framework of this thesis. The expert-based architecture does not replace the need for carefully curated datasets, nor does it solve the personalization problem addressed by fine-tuning. Instead, it occupies an intermediate methodological layer between global modeling and individual adaptation. Relative to conventional monolithic models, MoE provides a more principled way to incorporate sensor-location information and to preserve the modularity of heterogeneous wearable inputs. This is especially relevant in a multi-dataset setting, where forcing all signals into a single undifferentiated representation would ignore an important source of structure in the data. At the same time, the present results indicate that modularity alone is not sufficient for full cross-domain robustness.

A further strength of the proposed framework lies in its interpretability at the architectural level. Because each expert corresponds to a specific sensor configuration or body location, the model's behavior can in principle be analyzed in terms of which anatomical viewpoints contribute most strongly under different dataset conditions. Even when absolute performance remains limited, this structural transparency is valuable: it provides a more meaningful representation of the sensing problem than a black-box classifier trained on arbitrarily concatenated inputs. In this sense, the contribution of the MoE framework is not only predictive but also conceptual, as it introduces a position-aware formulation for wearable FoG detection that is directly aligned with the heterogeneous nature of the available datasets.

Taken together, the MoE experiments support a nuanced conclusion. They suggest that expert specialization by body location provides a conceptually coherent strategy for handling heterogeneous sensor configurations. However, the current results indicate that this approach, in its present form, is not sufficient to achieve robust cross-dataset performance. While some datasets show comparatively better outcomes, the overall variability and near-chance performance in several cases highlight the need for more expressive expert models and stronger domain adaptation mechanisms. At the same time, the residual variability in performance across left-out datasets makes clear that cross-dataset FoG detection remains an unresolved challenge. The results therefore position MoE not as a complete solution, but as an important step toward more flexible and structure-aware wearable models.

6.5 Toward Robust and Clinically Meaningful FoG Detection

The development of wearable-based digital biomarkers increasingly relies on structured validation frameworks that extend beyond algorithmic performance alone. In this context, the Digital Medicine Society (DiMe) has proposed the V3 and V3+ frameworks [151, 152], which describe a progressive validation pathway based on verification, analytical validation, clinical validation, and usability or feasibility assessment. Viewed through this framework, the work presented in this thesis primarily contributes to the verification and analytical validation stages. The extensive characterization of wearable datasets, sensor configurations, acquisition protocols, and harmonization strategies addresses verification-related aspects by examining the quality and comparability of the underlying measurements. Similarly, the methodological

developments presented throughout Chapter 5 — including patient-wise validation, cross-dataset evaluation, fine-tuning strategies, and Leave-One-Dataset-Out (LODO) experiments — contribute to analytical validation by assessing the robustness of the proposed approaches under heterogeneous sensing conditions. Nevertheless, clinical validation remains only partially addressed. Although the investigated datasets contain clinically annotated FoG episodes and represent meaningful Parkinson's disease populations, the relationship between model outputs and clinically actionable outcomes, as well as validation in longitudinal real-world settings, remains an important direction for future work. Likewise, usability and feasibility aspects—including patient acceptance, clinician interpretability, workflow integration, and long-term deployment—fall beyond the scope of this thesis and represent subsequent steps toward routine clinical adoption. From this perspective, the contributions presented here should be regarded as methodological building blocks within a broader digital biomarker validation pathway. An important methodological consideration concerns the ecological validity of the Freezing of Gait episodes used throughout this work. As described in Chapter 4, the FOGSTAR dataset, as well as other datasets described there, was collected using a structured clinical protocol specifically designed to increase the likelihood of observing FoG episodes through clinically established triggers, including OFF-medication assessments, turning tasks, constrained walking conditions, and dual-task paradigms. This strategy provides important advantages by enabling the acquisition of sufficient numbers of carefully annotated FoG episodes under controlled conditions, thereby improving data quality and reproducibility. However, elicited FoG episodes may not fully represent the complexity and variability of freezing events occurring during activities of daily living. In real-world environments, FoG may emerge through more complex interactions involving environmental constraints, emotional state, cognitive load, fatigue, medication fluctuations, and spontaneous behavioral adaptations. Consequently, datasets collected under structured protocols may overrepresent specific triggering conditions while underrepresenting the broader spectrum of daily-life manifestations. These considerations directly affect the interpretation of model generalizability. Although the datasets considered in this thesis include clinically relevant and carefully annotated FoG episodes, most recordings were acquired under controlled experimental conditions designed to maximize the occurrence of freezing events. Therefore, the performance obtained on these datasets should not be interpreted as direct evidence of equivalent performance during unrestricted daily-life monitoring. Rather, the presented results demonstrate the ability of the proposed methodologies to generalize across heterogeneous datasets

collected under different experimental settings, while highlighting that ecological generalization remains an open research challenge. The cross-dataset experiment performed in Sections 5 represents an important step toward addressing this issue by evaluating robustness across different populations, sensing configurations, and acquisition protocols. Nevertheless, the observed variability under the stringent Leave-One-Dataset-Out evaluation protocol indicates that robust generalization across all clinical scenarios has not yet been achieved. These findings reinforce the need for future validation using long-term, home-based, and unscripted recordings that better reflect everyday patient behavior. Beyond statistical performance, the clinical significance of wearable-based FoG detection depends on how model errors influence the intended application. Performance metrics such as sensitivity, specificity, precision, and F1-score provide complementary information, but their practical importance varies according to the clinical context. High sensitivity is desirable because missed FoG episodes (false negatives) may underestimate symptom burden and reduce the ability to accurately characterize disease progression. Conversely, high specificity minimizes false-positive detections, an aspect that becomes particularly important when model outputs are used to support clinical decision-making or quantify FoG frequency over extended monitoring periods. Importantly, the acceptable balance between false positives and false negatives depends on the intended application. For retrospective monitoring and clinical assessment, occasional false-positive detections may be acceptable provided that the majority of true FoG episodes are identified, as clinicians can interpret aggregated results together with complementary clinical information. In contrast, real-time interventions—such as auditory, visual, or haptic cueing systems—require much stricter control of false alarms, since excessive false-positive detections may reduce user confidence, increase cognitive burden, and ultimately compromise adherence to the technology. Likewise, while isolated false negatives may have limited impact during longitudinal symptom monitoring, failure to detect a freezing episode in real-time assistive systems could prevent the delivery of a potentially beneficial intervention when it is most needed. Accordingly, the results obtained in this thesis should not be interpreted solely as measures of classification accuracy, but rather as evidence of the trade-offs that future wearable FoG monitoring systems must manage. The proposed fine-tuning and Mixture-of-Experts strategies demonstrate promising approaches for improving personalization and robustness across heterogeneous sensing conditions; however, the selection of operating thresholds and evaluation criteria should ultimately be driven by the specific clinical application. From

