OPEN ACCESS

University of BRISTOL

Chan, L. L. Y., Yang, S., Aswani, M., Henderson, E. J., Lord, S. R., & Brodie, M. A. (2024). Development, Validation, and Limits of Freezing of Gait Detection Using a Single Waist-Worn Device. *IEEE Transactions on Biomedical Engineering*, 1-8. Advance online publication. https://doi.org/10.1109/TBME.2024.3407059

Peer reviewed version

License (if available):
CC BY

Link to published version (if available):
10.1109/TBME.2024.3407059

Link to publication record in Explore Bristol Research
PDF-document

This is the accepted author manuscript (AAM) of the article which has been made Open Access under the University of Bristol's Scholarly Works Policy. The final published version (Version of Record) can be found on the publisher's website. The copyright of any third-party content, such as images, remains with the copyright holder.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

Development, Validation, and Limits of Freezing of Gait Detection Using a Single Waist-Worn Device

Lloyd L.Y. Chan, *Member, IEEE*, Song Yang, *Member, IEEE*, Mira Aswani, *Member, IEEE*, Lauren Kark, *Member, IEEE*, Emily Henderson, Stephen R Lord, and Matthew A. Brodie*, *Member, IEEE*

Abstract*— *Objective: **Freezing of Gait (FOG) often described as the sensation of "the feet being glued to the ground" is prevalent in people with Parkinson's disease (PD) and severely disturbs mobility. In addition to tracking disease progression, precise detection of the exact boundaries for each FOG episode may enable new technologies capable of "breaking" FOG in real time. This study investigates the limits of sensitivity and performance for automatic device-based FOG detection.**

Methods: **Eight machine-learning classifiers (including Neural Networks, Ensemble & Support Vector Machine) were developed using (i) accelerometer and (ii) accelerometer and gyroscope data from a waist-worn device. While wearing the device, 107 people with PD completed a walking and mobility task designed to elicit FOG. Two clinicians independently annotated the precise FOG episodes using synchronized video according to international guidelines, which were incorporated into a flowchart algorithm developed for this study. Device-detected FOG episodes were compared to the annotated FOG episodes using 10-fold cross-validation to determine accuracy and with Interclass Correlation Coefficients (ICC) to assess level of agreement.**

Results: **Development used 50,962 windows of data representing over 10 hours of data and annotated activities. Very strong agreement between clinicians for precise FOG episodes was observed (90% sensitivity, 92% specificity and $ICC_{1,1}$ = 0.97 for total FOG duration). Device-based performance varied by method, complexity and cost matrix. The Neural Network that used only 67 accelerometer features provided a good balance between high sensitivity to FOG (89% sensitivity, 81% specificity and $ICC_{1,1}$ = 0.83) and solution stability (validation loss ≤ 5%).**

Conclusion: **The waist-worn device consistently reported accurate detection of precise FOG episodes and compared well to more complex systems. The superior agreement between clinicians indicates there is room to improve future device-based FOG detection by using larger and more varied data sets.**

Significance: **This study has clinical implications with regard to improving PD care by reducing reliance on clinical FOG assessments and time-consuming visual inspection. It shows high sensitivity to automatically detect FOG is possible.**

***Index Terms*— Freezing, gait, Parkinson's, device, detection.**

I. Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disease, characterized by neuronal loss in the substantia nigra, intracellular deposition of α-synuclein and subsequent dopamine deficiency [1]. Its cardinal presentations include tremors, rigidity, bradykinesia, and postural instability [1]. Freezing of gait (FOG) is a common and severe gait disturbance [2] that presents as an episodic inability to initiate and continue walking [3]. People with PD often describe FOG as the sensation of having "their feet being glued to the ground", and it can be characterized into three forms: moving with very small steps; trembling legs; and total akinesia [4]. By impairing mobility [5] and increasing fall risk [6], FOG significantly reduces quality of life.

Accurate gait disturbance assessments are important for disease management. While dopaminergic medication generally decreases the frequency of FOG, prolonged treatment may increase FOG [7]. Such divergent effects indicate accurate FOG assessments are required to improve the titration of levodopa dosage for optimal control of motor symptoms. Furthermore, accurate FOG assessments might also be used to improve the precision of PD staging, which could help target device-aided therapies such as deep brain stimulation to be delivered at the most appropriate juncture during the clinical course of the condition [8].

