Evaluating Temporal Windows and Subject-Specific Fine-Tuning for Freezing of Gait Detection with Classical Machine Learning Models

Clebson I. S. Silva
Signal Processing Laboratory
ITEC (UFPA)
Belém-PA, Brazil
clebson.silva@itec.ufpa.br

Ronaldo F. Zampolo
Signal Processing Laboratory
ITEC (UFPA)
Belém-PA, Brazil
zampolo@ufpa.br

Antônio Pereira, Jr
Signal Processing Laboratory
ITEC (UFPA)
Belém-PA, Brazil
apereira@ufpa.br

***Abstract*—Freezing of Gait (FoG) is a critical motor symptom in Parkinson's disease and a prime target for wearable-based detection. While global models trained across subjects offer scalability, they may fail to generalize to individuals with distinct gait signatures. This study investigates the effectiveness of classical machine learning models, explores the impact of temporal window size, and evaluates whether subject-specific fine-tuning improves detection performance. Using data from a dataset with 35 subjects performing a rotation-eliciting task, we show that ensemble models with 4–5 s windows achieve F1-scores up to 0.92. The results of the importance analysis of features highlight the role of signals from the SI-axis derived from a gyroscope. Fine-tuning with 10 windows yields significant classification gains in selected individuals but only marginal average improvement ($\triangle$F1 = +0.011), raising questions about its general viability. We argue for adaptive deployment strategies that combine selective personalization and task-aware design to balance accuracy, interpretability, and efficiency in wearable systems.**

***Index Terms*—Freezing of Gait (FoG), Parkinson's Disease, Inertial Measurement Unit (IMU), Machine Learning, Personalization, Fine-Tuning, Temporal Windowing**

I. Introduction

Freezing of Gait (FoG) is a disabling and episodic motor symptom affecting up to 60% of individuals with Parkinson's disease (PD). It is associated with increased risk of falls, reduced mobility, and lower quality of life. Wearable systems based on inertial measurement units (IMUs) have emerged as promising tools for continuous FoG detection, particularly in free-living or clinical settings [1], [2].

Machine learning models have shown considerable success in processing IMU data to detect FoG events [1], [2]. Most existing solutions rely on global models trained across multiple subjects, enabling scalability and generalization [1], [3]. However, such models may fail to account for inter-subject variability stemming from differences in gait patterns, disease stage, sensor placement, and medication effects.

Recent research has explored personalization strategies to address this limitation, including transfer learning, fine-tuning with few-shot examples, and adaptive modeling [2], [4]. However, the field lacks a systematic evaluation of the trade-offs involved—particularly in terms of performance gain, data requirements, and computational cost [4].

While some studies have employed deep learning architectures for FoG detection [1], [2], [5], [6], few have investigated the effectiveness of lightweight personalization techniques in real-world conditions [4]. This study seeks to fill that gap by comparing global and fine-tuned models across subjects, evaluating their effectiveness and feasibility for deployment in wearable systems.

In addition, the choice of temporal window size, an often overlooked factor, can significantly affect classifier performance. We examine this aspect in the context of a structured rotation-based task ("turning-in-place"), which has been shown to reliably elicit FoG events [7]. Using a public dataset with IMU recordings from 35 PD subjects, we extracted 99 temporal features per window in multiple domains.

Despite recent advances in FoG detection methods, few studies have systematically evaluated the impact of different personalization strategies and window size choices in embedded scenarios, where computational constraints and limited adaptation data are crucial considerations. Thus, this work aims to fill this gap by comprehensively analyzing the performance of classical models under different window configurations and lightweight personalization, providing practical insights for deploying FoG detection models on low-cost wearable devices.

II. Methods

A. Dataset and Preprocessing

The public dataset is available on the Figshare platform (www.figshare.com) [8]. The dataset consists of data from 35 individuals diagnosed with idiopathic Parkinson's Disease (PD) and a history of Freezing of Gait (FoG), comprising 16 women and 19 men, aged between 44 and 84 years. All participants were taking PD medication during data collection.