a translational perspective, the different methodological components developed throughout this thesis may support complementary aspects of Parkinson's disease management. The Human Activity Recognition framework presented in Section 5.1 provides contextual information that may enhance retrospective mobility assessment and facilitate clinical research into motor behavior surrounding FoG episodes. The subject-specific fine-tuning strategy described in Section 5.2 addresses one of the major challenges of wearable monitoring—inter-individual variability—and therefore represents a promising approach for personalized long-term home monitoring. Similarly, the Mixture-of-Experts framework proposed in Sections 5 was specifically designed to improve robustness across heterogeneous sensing configurations, potentially facilitating deployment in scenarios where sensor placement or hardware availability varies between users. Collectively, these methodologies may contribute to future applications including continuous home monitoring, retrospective symptom quantification, rehabilitation support, and large-scale clinical research studies. Although real-time cueing systems remain an attractive long-term objective, additional validation regarding latency, reliability, false-alarm rates, usability, and ecological robustness will be required before considering routine clinical deployment. Consequently, the contributions of this thesis should be interpreted primarily as methodological advances toward scalable, personalized, and clinically meaningful wearable-based FoG monitoring systems rather than as immediately deployable clinical solutions. Future research should therefore extend these developments through prospective real-world validation studies, user-centered evaluations, and broader clinical validation, ultimately completing the pathway from methodological innovation to clinically adopted digital biomarkers.

6.6 Toward an integrated perspective

In the broader context of this thesis, these findings reinforce the idea that robust FoG monitoring cannot rely on a single modeling strategy, but instead requires the integration of multiple complementary components. Dataset-aware modeling is necessary to explicitly account for the structural heterogeneity of wearable signals across studies, as explored through the MoE framework. At the same time, context-sensitive representations, such as those provided by HAR, play a key role in capturing the underlying motor activities that surround and characterize freezing events. Finally, subject-specific adaptation mechanisms, exemplified by fine-tuning, are essential to bridge the gap between population-level learning and individual variability in FoG

expression.

Rather than acting in isolation, these elements can be viewed as operating at different levels of the problem: dataset-level variability, activity-level context, and subject-level specificity. The results presented in this work suggest that addressing only one of these aspects is insufficient to achieve reliable performance, particularly under realistic conditions involving heterogeneous datasets and unseen individuals. Instead, a more integrated perspective is required, in which structural heterogeneity, behavioral context, and individual differences are jointly considered during model design and evaluation.

Chapter 7

Conclusions and future perspectives

Taken together, the results of this thesis suggest that the main challenges in wearable-based FoG detection are not purely algorithmic, but structural. The variability of Parkinsonian motor behavior, the limited availability of annotated data, and the diversity of experimental conditions collectively constrain model performance. Addressing these challenges requires methodological approaches that go beyond optimizing classification accuracy within a single dataset. Instead, future work should focus on integrating contextual modeling, data-efficient learning strategies, and adaptive mechanisms capable of accommodating variability across individuals and conditions.

Several limitations should be acknowledged. First, the analyses rely on a limited number of publicly available datasets, which, despite their diversity, do not fully capture the breadth of clinical variability observed in Parkinson's disease. Differences in sensor placement, acquisition protocols, and annotation procedures introduce sources of variability that are difficult to control and may affect the reproducibility of the results. In particular, the Leave-One-Dataset-Out evaluation highlights how strongly model performance depends on the compatibility between datasets, suggesting that current benchmarks may still underestimate the complexity of real-world deployment.

Second, although the proposed models provide a structured and interpretable framework, their representational capacity may be limited in the specific context of cross-dataset generalization. In particular, the convolutional architecture employed within each expert of the MoE framework is relatively shallow, which may restrict its ability to capture long-range temporal dependencies and subtle motor patterns associated with FoG. This limitation is less evident in the HAR and fine-tuning experiments, where the models operate under more homogeneous or subject-adapted

conditions, but becomes particularly critical in the LODO setting, where the combination of domain shift and heterogeneous sensor configurations demands more expressive temporal representations.

For these reasons, the proposed Mixture-of-Experts framework has to be interpreted primarily as a methodological contribution aimed at improving robustness across heterogeneous datasets and sensing configurations. Although the results demonstrate the potential of expert specialization and adaptive sensor-aware modeling, the observed variability under stringent cross-dataset evaluation indicates that reliable generalization across all clinical settings remains an open challenge. Consequently, the present work represents a step toward more deployable FoG detection systems rather than a clinically validated solution for routine use.

Third, the study focuses primarily on inertial sensing modalities. While IMUs offer a practical and widely adopted solution for wearable monitoring, they provide only a partial view of the physiological processes underlying freezing episodes. The absence of complementary modalities, such as plantar pressure, electromyography, or physiological signals, may limit the ability to fully characterize the onset and progression of FoG, especially in complex or ambiguous scenarios.

From a methodological perspective, several future directions emerge. A first promising avenue concerns the development of more expressive temporal models, including temporal convolutional networks [153], attention-based architectures [154], or hybrid approaches that combine convolutional and recurrent components [127,155]. Such models could improve the representation of temporal context and better capture the dynamic nature of freezing episodes.