Presently, the assessment of FOG is primarily based on clinical assessment [9] or self-report [10]. Annotation of video recordings is considered the gold standard [11, 12], but variations in assessments of FOG, time-consuming annotation and the need for skilled assessors limit its widespread use. Furthermore, FOG occurs most frequently during daily activities making it difficult to quantify the frequency and severity during assessment in an ambulatory care clinic [13, 14]. Self-report questionnaires provide estimates of FOG during daily activities but are subject to recall bias and have low resolution [15].

Recently, wearable devices have been used to detect FOG [16] with moderate-to-good agreement achieved between visual inspection and sensor-based methods in quantifying the total number of FOG episodes [17] with the aim of providing continuous low-cost FOG assessments. For example, a sensitivity of 74.7% and a specificity of 79.0% for detecting FOG has been reported using a Support Vector Machine (SVM) embedded into a device worn on the side of the waist in the home environment.[18]. These findings support the potential of continuous FOG monitoring as deployed in a currently available commercial device (Holter, SENSE4CARE) [19].

However, sample sizes utilized in most previous investigations, have been small (ranging in size from one to forty participants), which may limit the drawing of representative inferences [16-18]. Paired with machine learning (ML) approaches, a small sample size may lead to overfitting and difficulties accounting for individual variability, which

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

may limit accuracy when used in other cohorts.

In the one larger study with 71 participants, a sensitivity of 80.0% ± 19.2% and specificity of 82.5% ± 11.2% was reported for longer (>1s) FOG episodes using an SVM classifier [20] and independent validation set. The study used three devices placed above both ankles and on the lower back to improve performance, which demonstrates the advantage of combining data from multiple devices to detect FOG. However, methodological variances in general, make direct comparisons between study outcomes ambiguous. For example, in the above study, sensitivity was determined for each individual separately. For participants with no annotated FOG episodes, sensitivity was then considered to be 100% before the mean and standard deviation for sensitivity were calculated.

The use of multiple sensors or multiple devices [17, 20, 23-24] also has a disadvantage in that they may increase power consumption and the inconvenience of daily charging if used in a home environment. In comparison, a small user-friendly single device with fewer sensors could improve compliance by improving comfort, convenience and battery life for long-term remote applications. Regarding single device location, waist-worn sensors are acceptable to people with PD (Dynaport, McRoberts) [26], and have demonstrated lower false detection rates than wrist-worn sensors, potentially due to their more stable orientations [25].

Sensor-based FOG detection could assist in improving the usability of many emerging technologies, including laser-generated visual-cueing systems [21] and rhythmic auditory or sensory stimulation systems [21, 22]. However, several hurdles to seamless implementation remain. Currently, the practical upper limits of device-based FOG detection remain unknown and methodological designs have often focused on general remote monitoring applications rather than the precise detection of the exact boundaries of each FOG episode. For example, some studies have merged multiple data windows into single episodes of FOG or non-FOG [16-18] or excluded shorter FOG episodes (<1s) [20] or non-FOG episodes (<5s) [18].

Consequently, this may limit the efficacy of current FOG detection classifiers for use in new assistive devices [21-22] capable of "breaking" shorter or intermitted FOG episodes in real time regardless of FOG duration. If detection of the exact boundaries of each FOG episode were possible, then stimulation could be delivered more precisely at the onset of FOG. This would avoid the burden of people with PD needing to manually activate a device while experiencing FOG.

Considering the above complex issues, we first developed a flowchart algorithm based on consensus guidelines [2, 3, 27] to minimise the variances between the annotation of FOG episode onsets and endings [Fig 1B]. This gold standard was considered to provide an upper limit for subsequent device-based FOG detection. We then developed eight different FOG detection classifiers using (a) the accelerometer plus gyroscope data and (b) accelerometer data only from a single waist-worn device in 107 people with PD. To increase sensitivity to FOG we further investigated increased costs for missed FOG data windows. Algorithm performance was assessed regarding its ability to (1) identify precise FOG episodes (defined as the exact start and stop times of each FOG episode), and (2) quantify total FOG duration while participants completed a complex walking and mobility task designed to induce FOG.