Episodes of FoG were annotated by two movement disorder experts via synchronized video analysis using ELAN software [9], ensuring high-quality event labeling.

Participants performed a "turning-in-place" task, in which they walked in circles alternating directions at a

self-selected pace for two minutes. Each subject completed three experimental sessions spaced one month apart.

During the task, subjects wore a *Physilog* 5 inertial measurement unit (IMU) placed near the heel on their more affected side. The IMU recorded tri-axial linear accelerations and angular velocities at a sampling rate of 128 Hz. Raw IMU signals (accelerometer, gyroscope) were filtered using a 4th-order Butterworth filter with a cutoff frequency of 60 Hz. The filtered signals were then segmented into non-overlapping time windows. We investigated the effect of multiple window lengths (1–6 seconds) on model performance, ultimately selecting 2-second windows as a balance between resolution and event coverage. Each window received a binary label (FoG or non-FoG), based on whether the annotated FoG occupied at least 75 % of the window duration. The 99 features were:

- **Temporal Domain Features:** Mean, standard deviation, max, min, RMS, skewness, kurtosis, zero-crossings, and Signal Magnitude Area (SMA) from all six IMU axes.
- **Spectral Domain Features:** Using FFT, mean and max magnitude in the frequency domain, and spectral power in the 3–8 Hz band [6]. Additionally, a power ratio was computed between the 3–8 Hz band and the rest.
- **Wavelet Features:** Energy and standard deviation of discrete wavelet transform (DWT) coefficients using 'db4' at level 4.
- **Correlation Features:** Pearson correlations between orthogonal axis pairs (ACC and GYR).
- **Ground Reaction Force (GRF):** A kinematic feature estimated from acceleration in the vertical axis and subject weight (mean, standard deviation, and max GRF per window).

All features were standardized using z-score normalization (mean and standard deviation computed from the training set).

All experiments were executed on a laptop equipped with an Intel Core i5-13450HX (10-core, cache de 20MB, até 4.6GHz) and 16GB RAM.

In the following section, we describe the experimental setup used to compare global and personalized models.

Figure 1 provides a high-level summary of the signal processing and modeling stages described above, highlighting both the global and personalized model evaluation branches.

B. Experimental Protocol

In this study, two main sets of experiments were conducted to investigate different aspects of Freezing of Gait (FOG) detection using data from Inertial Measurement Units (IMUs).

1) Evaluation of Classical Classifiers and Temporal Windowing

The primary objective of this experiment was to examine the influence of temporal window size and the performance of classical machine learning classifiers for FOG detection. Four classical classifiers were evaluated: Support Vector Classifier (SVC), Random Forest (RForest), K-Nearest Neighbors (KNN), and XGBoost.

The experimental methodology involved segmenting the IMU signals into non-overlapping temporal windows. The effect of window duration was analyzed by testing windows ranging from 1 to 6 seconds. For window labeling, a segment was considered as FOG or non-FOG if the proportion of time annotated as FOG within the window was greater than or equal to 75

The performance of each classifier at each tested window size was evaluated using Leave-One-Subject-Out (LOSO) cross-validation. In this scheme, data from one subject are held out for testing, while the data from all other subjects are used to train the model. Model performance was quantified using a comprehensive set of metrics, including Accuracy, F1-score (F1), Sensitivity to FOG episodes (Sens.), Specificity to non-FOG segments (Spec.), and Area Under the Curve (AUC).

2) Evaluation of Global vs. Personalized Models

The second set of experiments focused on comparing the performance of a globally trained model with a fine-tuned (personalized) version for each individual subject in the FOG detection task. A Random Forest model with 100 trees was used for this comparison.

In this experiment, non-overlapping temporal windows of 2 seconds in duration were employed. Features were extracted from each window, covering temporal, frequency, and wavelet domains.