A second direction involves the integration of domain adaptation and domain generalization techniques within the proposed MoE framework. The results obtained under the LODO protocol show that expert specialization alone is not sufficient to compensate for cross-dataset variability, particularly when the target dataset exhibits sensing configurations or signal characteristics not adequately represented in the training pool. In wearable HAR, domain adaptation approaches such as adversarial and contrastive transfer learning have been proposed to address cross-user, cross-body, and cross-sensor mismatch [156]. Similarly, domain generalization methods have been developed to learn representations that remain robust across unseen target domains without requiring access to target-domain data during training [157]. Integrating such mechanisms directly into the MoE architecture—either at the expert level or within the gating function—could improve the transferability of the learned representations across heterogeneous sensing environments.

A third line of research concerns the combination of personalization and global modeling. The fine-tuning experiments demonstrate that subject-specific adaptation is highly effective, but also highlight that it does not fully resolve all cases. Future systems could explore more flexible adaptation mechanisms, such as continual learning [158] or few-shot learning [159], enabling models to update progressively as new patient data become available without requiring full retraining.

An additional important direction concerns the standardization of dataset construction and acquisition protocols. As highlighted throughout this work, a significant portion of the variability observed in model performance originates from differences in sensor placement, sampling frequencies, annotation procedures, and experimental protocols across datasets. These inconsistencies complicate cross-dataset evaluation and limit the reproducibility and transferability of learned models. Future efforts should therefore focus on defining shared guidelines for wearable data collection, including standardized sensor configurations, consistent labeling strategies, and harmonized preprocessing pipelines. Such standardization would not only facilitate more reliable benchmarking, but also enable the development of models that generalize more effectively across studies and clinical settings.

Not less importantly, the integration of multimodal sensing represents a key opportunity for advancing FoG detection. Combining inertial data with complementary sources of information, such as pressure insoles, video-based pose estimation, or physiological signals, could provide a richer and more robust representation of motor behavior. However, this also introduces challenges related to synchronization, data fusion, and increased system complexity, which must be carefully addressed.

Finally, the transition from experimental research to real-world deployment introduces additional constraints related to hardware, energy consumption, latency, and user compliance. Models must be not only accurate but also efficient, interpretable, and robust to noise and missing data. Edge computing solutions, model compression techniques, and privacy-preserving learning paradigms, such as federated learning, will play a crucial role in enabling practical clinical applications.

In conclusion, this study provides a comprehensive analysis of the challenges and opportunities associated with wearable-based monitoring of Parkinsonian motor behavior. By linking dataset characteristics, activity recognition, and personalized learning strategies, it contributes to a more integrated understanding of FoG detection. At the same time, the results emphasize that robust and clinically meaningful solutions will require a combination of advances in data collection, model design, and evaluation methodology. Rather than relying on a single paradigm, future progress will likely

emerge from the coordinated development of dataset-aware, context-sensitive, and adaptive systems capable of operating reliably in real-world conditions.

Bibliography

[1] GBD 2019 Neurological Disorders Collaborators. Global, regional, and national burden of neurological disorders, 1990–2019. *Lancet Neurology*, 20(11):918–947, 2021.

[2] Tamara Pringsheim, Nathalie Jette, Allan Frolkis, and Thomas D. L. Steeves. The prevalence of parkinson's disease: a systematic review and meta-analysis. *Movement Disorders*, 29(13):1583–1590, 2014.

[3] L. M. L. De Lau and M. M. B. Breteler. Epidemiology of parkinson's disease. *Lancet Neurology*, 5(6):525–535, 2006.

[4] S. von Campenhausen, B. Bornschein, R. Wick, et al. Prevalence and incidence of parkinson's disease in europe. *European Neuropsychopharmacology*, 15(4):473–490, 2005.

[5] Nicole Ball, David Teo, Arjune Chandra, Adam Chapman, and Edwin Choudhury. Environmental exposures and parkinson's disease: a review. *NeuroToxicology*, 70:132–141, 2019.

[6] Caroline M. Tanner, Freya Kamel, Gail W. Ross, et al. Rotenone, paraquat, and parkinson's disease. *Environmental Health Perspectives*, 119(6):866–872, 2011.

[7] Alberto Ascherio and Michael A. Schwarzschild. The epidemiology of parkinson's disease: risk factors and prevention. *Lancet Neurology*, 15(12):1257–1272, 2016.

[8] Christine Klein and Anette Westenberger. Genetics of parkinson's disease. *Cold Spring Harbor Perspectives in Medicine*, 2(1):a008888, 2012.

[9] Damien G. Healy, Mario Falchi, Seamus S. O'Sullivan, et al. Phenotype, genotype, and worldwide genetic penetrance of lrrk2-associated parkinson's disease. *Neurology*, 70(11):908–913, 2008.

[10] Lorraine V. Kalia and Anthony E. Lang. Parkinson's disease. *Lancet*, 386(9996):896–912, 2015.

[11] Maria Grazia Spillantini, Michael L. Schmidt, Virginia M.-Y. Lee, John Q. Trojanowski, Roy Jakes, and Michel Goedert. Alpha-synuclein in lewy bodies. *Nature*, 388(6645):839–840, 1998.

[12] Heiko Braak, Kelly Del Tredici, Udo Rüb, Rob A. I. de Vos, Enno N. H. Jansen Steur, and Eva Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of Aging*, 24(2):197–211, 2003.

[13] Roger L. Albin, Andrew B. Young, and John B. Penney. The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, 12(10):366–375, 1989.

[14] Mahlon R. DeLong. Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, 13(7):281–285, 1990.

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

[16] Nikolaus I. Bohnen, Kirk A. Frey, Stephanie Studenski, et al. Gait speed in parkinson disease correlates with cholinergic denervation. *Neurology*, 81(18):1611–1616, 2013.

[17] Glenda M. Halliday, James B. Leverenz, Julie S. Schneider, and Charles H. Adler. The neurobiological basis of nonmotor symptoms in parkinson's disease. *Movement Disorders*, 29(5):634–650, 2014.

[18] Nicole Exner, Bernhard Lutz, Christian Haass, and Konstanze F. Winklhofer. Mitochondrial dysfunction in parkinson's disease: molecular mechanisms and pathophysiological consequences. *EMBO Journal*, 31(14):3038–3062, 2012.