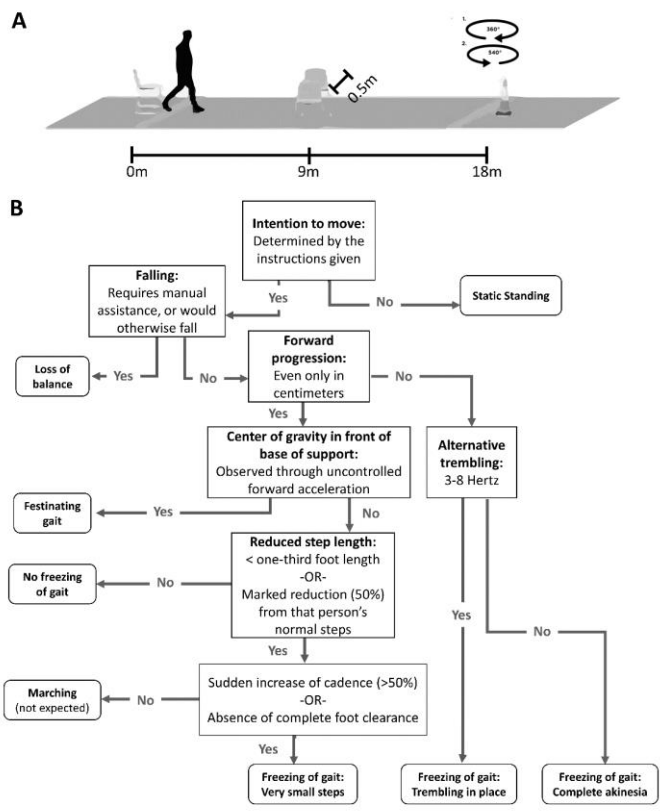

Fig. 1. [A] Walking task designed to elicit FOG. Participants stood up from a chair, walked through two chairs placed 50 centimeters apart, turned 360° to the right, 540° to the left, and then returned. [B] Flowchart developed for the video annotation and based on international consensus guidelines [2, 3, 27]. For this study we adopted and implemented the definition for FOG described as follows: "Brief, episodic absence or marked reduction of forward progression of the feet despite the intention to walk."

II. Methods

A. Participants

One hundred and thirty people with moderate Parkinson's disease (Hoehn and Yahr stage 2-3) were recruited from a single center in South-West England to participate in the ReSPonD trial [28]. Data from the 107 participants who completed a walking task designed to induce FOG [Fig 1A] without the use of walking aids were extracted. Participants were aged between 46 to 90 years and had a disease duration of 5 to 13 years. Detailed participant characteristics are presented in Table 1. Ethics approval was granted from the South West Research Ethics Committee, UK in September 2011 and the Medicines and Healthcare Regulatory Agency (MHRA) in June 2012. Informed consent was obtained from all participants before participation.

B. Data collection

Participants first completed a series of questionnaires, including the Montreal Cognitive Assessment (MoCA), Geriatric Depression Scale (GDS), Frontal Assessment Battery

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

TABLE I
PARTICIPANT CHARACTERISTICS (N=107)

	Non-freezers (N=10)	Freezers (N=97)	Total (N=107)	P-value
Female	5 (50.0%)	34 (35.1%)	39 (36.4%)	0.35[1]
Age(Years)	65.6 (5.88)	70.2 (8.32)	69.7 (8.21)	0.05[2]
Disease Duration (Years)	6.6 (5.63)	10.3 (6.64)	9.9 (6.62)	0.07[2]
Requires Walking Aid	5 (50.0%)	70 (72.2%)	75 (70.1%)	0.15[1]
MoCA Score	25.4 (3.95)	24.2 (3.50)	24.3 (3.54)	0.21[2]
GDS	3.8 (3.77)	4.4 (3.00)	4.3 (3.06)	0.25[2]
FAB	14.5 (2.72)	13.7 (2.56)	13.8 (2.57)	0.55[2]
CFQ	35.2 (11.70)	40.9 (14.77)	40.4 (14.56)	0.22[2]
Number of Falls in the past 12 months	2.5 (2.0, 4.0)	6.0 (2.0, 20.0)	6.0 (2.0, 20.0)	0.03[3]
UPDRS total	72.9 (24.03)	89.9 (24.73)	88.3 (25.05)	0.08[2]
UPDRS Part 3	31.6 (9.63)	41.5 (13.20)	40.6 (13.20)	0.04[2]
NFOGQ	12.0 (9.0, 20.5)	17.5 (13.0, 22.5)	17.0 (13.0, 22.5)	0.33[3]
BMI	32.3 (8.27)	27.2 (4.22)	27.7 (4.92)	0.03[2]

[1] Chi-Square p-value; [2] Independent t-test p-value; [3] Mann-Whitney U-test p-value; BMI, body mass index CFQ, cognitive failure questionnaire; FAB, frontal assessment battery; GDS, geriatric depression scale; MoCA, Montreal cognitive assessment; NFOGQ, the new freezing of gait questionnaire; UPDRS, unified Parkinson's disease rating scale. Freezers were defined as individuals who exhibited any episodes of FOG during the walking task. Conversely, non-freezers were participants who did not present any FOG while performing the same task.