The Global model was evaluated using Leave-One-Subject-Out (LOSO) cross-validation, with the model trained on data from 34 subjects and tested on the remaining subject, with the process repeated for each of the 35 subjects.

For the Personalized approach, the previously trained Global model was fine-tuned individually for each of the 35 subjects. This fine-tuning process was performed using a small subject-specific dataset composed of only 10 labeled windows (5 FOG and 5 non-FOG). After fine-tuning, the performance of the Personalized model was evaluated using the remaining subject's data as the test set.

III. Results

A. Classifier Performance and Feature Ranking

In this subsection, we present the findings of Experiment 1. Table I lists the metrics Accuracy, F1-score, Sensitivity, Specificity, and AUC for temporal windows of 1 to 6 s. Overall performance improves with increasing window size, with Random Forest performing best at 4 s (F1 = 0.9179) and XGBoost at 6 s (F1 = 0.9207). The evolution of sensitivity as a function of window duration is illustrated in Figure 2. Additionally, the ranking of the top 30 features computed via Mutual Information is shown in Figure 3, highlighting the predominance of gyroscope-derived variables along the SI axis.

B. Global vs. Personalized Model Performance and Fine-Tuning Time

A total of 26 subjects were included after excluding noisy or incomplete data. The global model achieved a macro-average F1-score of 0.864, while the fine-tuned models reached 0.875, resulting in a mean ΔF1 of +0.011.