[19] Anthony H. V. Schapira and Michael Gegg. Mitochondrial contribution to parkinson's disease pathogenesis. *Parkinson's Disease*, 2011:159160, 2011.

[20] Rosario Di Maio, Patrick J. Barrett, Emily K. Hoffman, et al. Alpha-synuclein binds to tom20 and inhibits mitochondrial protein import in parkinson's disease. *Science Translational Medicine*, 8(342):342ra78, 2016.

[21] Thomas G. Beach, C. Lawrence White, Charles L. Hladik, et al. Validation of the braak staging scheme for parkinson's disease by a prospective study of autopsy cases. *Acta Neuropathologica*, 117(6):613–622, 2009.

[22] Ronald B. Postuma, Jean-François Gagnon, and Jacques Montplaisir. Prodromal parkinson's disease: the decade past, the decade to come. *Movement Disorders*, 27(5):641–650, 2012.

[23] Sebastian Heinzel, Daniela Berg, Thomas Gasser, Honglei Chen, Chen Yao, and Ronald B. Postuma. Update of the mds research criteria for prodromal parkinson's disease. *Movement Disorders*, 34(10):1464–1470, 2019.

[24] Julian M. Fearnley and Andrew J. Lees. Aging and parkinson's disease: substantia nigra regional selectivity. *Brain*, 114(5):2283–2301, 1991.

[25] Benjamin S. Connolly and Anthony E. Lang. Pharmacological treatment of parkinson disease: a review. *JAMA*, 311(16):1670–1683, 2014.

[26] J. Eric Ahlskog. Cheaper, simpler, and better: tips for treating seniors with parkinson disease. *Mayo Clinic Proceedings*, 86(12):1211–1216, 2011.

[27] C. Warren Olanow, Jose A. Obeso, and Fabrizio Stocchi. Continuous dopaminergic stimulation in the treatment of parkinson's disease: scientific rationale and clinical implications. *Lancet Neurology*, 5(8):677–687, 2006.

[28] Alim-Louis Benabid, Stephan Chabardes, John Mitrofanis, and Pierre Pollak. Deep brain stimulation of the subthalamic nucleus for the treatment of parkinson's disease. *Lancet Neurology*, 8(1):67–81, 2009.

[29] John G. Nutt, Bastiaan R. Bloem, Nir Giladi, Mark Hallett, Fay B. Horak, and Alice Nieuwboer. Freezing of gait: moving forward on a mysterious clinical phenomenon. *Lancet Neurology*, 10(8):734–744, 2011.

[30] Jean D. Schaafsma, Y. Balash, T. Gurevich, et al. Characterization of freezing of gait subtypes and the response of each to levodopa in parkinson's disease. *European Journal of Neurology*, 10(4):391–398, 2003.

[31] Nir Giladi, Haim Shabtai, Ephraim S. Simon, Stavit Biran, Jochanan Tal, and Amos D. Korczyn. Construction of freezing of gait questionnaire for patients with parkinsonism. *Parkinsonism & Related Disorders*, 6(3):165–170, 2000.

[32] Elke Heremans, Alice Nieuwboer, and Sarah Vercruysse. Freezing of gait in parkinson's disease: a persisting unsolved mystery. *Journal of the Neurological Sciences*, 331(1-2):194–200, 2013.

[33] Bastiaan R. Bloem, Jeffrey M. Hausdorff, Marjolein Visser, and Nir Giladi. Falls and freezing of gait in parkinson's disease: a review of two interconnected, episodic phenomena. *Movement Disorders*, 19(8):871–884, 2004.

[34] Nir Giladi, Michael P. McDermott, Stanley Fahn, et al. Freezing of gait in pd: Prospective assessment in the datatop cohort. *Neurology*, 56(12):1712–1721, 2001.

[35] Albertine H. Snijders and Bastiaan R. Bloem. Freezing of gait in parkinson's disease: a review of pathophysiology, risk factors and treatments. *Nature Reviews Neurology*, 12(10):657–670, 2016.

[36] James M. Shine, E. Matar, Peter B. Ward, et al. Freezing of gait in parkinson's disease is associated with functional decoupling between the cognitive control network and the basal ganglia. *Brain*, 136(12):3671–3681, 2013.

[37] Simon J. G. Lewis and James M. Shine. The next step: a common neural mechanism for freezing of gait. *Neuroscientist*, 22(1):72–82, 2016.

[38] Alice Nieuwboer, Lynn Rochester, Talia Herman, et al. Reliability of the new freezing of gait questionnaire: agreement between patients with parkinson's disease and their carers. *Gait & Posture*, 30(4):459–463, 2009.

[39] Christopher G. Goetz, Brent C. Tilley, Sherrie R. Shaftman, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170, 2008.

[40] Robert A. Hauser and Nick H. G. Holford. Quantitative description of motor fluctuations in parkinson's disease: time series analysis of self-reported motor states. *Movement Disorders*, 15(4):661–670, 2000.

[41] Shyamal Patel, David M. Sherrill, and Paolo Bonato. Wearable sensors and systems in parkinson's disease: a review of recent developments. *IEEE Reviews in Biomedical Engineering*, 14:18–35, 2021.

[42] Silvia Del Din, Alan Godfrey, and Lynn Rochester. Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and parkinson's disease: toward clinical and at home use. *IEEE Journal of Biomedical and Health Informatics*, 20(3):838–847, 2016.

[43] Catarina Godinho, Joana Domingos, Luís Cunha, et al. A systematic review of the characteristics and validity of monitoring technologies to assess parkinson's disease. *Journal of NeuroEngineering and Rehabilitation*, 13(1):24, 2016.

[44] Daniel Rodríguez-Martín, Albert Samà, Carlos Pérez-López, et al. Home detection of motor fluctuations in parkinson's disease patients based on inertial signals. *Sensors*, 17(11):2507, 2017.

[45] Luca Baratto, Silvia Caravita, Massimo F. Piepoli, and Michele Senni. Heart rate variability: a diagnostic and prognostic tool in cardiovascular disease and beyond. *European Heart Journal Supplements*, 24:D20–D24, 2022.