(FAB), the Cognitive Failures Questionnaire (CFQ), the Unified Parkinson's Disease Scale (UPDRS), the New Freezing of Gait Questionnaire (NFOGQ) and questions regarding falls in the past twelve months. Gait assessments were subsequently conducted while participants were in the 'on' medication state. Wearing their usual shoes, participants were instructed to undertake a 36-meter complex walking and mobility task twice, designed to induce FOG [Fig 1A]. Recording started with participants sitting for 3 seconds. Participants then stood up from a chair, walked through two chairs placed 50 centimeters apart, turned 360° to the right, turned 540° to the left, and then walked back to the chair and sat down. The recording was stopped after three seconds of sitting. Participants were video recorded and wore a small waist-worn device (106.6 x 58 x 11.5mm, 55 grams) attached to a belt that recorded triaxial accelerations and gyroscopic data at 100Hz (Dynaport Hybrid, McRoberts, Netherlands).

Annotation of FOG as the gold standard was undertaken by a physiotherapist and an exercise physiologist, both with substantial experience in PD. Both clinicians watched video recordings of the participants' progress through the walking task and annotated the start and end of each FOG episode. To facilitate accurate and consistent annotation of FOG, a flowchart algorithm [Fig 1B] based on scientific literature and consensus outcomes from an international workshop [2, 3, 27] was developed. We considered that agreement between the two independent clinicians provided a practical upper limit for our subsequent sensor-based FOG detection.

Four annotated movement patterns, including (1) walking with very small steps, (2) trembling in place, (3) complete akinesia and (4) festinating gait were grouped into the single activity class of FOG. This ensured an adequate number of windows in each activity class and improved the clinical relevance of data by capturing a broad spectrum of FOG manifestations. Videos of 64 participants were assigned to each assessor, of which 20 were assigned to both assessors allowing examination of interrater reliability. Disagreements between assessors were resolved by a third reviewer.

C. Data processing

Linear acceleration and angular velocity (gyroscope) data from the waist-worn device were allocated into non-overlapping windows of adaptive length defined by heel strikes and impact events, which were identified by peaks in the vertical acceleration data with a method described previously [29]. Our adaptive window approach was developed to capture the characteristics of each step, attempted step or movement. The adaptive window duration was set to range between 0.3 (shortest anticipated festinating step) to 3 seconds (longest anticipated slow step) to ensure the relevance and consistency of the data within each window.

For each data window, a pool comprising 118 features was extracted using a combination of signal processing techniques from both the triaxial gyroscope and triaxial accelerometer. This included single-step features extracted from step times, autocorrelation coefficients, acceleration amplitudes, and wavelet coefficients, and 8-step features used to provide context in the form of the preceding movement patterns for the machine learning classifiers. MATLAB's feature ranking tool (The Math Works Inc., Version 2018b) was used to calculate and rank feature importance [Supplementary Table 1].

Data extraction was repeated using the first 67 features derived from the single triaxial accelerometer only to assess potential for low-power remote FOG detection with limited sensor data. Wavelet coefficients featured strongly in both feature sets, which included frequency information about linear accelerations (from the accelerometer) using the Daubechies 5 wavelet and angular velocities (from the gyroscope) using the Morlet wavelet. It was evident wavelet-transformed anteroposterior acceleration signals had increased power at lower scales (high frequencies) during FOG, increased power in mid scales during walking, and increased power at higher scales (lower frequencies) during turning and transitions [Fig 2A]. Discrete wavelet coefficients were calculated prior to data windowing, which meant all features could be extracted regardless of window duration.