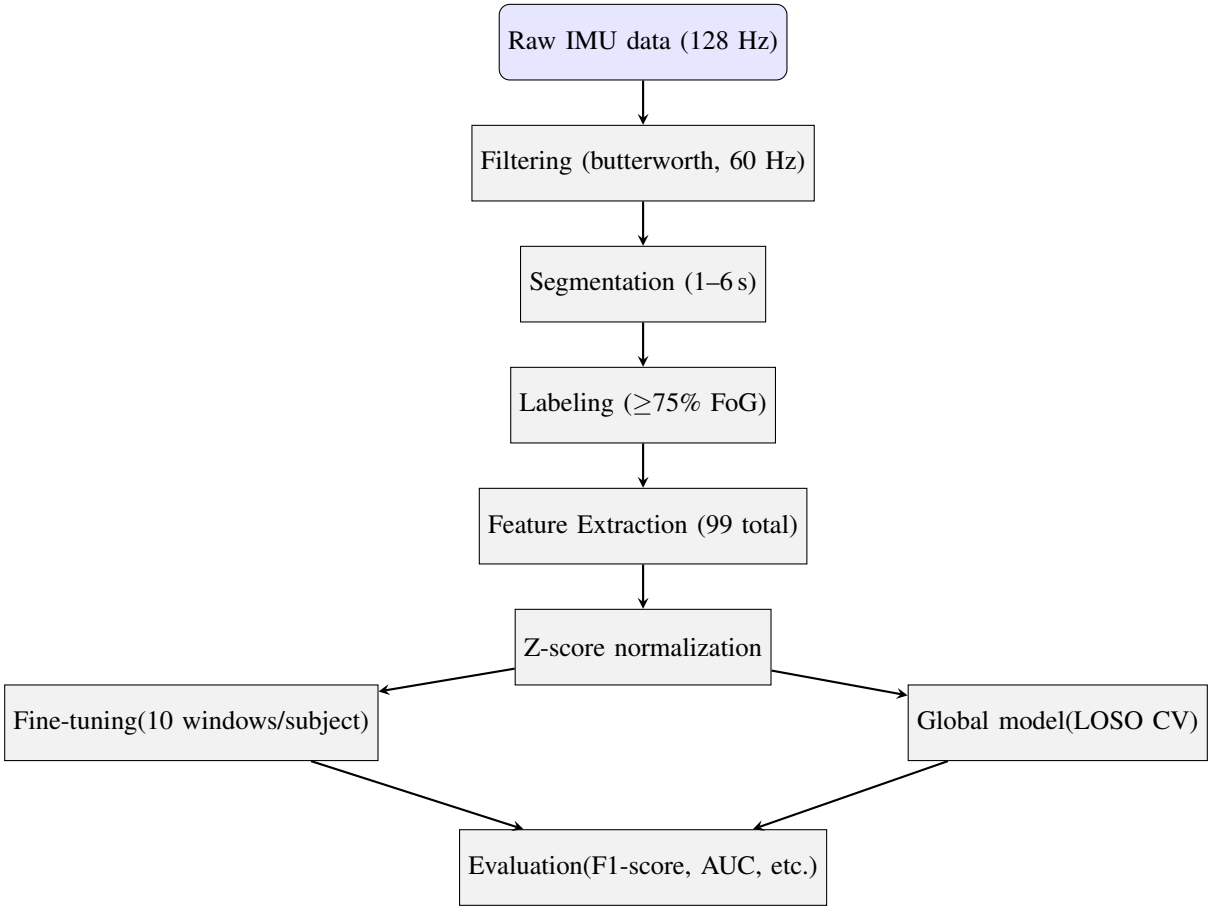

Fig. 1: Overview of the FoG detection pipeline, from raw IMU acquisition to evaluation of global and personalized models.

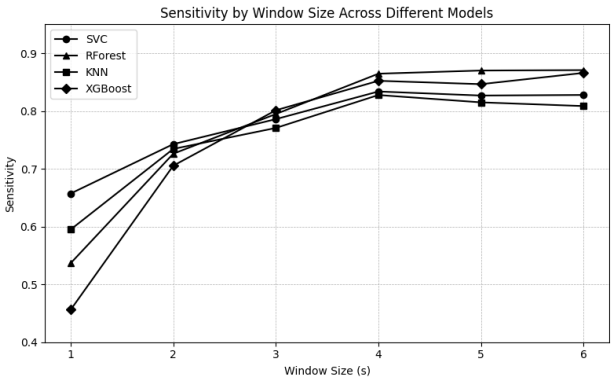

Fig. 2: Evolution of classifier sensitivity as a function of temporal window size.

Performance improved in 38 % of subjects—with Subjects 6, 7, and 15 exhibiting gains greater than +0.5 (each reaching an F1 of 1.0)—and declined in 62 %, the largest reductions being ΔF1 = -0.26 for Subject 2 and ΔF1 = -0.25 for Subject 31.

The average fine-tuning time was 3.8 s per subject (SD $\pm$ 0.4 s). Detailed per-subject F1-scores, ΔF1 values, and training times are provided in Table II. Figure 4 summarizes individual gains and losses.

IV. Discussion

The results of Experiment 1 show the decisive impact of temporal window size on FOG detection using classical classifiers. Larger windows (4–5 s) enabled better capture of the complex dynamics of FOG episodes, reflected in F1-scores of up to 0.92 and AUC values of 0.92. In particular, the Random Forest with a 5 s window achieved the best balance between sensitivity and specificity, while XGBoost reached its highest sensitivity with a 6 s window. These findings are consistent with the literature consensus, which indicates durations of 2–4 s as a trade-off between latency and FOG pattern coverage [10], [11], but suggest that slightly longer durations may enhance performance in 360° turning tasks. Furthermore, the feature ranking by Mutual Information revealed a predominance of gyroscope-derived variables on the SI axis, reinforcing the importance of capturing vertical ankle motion to discriminate freezing episodes [12].

In comparison, previous threshold-based or linear model approaches have reported sensitivities around 73 % and specificities of 82 % using only spectral features [13], whereas studies employing leave-one-subject-out validation tend to show performance drops in real-world generalization scenarios [10]. Our classical models, powered by a comprehensive set of 99 features, surpass these benchmarks by balancing high accuracy with low computational cost—a crucial factor for low-power wearable devices.