[46] Vera Ruonala, Andrej Meigal, Simo Rissanen, et al. Emg signal morphology and kinematic parameters in parkinson's disease during repetitive elbow flexion–extension task. *Clinical Neurophysiology*, 129(4):748–756, 2018.

[47] A. M. A. Handojoseno, J. M. Shine, T. N. Nguyen, et al. The prediction of freezing of gait in parkinson's disease using eeg and ecg signals. *IEEE Journal of Biomedical and Health Informatics*, 19(2):682–690, 2015.

[48] Ki H. Kim, Inhye Park, and J. Kim. A wearable plantar pressure monitoring system for gait analysis. *Sensors*, 19(17):3675, 2019.

[49] Marc Bächlin, Daniel Roggen, Meir Plotnik, Nira Inbar, Idan Meidan, Jeffrey M. Hausdorff, Nir Giladi, and Gerhard Tröster. Wearable assistant for parkinson's disease patients with the freezing of gait symptom. In *Proc. 4th International Conference on Pervasive Computing Technologies for Healthcare*, pages 58–61, 2010.

[50] Empatica Inc. E4 wristband: real-time physiological data streaming and storage (technical documentation). https://www.empatica.com/research/e4, 2018. Accessed 2025-11-28.

[51] Terry T. Um, Franz M. J. Pfister, Daniel Pichler, et al. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(2):1–17, 2017.

[52] Adrià Samà, Carlos Pérez-López, Daniel Rodríguez-Martín, Andreu Català, Joan Cabestany, Àngel Bayés, Sara Alcaine, Berta Mestre, Anna Prats, María C. Crespo, Patricia Quispe, Carla Gálvez-Barrón, Oriol de Fabregues, Adrià Moreno, and Alejandro Rodríguez-Molinero. A double-blind, randomized, controlled study of freezing of gait detection using accelerometers in parkinson's disease. *Journal of NeuroEngineering and Rehabilitation*, 18(1):40, 2021.

[53] Arnaud Delval, Luc Defebvre, Bastiaan R. Bloem, and Alice Nieuwboer. Gait analysis and freezing of gait in parkinson's disease: insights from wearable sensors and deep learning approaches. *Frontiers in Neurology*, 13:839203, 2022.

[54] N. L. W. Keijsers, M. W. I. M. Horstink, and S. C. A. M. Gielen. Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks. *Movement Disorders*, 21(1):34–44, 2006.

[55] Evanthia E. Tripoliti, Athanasios T. Tzallas, Markos G. Tsipouras, et al. Automatic detection of freezing of gait events in parkinson's disease patients: a pattern recognition approach. *Computer Methods and Programs in Biomedicine*, 110(1):12–26, 2013.

[56] Bjoern M. Eskofier, Seung Lee, Daniel Baron, et al. An overview of smart shoes in the context of parkinson's disease. *Sensors*, 16(11):1693, 2016.

[57] Simon T. Moore, Hamish G. MacDougall, and William G. Ondo. Ambulatory monitoring of freezing of gait in parkinson's disease. *Journal of Neuroscience Methods*, 167(2):340–348, 2008.

[58] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BioData Mining*, 13(1):11, 2020.

[59] Douglas G. Altman and J. Martin Bland. Diagnostic tests 3: receiver operating characteristic (roc) curves. *BMJ*, 309(6948):188, 1994.

[60] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers (ISWC)*, pages 1–10, 2016.

[61] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[62] Marco Matarazzo, Giacomo C. De Luca, Luca Palmerini, and Lorenzo Chiari. Transfer learning for the classification of parkinson's disease motor states from wearable sensor data. *Sensors*, 21(13):4381, 2021.

[63] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[64] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[65] Simon T. Moore, Marta Martínez-Manzanera, John K. Petersen, Carolin Curtze, Jeffrey M. Hausdorff, and Bastiaan R. Bloem. Recent advances in the detection and prediction of freezing of gait in parkinson's disease. *NPJ Parkinson's Disease*, 9(1):52, 2023.

[66] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[67] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[68] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[69] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, and Noam Shazeer. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

[70] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[71] Wensheng Gan, Zhenyao Ning, Zhenlian Qi, and Philip S. Yu. Mixture of experts (moe): A big data perspective. *Information Fusion*, 127:103664, 2025.

[72] Lala Shakti Swarup Ray, Qingxin Xia, Vitor Fortes Rey, Kaishun Wu, and Paul Lukowicz. Improving imu based human activity recognition using simulated multimodal representations and a moe classifier. *Frontiers in Computer Science*, 7, 2025.

[73] Debaditya Roy and Sarunas Girdzijauskas. Mixing temporal experts for human activity recognition. In *2022 Swedish Artificial Intelligence Society Workshop (SAIS)*, 2022.

[74] Fahrurrozi Rahman, Martin Schiemer, Andrea Rosales Sanabria, and Juan Ye. Continual learning in sensor-based human activity recognition with dynamic mixture of experts. *Pattern Recognition Letters*, 2025.

[75] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025. OpenReview paper page: {https://openreview.net/forum?id=e1wDDFmlVu}.

[76] Francesca Demrozi, Raffaella Bacchin, Stefano Tamburin, and et al. Human activity recognition for parkinson's disease using wearable sensors: A systematic review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2733–2746, 2020.

[77] Anjali Singh, Nikhil Gupta, and et al. Cnn-based human activity recognition: Applications and challenges. *Biomedical Signal Processing and Control*, 85:104933, 2023.

[78] Mykael V. Albert, Sandra Toledo, Matthew Shapiro, and Konrad Kording. Using mobile phones for activity recognition in parkinson's patients. *Frontiers in Neurology*, 3:158, 2012.

[79] V. Aswar, A. Khedkar, and et al. Machine learning-based activity recognition in parkinson's disease using inertial sensors. *Biomedical Signal Processing and Control*, 82:104558, 2023.

[80] L. Davidashvilly, A. Khelil, and et al. Domain-adversarial neural networks for parkinson's disease activity recognition. *Sensors*, 24(3):912, 2024.