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

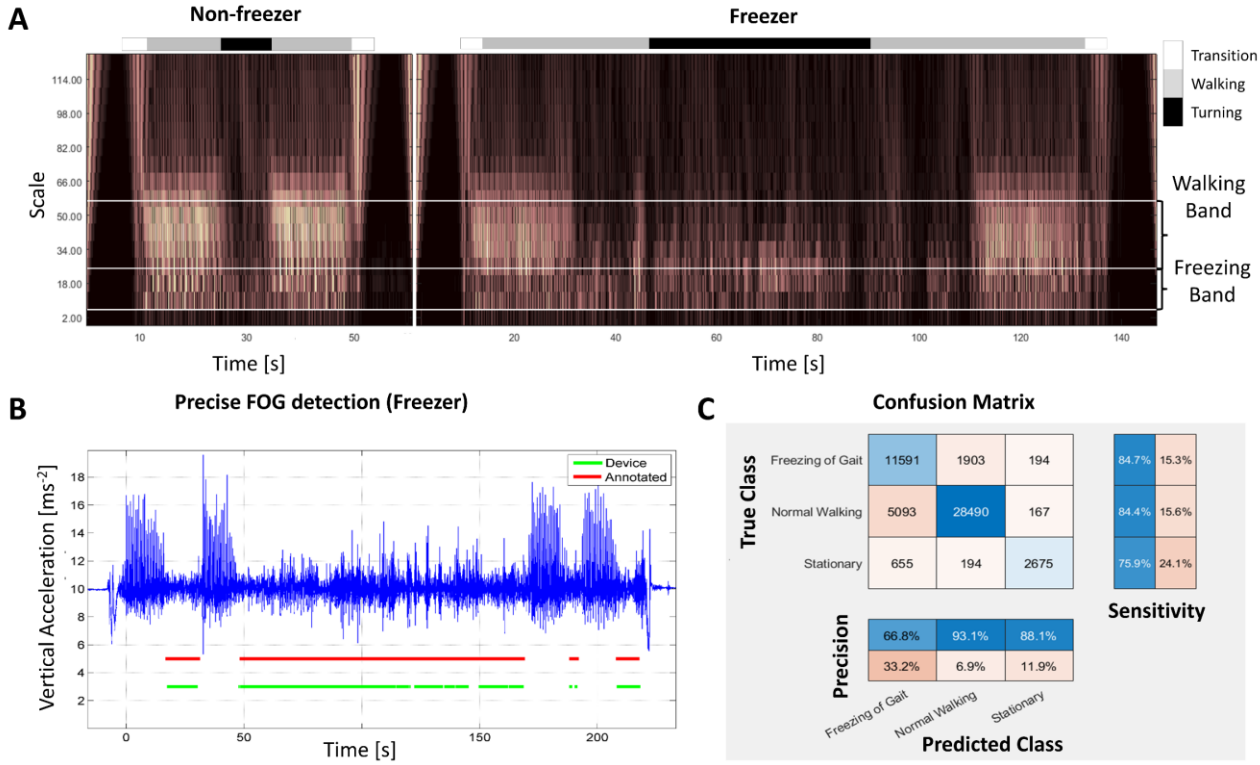

Figure 2. [A] Examples of continuous wavelet-transformed antero-posterior acceleration signals from a person with FOG (Freezer) and a Non-freezer. For illustrative purposes, the "Walking band" refers to mid frequency content (scales between 27 and 57) and the "Freezing band" refers to higher frequency content (scales between 4 and 22). [B] Example of the agreement between waist-worn device detection of FOG (green) and the annotated FOG reference (red) for the complex task. [C] Confusion matrix comparing true class (video annotation) vs predicted class (wearable device-based) assessment of FOG.

D. FOG Classification and Analysis

Data windows were classified into three activity classes: (1) Walking, (2) FOG (including festination), and (3) Transition or stationary for algorithm training using the referenced video annotations. Eight device-based FOG classification algorithms were then developed using an Ensemble of trees with adaptive boosting (Ensemble), Neural Networks (NNW) and Support Vector Machine second and third order (SVM 2 and SVM 3) to identify precise FOG episodes. The sensitivity to FOG episode onset was prioritized over specificity to non-FOG by increasing the cost of missed FOG windows [Table 2].

Because of the participant pool size, ten-fold cross-validation was selected as the most appropriate training and validation protocol to minimize overfitting and reduce risk of overestimating reported performance. For ten-fold validation, each participant's data was randomly assigned to one of ten folds. Each fold comprises data from ten or eleven participants. Stratification by number of data windows annotated as FOG was used to ensure each fold included similar signal of interest (FOG window) relative to background signal (non-FOG windows, defined by the 'Walking' and 'Transition or Stationary' activity classes).