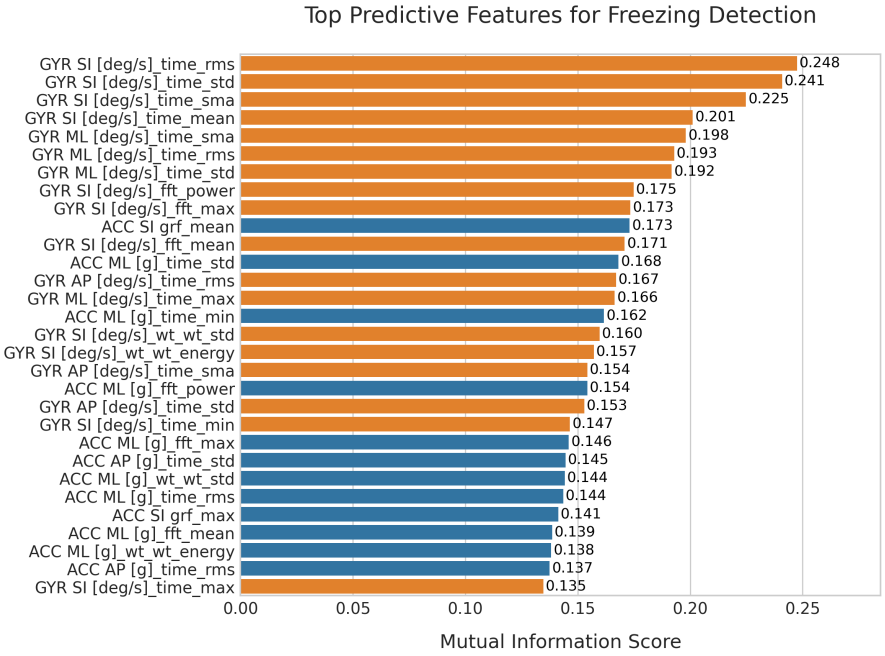

Fig. 3: Importance ranking of the top 30 features for freezing-of-gait detection, computed via Mutual Information. Features are derived from the gyroscope (GYR) and accelerometer (ACC) along the superior–inferior (SI), medio–lateral (ML), and antero–posterior (AP) axes.

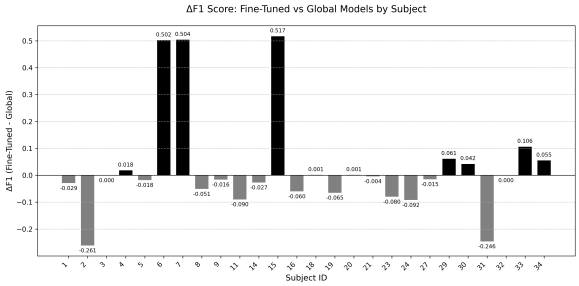

Fig. 4: Summary of global and fine-tuned F1-scores per subject, along with training time.

Experiment 2 illustrated the complexity of classifier personalization, although subject-specific fine-tuning yielded significant gains for some individuals (up to +0.52 in F1-score), most subjects exhibited neutral or even declining performance. The average gain of only +0.011 in F1-score suggests occasional benefits but raises concerns about the feasibility of universal personalization in embedded environments, where a few extra seconds of training may be critical. The paired analysis confirms this impression: the mean ΔF1 gain of 0.029 was not statistically significant (95% CI $[-0.049, 0.107]$; $t_{25} = 0.762$, $p = 0.453$; Wilcoxon $W = 132.5$, $p = 0.617$). These results align with meta-analyses emphasizing performance degradation in uncontrolled environments and point to adaptive strategies—such as validation-based triggers or hybrid ensemble models—to conserve resources and mitigate overfitting [14].

Additionally, the structured task of turn-in-place, recognized as a strong FOG elicitor in the review by Ribeiro De Souza et al. [7], underscores the need for protocol-specific evaluations. In clinical scenarios, selectively applying fine-tuning only to subjects whose global performance is unsatisfactory may optimize the trade-off between robustness, interpretability, and computational cost. Future investigations should explore regularization techniques, Bayesian fine-tuning, or meta-learning to enable safer personalization, as well as extend modeling efforts toward FOG prediction using sequential architectures such as LSTMs or Transformers.