[81] Rafael Gonçalves and et al. Deep learning-based recognition of daily activities in parkinson's disease using a single inertial sensor. *Sensors*, 25(1):112, 2025.

[82] Luis Sigcha, Luigi Borzì, Federica Amato, Irene Rechichi, Carlos Ramos-Romero, Andrés Cárdenas, Luis Gascó, and Gabriella Olmo. Deep learning and wearable sensors for the diagnosis and monitoring of parkinson's disease: A systematic review. *Expert Systems with Applications*, 229:120541, 2023.

[83] Bo Yuan, Tiago Nunes, and et al. A study on fine-tuning recurrent neural networks for freezing of gait prediction in parkinson's disease. In *Proceedings of the IEEE Engineering in Medicine and Biology Conference*, pages 5555–5559, 2020.

[84] Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M Hausdorff, Nir Giladi, and Gerhard Troster. Wearable assistant for parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2009.

[85] Vikas Torvi, S. Bhattacharya, and et al. Deep learning approaches for personalized freezing of gait detection in parkinson's disease. *IEEE Journal of Biomedical and Health Informatics*, 22(6):1762–1771, 2018.

[86] N. Naghavi and E. Wade. Towards real-time prediction of freezing of gait in patients with parkinson's disease: A novel deep one-class classifier. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1726–1736, 2022.

[87] Shubham Bajpai, Rahul Goyal, and et al. Multimodal deep learning for freezing of gait detection using imu and eeg data. *Biomedical Signal Processing and Control*, 84:104848, 2023.

[88] Florent Imbert, Eric Anquetil, Yann Soullard, and Romain Tavenard. Mixture-of-experts for handwriting trajectory reconstruction from imu sensors. *Pattern Recognition*, 2024.

[89] Benjamin J. Choi and Jian Liu. A brainwave- and head-gesture-controlled neuroprosthesis using a mixture-of-experts learning framework. *Journal of Neural Engineering*, 22:016031, 2025.

[90] Xiangnan Xu, Michal Lubomski, Andrew J. Holmes, Carolyn M. Sue, Ryan L. Davis, Samuel Muller, and Jean Y. H. Yang. Nemoe: a nutrition aware regularized mixture of experts model to identify heterogeneous diet–microbiome–host health interactions. *Microbiome*, 11:51, 2023.

[91] Konstantinos Sechidis, Riccardo Fusaroli, Juan Rafael Orozco-Arroyave, Detlef Wolf, and Yan-Ping Zhang. A machine learning perspective on the emotional content of parkinsonian speech. *Artificial Intelligence in Medicine*, 115:102061, 2021.

[92] Rizhi Ding, Hui Lu, and Manhua Liu. Denseformer-moe: A dense transformer foundation model with mixture of experts for multi-task brain image analysis. *IEEE Transactions on Medical Imaging*, 44(10):4037–4050, 2025.

[93] Michele Tebaldi, Luigi Borzì, Gabriella Olmo, Rosalba Giugno, Graziano Pravadelli, and Florenc Demrozi. Exploring parkinson's disease datasets: Key findings, challenges, and recommendations for motor symptom analysis. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7, 2025.

[94] Luigi Borzì, Florenc Demrozi, Ruggero Angelo Bacchin, Cristian Turetta, Michele Tebaldi, Luis Sigcha, Samaneh Zolfaghari, Domiziana Rinaldi, Giuliana Fazzina, Giulio Balestro, Alessandro Picelli, Graziano Pravadelli, Gabriella Olmo, Stefano Tamburin, Leonardo Lopiano, and Carlo Alberto Artusi. A multi-level annotated sensor dataset of gait freezing manifestations and severity in parkinson's disease. *Scientific Data*, 13:305, 2026.

[95] Tebaldi Michele Borzì, Luigi and Florenc Demrozi. Ai-based recognition of daily activities in parkinson's disease using wearable inertial sensors. In *To appear in International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, pages XXX–XXX. Springer, 2025.

[96] Michele Tebaldi, Graziano Pravadelli, Florenc Demrozi, Rosalba Giugno, and Cristian Turetta. Enhancing freezing of gait detection in parkinson's through fine-tuned deep learning models. In *2024 IEEE International Conference on Digital Health (ICDH)*, pages 87–94, 2024.

[97] Anthony J Anderson, David Eguren, Michael A Gonzalez, Naima Khan, et al. Weargait-pd: An open-access wearables dataset for gait in parkinson's disease and age-matched controls. *medRxiv*, pages 2024–09, 2024.

[98] Wei Zhang, Zhuokun Yang, Hantao Li, Debin Huang, Lipeng Wang, Yanzhao Wei, Lei Zhang, Lin Ma, Huanhuan Feng, Jing Pan, et al. Multimodal data for the detection of freezing of gait in parkinson's disease. *Scientific data*, 9(1):606, 2022.

[99] Sinziana Mazilu, Alberto Calatroni, Eran Gazit, Anat Mirelman, Jeffrey M. Hausdorff, and Gerhard Tröster. Prediction of freezing of gait in parkinson's from physiological wearables: An exploratory study. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1843–1854, 2015.

[100] Johanna O'Day, Marissa Lee, Kirsten Seagers, Shannon Hoffman, et al. Assessing inertial measurement unit locations for freezing of gait detection and patient preference. *Journal of neuroengineering and rehabilitation*, 19(1):20, 2022.

[101] Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, et al. Wearable assistant for parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2010.

[102] Caroline Ribeiro De Souza, Runfeng Miao, Júlia Ávila De Oliveira, Andrea Cristina De Lima-Pardini, Débora Fragoso De Campos, et al. A public data set of videos, inertial measurement unit, and clinical scales of freezing of gait in individuals with parkinson's disease during a turning-in-place task. *Frontiers in Neuroscience*, 16:832463, 2022.

[103] Daniel Rodríguez-Martín, Albert Samà, Carlos Pérez-López, Andreu Català, Joan M Moreno Arostegui, Joan Cabestany, Àngels Bayés, Sheila Alcaine, Berta Mestre, Anna Prats, et al. Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. *PloS one*, 12(2):e0171764, 2017.