Classifier training was repeated 10 times using 10 different 9-fold combinations of data followed by validation using the remaining unseen 1-fold for that training iteration. We combined the results from all data windows from the 10 unseen folds before calculating and reporting the outcomes.

Participant characteristics between those who demonstrated at least one FOG episode and those who did not were compared with chi-square tests, independent t-tests, and Mann-Whitney U tests for categorical, parametric continuous, and non-parametric continuous variables, respectively.

Performance of the waist-worn device in identifying the precise FOG episodes was assessed using **(1)** sensitivity (FOG windows correctly identified / total FOG windows), **(2)** specificity (non-FOG windows correctly identified/ total non-FOG windows), **(3)** accuracy (windows correctly identified / total windows), **(4)** precision (FOG windows correctly identified / total predicted FOG windows) and **(5)** the F1-score, a harmonic mean of precision and sensitivity. Non-FOG

TABLE II
MACHINE LEARNING PARAMETERS

Method	Parameters
Ensemble of Trees (both 67 and 118 features)	Method bag; Learners 30; Max Num Splits 50; Cost [0 2 1; 3 0 3; 1 2 0] (x3 missed FOG, x2 missed nonFOG)
Neural Network (118 features)	Layer Sizes [236 118] for 118 features; Lamba 0.001; Initial Step Size 0.01; Cost [0 2 1; 8 0 8; 1 2 0] (x8 missed FOG, x2 missed nonFOG)
Neural Network (67 features)	Layer Sizes [134 67] for 67 features; Lamba 0.001; Initial Step Size 0.01; Cost [0 2 1; 6 0 6; 1 2 0] (x6 missed FOG, x2 missed nonFOG)
Support Vector Machine 2 (both 67 and 118 features)	Polynomial Order 2; Box Constraint 1; Standardise true; Cost [0 1 1; 1 0 1; 1 1 0] (default cost, all classes equal)
Support Vector Machine 3 (both 67 and 118 features)	Polynomial Order 3; Box Constraint 1; Standardise true; Cost [0 2 1; 3 0 3; 1 2 0] (x3 missed FOG, x2 missed nonFOG)

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

TABLE III
WAIST-WORN DEVICE (USING ALL 118 FEATURES) PERFORMANCE AFTER 10-FOLD VALIDATION AND VIDEO ANNOTATION INTERRATER AGREEMENT. VALIDATION LOSS (LOSS) WAS CALCULATED BY SUBTRACTING THE REPORTED VALIDATED PERFORMANCE FROM THE TRAINED PERFORMANCE.

	Sensitivity (Loss)	Specificity (Loss)	Accuracy (Loss)	Precision (Loss)	F1 Score (Loss)	$ICC_{1,1}$ of Total FOG Duration
FOG Episodes Interrater Agreement Between Clinicians (referenced standard)						
Human	90%	92%	92%	84%	87%	0.97
Machine Learning FOG Detection (118 All Features)						
Ensemble	82% (5%)	86% (2%)	85% (3%)	65% (5%)	72% (5%)	0.84
NNW	75% (24%)	86% (10%)	83% (14%)	63% (27%)	68% (26%)	0.88
SVM 2	76% (11%)	91% (5%)	88% (6%)	74% (14%)	75% (13%)	0.85
SVM 3	76% (20%)	90% (8%)	86% (11%)	70% (22%)	73% (21%)	0.87
Machine Learning Low Power Potential (67 Accelerometer Features)						
Ensemble	80% (5%)	85% (2%)	84% (3%)	64% (5%)	71% (5%)	0.83
NNW	89% (5%)	81% (3%)	83% (3%)	59% (5%)	71% (5%)	0.83
SVM 2	74% (10%)	91% (4%)	87% (6%)	72% (12%)	73% (11%)	0.87
SVM 3	74% (19%)	89% (7%)	86% (10%)	72% (20%)	69% (20%)	0.86

included normal walking, postural transitions and stationary periods. Validation loss, as a measure of solution stability and potential generalizability to other cohorts, was calculated by subtracting the above validated performance (using outcome data from the unseen folds) from the trained performance (using outcome data from the training folds).