V. Conclusion

This study demonstrates that classical ensemble models, when paired with adequately sized temporal windows ($\geq$ 4 s) and comprehensive feature sets, can rival the performance of deep learning approaches for Freezing of Gait detection, while maintaining lower computational complexity. The impact of window size was shown to be critical, with larger windows improving sensitivity to the complex temporal patterns of FoG.

Subject-specific fine-tuning offered substantial benefits in a subset of participants, with F1-score improvements exceeding +0.5 in select cases. However, the average gain among all subjects was minimal (ΔF1 = +0.011), and in several instances performance decreased, highlighting the risks of indiscriminate personalization in embedded systems with limited resources.

Given these findings, we propose selective personalization strategies triggered by performance thresholds, and task-specific protocols such as turning-in-place to enhance detection consistency. Future work should explore meta-learning, Bayesian fine-tuning, and sequential models (e.g., LSTMs, Transformers) for safer, more adaptive FoG detection and prediction in real-world settings.

TABLE I: Model performance by time window. Best value per metric in blue; worst value in red.

Window	Model	Accuracy	F1	Sens.	Spec.	AUC
1s	SVC	0.877	0.823	0.657	0.949	0.803
	RForest	0.854	0.777	0.537	0.958	0.748
	KNN	0.832	0.764	0.595	0.9102	0.753
	XGBoost	0.8155	0.7168	0.4561	0.933	0.6946
2s	SVC	0.912	0.872	0.743	0.965	0.854
	RForest	0.906	0.864	0.726	0.963	0.845
	KNN	0.894	0.849	0.734	0.944	0.839
	XGBoost	0.897	0.850	0.705	0.957	0.831
3s	SVC	0.923	0.889	0.786	0.965	0.875
	RForest	0.925	0.891	0.795	0.964	0.879
	KNN	0.909	0.869	0.771	0.951	0.861
	XGBoost	0.921	0.888	0.801	0.958	0.880
4s	SVC	0.941	0.913	0.834	0.972	0.903
	RForest	0.943	0.918	0.865	0.966	0.915
	KNN	0.935	0.906	0.828	0.967	0.897
	XGBoost	0.939	0.913	0.852	0.965	0.909
5s	SVC	0.937	0.908	0.827	0.969	0.898
	RForest	0.9471	0.9233	0.870	0.969	0.9197
	KNN	0.933	0.901	0.815	0.967	0.891
	XGBoost	0.939	0.911	0.847	0.966	0.906
6s	SVC	0.939	0.910	0.828	0.971	0.899
	RForest	0.944	0.920	0.8708	0.966	0.918
	KNN	0.938	0.907	0.809	0.9752	0.892
	XGBoost	0.945	0.921	0.866	0.968	0.917

TABLE II: Per-subject performance of global vs. fine-tuned models. Overall difference: $\Delta\text{F1}_{\text{mean}} = 0.029$ (IC 95% -0.049–0.107), $t = 0.762$, $p = 0.453$; Wilcoxon $W = 132.5$, $p = 0.617$.

Subject	Global F1	Fine-tuned F1	ΔF1	Train Time (s)
1	0.471	0.442	−0.029	3.730
2	0.708	0.447	−0.261	3.680
3	1.000	1.000	0.000	4.420
4	0.464	0.482	0.018	3.700
5	0.482	0.464	−0.018	3.490
6	0.498	1.000	0.502	4.000
7	0.496	1.000	0.504	4.050
8	0.744	0.693	−0.051	3.620
9	0.608	0.592	−0.016	3.490
11	0.849	0.759	−0.090	4.230
14	0.510	0.482	−0.027	3.680
15	0.483	1.000	0.517	3.510
16	0.637	0.576	−0.060	3.500
18	0.496	0.497	0.001	4.400
19	0.538	0.473	−0.065	3.460
20	0.497	0.498	0.001	3.360
21	0.489	0.485	−0.004	4.360
23	0.561	0.482	−0.080	4.690
24	0.535	0.443	−0.092	3.580
27	0.791	0.776	−0.015	3.510
29	0.445	0.506	0.061	4.220
30	0.717	0.759	0.042	4.080
31	0.744	0.498	−0.246	3.330
32	1.000	1.000	0.000	3.460
33	0.508	0.614	0.106	4.340
34	0.671	0.726	0.055	3.840