[104] Luigi Borzì, Marilena Varrecchia, Stefano Sibille, Gabriella Olmo, Carlo Alberto Artusi, Margherita Fabbri, Mario Giorgio Rizzone, Alberto Romagnolo, Maurizio Zibetti, and Leonardo Lopiano. Smartphone-based estimation of item 3.8 of the mds-updrs-iii for assessing leg agility in people with parkinson's disease. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:140–147, 2020.

[105] Addison Howard, amit salomon, eran gazit, Jeff Hausdorff, Leslie Kirsch, Maggie, Pieter Ginis, Ryan Holbrook, and Yasir F Karim. Parkinson's freezing of gait prediction. https://kaggle.com/competitions/tlvmc-parkinsons-freezing-gait-prediction, 2023. Kaggle.

[106] Alexandros T Tzallas, Markos G Tsipouras, Georgios Rigas, Dimitrios G Tsalikakis, et al. Perform: a system for monitoring, assessment and management of patients with parkinson's disease. *Sensors*, 14(11):21329–21357, 2014.

[107] Jean-Francois Daneault, Gloria Vergara-Diaz, Federico Parisi, Chen Admati, Christina Alfonso, Matilde Bertoli, Edoardo Bonizzoni, Gabriela Ferreira Carvalho, Gianluca Costante, Eric Eduardo Fabara, et al. Accelerometer data collected with a minimum set of wearable sensors from subjects with parkinson's disease. *Scientific Data*, 8(1):48, 2021.

[108] Catherine Morgan, Emma L Tonkin, Alessandro Masullo, Ferdian Jovan, et al. A multimodal dataset of real world mobility activities in parkinson's disease. *Scientific data*, 10(1):918, 2023.

[109] Franz MJ Pfister, Terry Taewoong Um, Daniel C Pichler, Jann Goschenhofer, Kian Abedinpour, Muriel Lang, Satoshi Endo, Andres O Ceballos-Baumann, Sandra Hirche, Bernd Bischl, et al. High-resolution motor state detection in parkinson's disease using convolutional neural networks. *Scientific reports*, 10(1):5860, 2020.

[110] Luigi Borzì, Gabriella Olmo, Carlo Alberto Artusi, Margherita Fabbri, Mario Giorgio Rizzone, Alberto Romagnolo, Maurizio Zibetti, and Leonardo Lopiano. A new index to assess turning quality and postural stability in patients with parkinson's disease. *Biomedical signal processing and control*, 62:102059, 2020.

[111] Matthias Löhle, Jonathan Timpka, Alexander Bremer, Hamid Khodakarami, Florin Gandor, Malcom Horne, Georg Ebersbach, Per Odin, and Alexander Storch. Application of single wrist-wearable accelerometry for objective motor diary assessment in fluctuating parkinson's disease. *npj Digital Medicine*, 6(1):194, 2023.

[112] Claudia Mazzà, Lisa Alcock, Kamiar Aminian, Clemens Becker, Stefano Bertuletti, Tecla Bonci, Philip Brown, Marina Brozgol, Ellen Buckley, Anne-

Elie Carsin, et al. Technical validation of real-world monitoring of gait: a multicentric observational study. *BMJ open*, 11(12):e050785, 2021.

[113] Luigi Borzì, Marilena Varrecchia, Gabriella Olmo, Carlo Alberto Artusi, Margherita Fabbri, Mario Giorgio Rizzone, Alberto Romagnolo, Maurizio Zibetti, and Leonardo Lopiano. Home monitoring of motor fluctuations in parkinson's disease patients. *Journal of Reliable Intelligent Environments*, 5(3):145–162, 2019.

[114] Daniel Rodríguez-Martín, Joan Cabestany, Carlos Pérez-López, Marti Pie, et al. A new paradigm in parkinson's disease evaluation with wearable medical devices: A review of stat-ontm. *Frontiers in Neurology*, 13, 2022.

[115] Florenc Demrozi, Cristian Turetta, Philipp H Kindt, Fabio Chiarani, et al. A low-cost wireless body area network for human activity recognition in healthy life and medical applications. *IEEE Trans Emerg Top Comput*, 11:839–850, 2023.

[116] Rhoda Weiss-Lambrou Louise Demers and Bernadette Ska. Development of the quebec user evaluation of satisfaction with assistive technology (quest). *Assist Technol*, 8:3–13, 1996.

[117] Nir Giladi and Alice Nieuwboer. Understanding and treating freezing of gait in parkinsonism, proposed working definition, and setting the stage. *Mov Disord*, 23:423–425, 2008.

[118] Riccardo Ribeiro, Hugo Cachitas, Carlos Mao de Ferro, and Luis Teixeira. Python video annotator, 2019.

[119] Rakesh Aggarwal and Priya Ranganathan. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspect Clin Res*, 7:187–190, 2016.

[120] Leslie Gross Portney, Mary P Watkins, et al. *Foundations of clinical research: applications to practice*, volume 892. Pearson/Prentice Hall Upper Saddle River, NJ, 2009.

[121] Luigi Borzì, Florenc Demrozi, Ruggero Bacchin, Cristian Turetta, Michele Tebaldi, Luis Sigcha, Samaneh Zolfagharian, Domiziana Rinaldi, Giuliana Fazzina, Giulio Balestro, Alessandro Picelli, Graziano Pravadelli, Gabriella

OLMO, Stefano Tamburin, Leonardo Lopiano, and Carlo Alberto Artusi. Fog-star: Freezing of gait severity, tasks, activities, and ratings, August 2025.

[122] Luis Sigcha, Luigi Borzì, and Gabriella Olmo. Deep learning algorithms for detecting freezing of gait in parkinson's disease: A cross-dataset study. *Expert Systems with Applications*, 255:124522, 2024.

[123] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[124] Cristian Turetta, Florenc Demrozi, and Graziano Pravadelli. B-har: an open-source baseline framework for in-depth study of human activity recognition datasets and workflows. *IEEE Access*, 2024.