Performance for total FOG duration was assessed by combining data from the unseen folds and using the intra-class correlation coefficient, $ICC_{1,1}$ (one-way random effects, absolute agreement and single measurement). Interrater reliability between the two clinicians was assessed and considered the benchmark for the practical upper limit for device-based FOG detection. Algorithm development and statistical analyses were performed in MATLAB Statistics and Machine Learning Toolbox (Math Works Inc., Version 2020a).

III. RESULTS

Of the 107 included participants, 97 had at least one FOG episode while completing the walking task (91%). Compared to people without FOG, those with FOG had a lower BMI, more severe motor-related PD symptoms and reported more falls in the past twelve months. Participant demographics are summarized in Table 1. The mean total FOG duration (all episodes combined) for people who experienced FOG was 93 seconds with a standard deviation (SD) of 143 seconds.

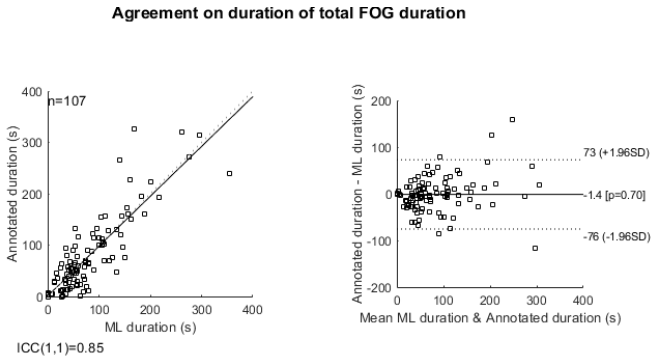

Figure 3. Agreement between annotated videos and "unseen" automatic wearable device-based assessment of total FOG duration using the Support Vector Machine 2 classifier with 118 features ($ICC_{1,1}$ = 0.85, $p \leq 0.05$).

A. FOG detection accuracy

The 10-fold cross-validation set for the device-based method development used 50,962 windows of data representing 37,809 seconds (10 hours, 30 minutes and 9 seconds) of activities. For each complex walking task, true activity classes (video annotation) and predicted activity classes (wearable device-based) were compared [Fig 2B] with overall performance calculated using a confusion matrix [Fig 2C]. The eight machine learning classifiers consistently reported good performances in distinguishing FOG windows from non-FOG windows [Table 3].

Agreements between device-based and video annotation methods were consistently strong across methods, i.e. $ICC_{1,1}$ 0.83 to 0.88 for total FOG duration [Fig 3]. Regarding our goal to prioritise sensitivity to detect FOG events, the relatively simple Neural Network with 67 accelerometer features provided a good balance between performance (89% sensitivity and 81% Specificity) and solution stability (validation loss $\leq$ 5%). Regarding overall performance, the second order SVM 2 with a standard cost matrix and 118 features reported the highest accuracy of 88%. Advantageously, limiting the Neural Network method to the 67 features derived from the accelerometer sensor resulted in only marginal performance changes while providing substantial reductions in validation loss (from loss $\leq$ 27% to loss $\leq$ 5%). Similarly, the second-order SVM 2 classifier performed better than the more complex third-order SVM 3 (from loss $\leq$ 22% to loss $\leq$ 14%) [Table 3].

The associated reduction in validation loss from using a less complicated classifier indicates increased solution stability and likely better generalisability when used in independent cohorts.

B. Interrater reliability

A total of 8,090 windows from the data of the 20 participants annotated by both assessors were included in the interrater reliability analysis. This revealed 92% accuracy in identifying precise FOG episodes (90% sensitivity and 92% specificity) and very strong agreement between assessors ($ICC_{1,1}$ = 0.97 for total FOG duration) [Fig 4].

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

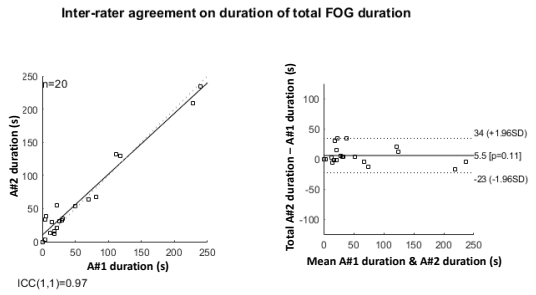

Figure 4. Very strong agreement between clinicians for total FOG duration ($ICC_{1,1}$ = 0.97, $p \leq 0.05$) was possible using the flowchart algorithm to guide video annotation. This provides a practical upper limit to compare the subsequent automatic device-based FOG detection performances against.