REFERENCES

[1] Theodoros Bikias et al., "Deepfog: An imu-based detection of freezing of gait episodes in parkinson's disease using deep learning and edge computing," *Frontiers in Robotics and AI*, vol. 8, 2021.

[2] Luis Sigcha et al., "A transformer-based deep learning approach for detecting freezing of gait in parkinson's disease using a single waist-mounted sensor," *Sensors*, vol. 24, no. 7, pp. 1895, 2024.

[3] Yu-Chuan Lin et al., "Freezing of gait detection based on deep learning using edge ai," *arXiv preprint arXiv:2212.00729*, 2022.

[4] Shyam Barua et al., "Lift-pd: Learning interpretable features with self-supervision for freezing of gait detection," *arXiv preprint arXiv:2410.20715*, 2024.

[5] S. T. Moore, H. G. MacDougall, and W. G. Ondo, "A wearable system for detecting freezing of gait in parkinson's disease,"

Movement Disorders, vol. 23, no. 6, pp. 836–839, 2008.

[6] Martina Mancini, Arash Salarian, Patricia Carlson-Kuhta, Cristina Zampieri, Laurie King, Lorenzo Chiari, and Fay B Horak, "Mobility lab to assess balance and gait with synchronized body-worn sensors," *Journal of bioengineering biomedical science*, vol. Suppl 1, pp. 007, 2011.

[7] C. Ribeiro de Souza, R. Miao, I. Ávila de Oliveira, A. C. De Lima-Pardini, D. F. De Campos, C. Silva-Batista, et al., "A public dataset of video, acceleration, and angular velocity in individuals with parkinson's disease during the turning-in-place task," 2022, Dataset.

[8] C. Ribeiro De Souza, E. C. Figueiredo, G. Almeida, J. Oliveira, L. Brito, F. Maia, C. S. Oliveira, A. Santos, S. M. C. Silva, A. Batista, 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 Aging Neuroscience*, vol. 16, pp. 832463, 2022.

[9] M. Gilat, "How to annotate freezing of gait from video: a standardized method using open-source software," *Journal of Parkinson's Disease*, vol. 9, pp. 821–824, 2019.

[10] Silas Mazilu, Markus Hardegger, Gerhard Tröster, Elad Gazit, and Jeffrey M. Hausdorff, "Online detection of freezing of gait with smartphones and machine learning techniques," in *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*, 2012, pp. 123–130.

[11] Luis Sigcha, Nuria Costa, Iván Pavón, Ricardo Costa, Pedro Arezes, and Jorge López, "Deep learning approaches for detecting freezing of gait in parkinson's disease patients through on-body acceleration sensors," *Sensors*, vol. 20, no. 7, pp. 1895, 2020.

[12] Yifan Guo, Jian Yang, Yifan Liu, Xian Chen, and Guang-Zhong Yang, "Detection and assessment of parkinson's disease based on gait analysis: A survey," *Frontiers in Aging Neuroscience*, vol. 14, pp. 916971, 2022.

[13] M. Bächlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Tröster, "Wearable assistant for parkinson's disease patients with the freezing of gait symptom," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 436–446, 2010.

[14] Mieke Pardoel, Gurleen Shalin, Julie Nantel, and Edward D. Lemaire, "Insights into parkinson's disease-related freezing of gait detection and prediction: A comprehensive review," *Sensors*, vol. 24, no. 12, pp. 3959, 2024.