[125] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[126] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 3995–4001. AAAI Press, 2015.

[127] Francisco J. Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[128] Bo Sun, Meiqin Liu, Ronghao Zheng, and Senlin Zhang. Attention-based lstm network for wearable human activity recognition. In *2019 Chinese Control Conference (CCC)*, pages 8677–8682, 2019.

[129] Shengjiao Zhao, Haikun Wei, and Kanjian Zhang. Deep bidirectional gru network for human activity recognition using wearable inertial sensors. In *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 238–242, 2022.

[130] Saif Mahmud, M Tanjid Hasan Tonmoy, Kishor Kumar Bhaumik, A K M Mahbubur Rahman, M Ashraful Amin, Mohammad Shoyaib, Muhammad Asif Hossain Khan, and Amin Ahsan Ali. Human activity recognition from wearable sensor data using self-attention, 2020.

[131] Luigi Borzì, Florenc Demrozi, Ruggero Angelo Bacchin, Cristian Turetta, Luis Sigcha, Domiziana Rinaldi, Giuliana Fazzina, Giulio Balestro, Alessandro Picelli, Graziano Pravadelli, et al. Freezing of gait detection: The effect of sensor type, position, activities, datasets, and machine learning model. *Journal of Parkinson's Disease*, page 1877718X241302766, 2025.

[132] Florenc Demrozi, Cristian Turetta, and Graziano Pravadelli. B-HAR: an open-source baseline framework for in depth study of human activity recognition datasets and workflows. *arXiv preprint arXiv:2101.10870*, 2021.

[133] Florenc Demrozi, Cristian Turetta, and Graziano Pravadelli. Fostering human activity recognition workflows: An open-source baseline framework. In *2023 IEEE International Conference on Digital Health (ICDH)*, pages 75–80. IEEE, 2023.

[134] Jonatan Fridolfsson, Mats Börjesson, Christoph Buck, örjan Ekblom, Elin Ekblom-Bak, Monica Hunsberger, Lauren Lissner, and Daniel Arvidsson. Effects of frequency filtering on intensity and noise in accelerometer-based physical activity measurements. *Sensors*, 19(9), 2019.

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

[136] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[137] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[138] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Transfer learning for time series classification. In *2018 IEEE international conference on big data (Big Data)*, pages 1367–1376. IEEE, 2018.

[139] Honglei Zhang, Sidi Mohamed Sid'El Moctar, Sofiane Boudaoud, and Imad Rida. A comprehensive review of semg-imu sensor fusion for upper limb movements pattern recognition. *Information Fusion*, 125:103422, 2026.

[140] Angelo Maria Sabatini. Estimating three-dimensional orientation of human body parts by inertial/magnetic sensing. *Sensors*, 11(2):1489–1525, 2011.

[141] Peter H Veltink, HansB J Bussmann, Wiebe De Vries, WimL J Martens, and Rob C Van Lummel. Detection of static and dynamic activities using uniaxial accelerometers. *IEEE Transactions on Rehabilitation Engineering*, 4(4):375–385, 1996.

[142] Justin J. Kavanagh and Hylton B. Menz. Accelerometry: A technique for quantifying movement patterns during walking. *Gait & Posture*, 28(1):1–15, 2008.

[143] Michael W. Whittle. *Gait Analysis: An Introduction*. Butterworth-Heinemann, 2007.

[144] Oscar D. Lara and Miguel A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013.

[145] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 10(3), 2015.

[146] Y. et al. Sun. Imu-based quantitative assessment of stroke from gait. *Science Reports*, 2025.

[147] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[148] Tinghuai Huang, Meng Li, and Jianwei Huang. Recent trends in wearable device used to detect freezing of gait and falls in people with parkinson's disease: A systematic review. *Frontiers in aging neuroscience*, 15:1119956, 2023.

[149] A Keogh, L Alcock, P Brown, E Buckley, et al. Acceptability of wearable devices for measuring mobility remotely: Observations from the mobilise-d technical validation study. *Digital Health*, 9, 2023.

[150] C. Moreau, T. Rouaud, D. Grabli, et al. Overview on wearable sensors for the management of parkinson's disease. *npj Parkinsons Dis.*, 9:153, 2023.

[151] Jennifer C. Goldsack, Andrea Coravos, Jessie P. Bakker, Brinnae Bent, Ariel V. Dowling, Cheryl Fitzer-Attas, Alan Godfrey, Job G. Godino, Ninad Gujar, Elena Izmailova, Christine Manta, Barry Peterson, Benjamin Vandendriessche, William A. Wood, Ke Li Wang, and Jessilyn Dunn. Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets). *npj Digital Medicine*, 3(1):55, 2020.

[152] Jessie P. Bakker, Roland Barge, Jacob Centra, Bryan Cobb, Chas Cota, Christine C. Guo, Bert Hartog, Nathalie Horowicz-Mehler, Elena S. Izmailova, Nikolay V. Manyakov, Samantha McClenahan, Stéphane Motola, Smit Patel, Oana Paun, Marian Schoone, Emre Sezgin, Thomas Switzer, Animesh Tandon, Willem van den Brink, Srinivasan Vairavan, Benjamin Vandendriessche, Bernard Vrijens, and Jennifer C. Goldsack. V3+ extends the v3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. *npj Digital Medicine*, 8:51, 2025.

[153] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[154] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[155] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[156] Andrea Rosales Sanabria, Franco Zambonelli, Simon Dobson, and Juan Ye. Contrasgan: Unsupervised domain adaptation in human activity recognition via adversarial and contrastive learning. *Pervasive and Mobile Computing*, 78:101477, 2021.

[157] Wang Lu, Jindong Wang, Yiqiang Chen, Sinno Jialin Pan, Chunyu Hu, and Xin Qin. Semantic-discriminative mixup for generalizable sensor-based cross-domain activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–27, 2022.

[158] Saurav Jha, Martin Schiemer, Franco Zambonelli, and Juan Ye. Continual learning in sensor-based human activity recognition: An empirical benchmark analysis. *Information Sciences*, 575:1–21, 2021.

[159] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.