IV. Discussion

In this study involving 107 people with PD, we investigated the performance of eight different machine learning classifiers to detect precise FOG episodes. The single waist-worn device consistently provided strong agreement with the reference video annotation (Accuracies between 83% and 88% for FOG windows and $ICC_{1,1}$ between 0.83 and 0.88 for total FOG duration) [Table 3] regardless of classifier parameters. These findings suggest automated FOG detection has the potential to complement clinical FOG assessments and enable new assistive technologies to "break" FOG in real time.

Regarding our focus on improving sensitivity to FOG, the Neural Network with 67 features provided a good balance between performance and solution stability (loss ≤ 5%). The sensitivity of 89% was achieved by increasing the cost of missed FOG at the expense of a lower 81% specificity. For context, the 2nd order SVM 2 with default cost settings had lower (75%) sensitivity but higher (91%) specificity.

Our flowchart algorithm [Fig 2B] based on consensus guidelines [2, 3, 27] to guide precise annotation was successful. Very strong agreement between clinicians for FOG episodes was observed (90% sensitivity, 92% specificity and $ICC_{1,1}$ = 0.97 for total FOG duration) [Fig 4]. Compared to the performances of our eight device-based classifiers, we conclude that our skilled clinicians were better at identifying precise FOG episodes than our machine-learning classifiers.

Small differences were also observed in how the device-based methods identified FOG episodes [Fig 2B]. Where the clinical assessors typically annotated longer continuous FOG episodes [red line, Fig 2B], the device-based methods had the tendency to "tirelessly" consider each window of data according to the objectively trained criteria resulting in shorter FOG episodes [broken green line, Fig 2B].

Previous studies have shown good discrimination can be obtained between people with and without any freezing of gait using a validated questionnaire [10] or gait data derived from five inertial sensors attached to the feet, shins and lumbar region [17]. Identification of precise FOG episodes (defined as the exact boundaries of each FOG episode) may offer different opportunities for the clinical management of FOG.

Putting aside methodological variances across studies, our Neural Network classifier with a sensitivity of 89% and specificity of 81% compares favourably to previous studies. This includes a sensitivity of 75% and specificity of 76% for a waist-worn device [30], a sensitivity of 75% and specificity of 79% for a side waist-worn device [18], a sensitivity of 71%, a specificity of 82% for ankle-worn sensors [31] and a sensitivity of 80% and a specificity of 83% for a system comprising both waist-worn and ankle-worn sensors [20].

Automatic FOG detection could be used clinically in remote monitoring systems and incorporated into commercially available devices [19, 26]. The provision of remote FOG assessments may help inform and optimise intervention strategies including the timing and dose of medication. For this type of application, FOG frequency and time-of-day may be more important than the precise boundaries of each FOG episode and the cost matrix [Table 2] may be adjusted to achieve the optimum balance between the sensitivity and specificity of any classifier.

In contrast to most previous studies, we prioritized the precise detection of the exact boundaries of each FOG episode and adjusted the cost matrix to increase sensitivity. In this way, any technologies capable of "breaking" FOG episodes may be tuned to respond earlier before FOG becomes a fall. The related increase in false positive rates may be acceptable to people with PD provided the intervention technology is unobtrusive and can help stabilise gait before the onset of FOG.

Diverging from previous studies with fixed-length data windows, we effectively combined adaptive-length windows with wavelet analyses. While shorter windows have fewer data points to assess low-frequency information, longer windows may cover two or more confounding activity classes. Based on heel-strikes [29], we implemented our adaptive windows to capture the characteristics of each step, attempted step or movement. The adaptive window duration was set to range between 0.3 (shortest anticipated festinating step) to 3 seconds (longest anticipated slow step) to avoid covering both FOG and non-FOG activity classes in any one data window.

To balance the possibility that shorter data windows may provide insufficient information, we also used 8-step averages to provide preceding context. In our study, wavelet coefficients also ranked highly by importance [Supplementary Table 1], which agrees with previous studies that used frequencies attributed to trembling and walking to identify FOG [17, 32].

A previous study found FOG detection accuracy could be increased from 77% to 84% by personalization of a side-mounted waist-device [18]. Our trained accuracy results [Table 3, Accuracy + Loss = 86% to 97%] corroborate that a waist device provides enough information to develop accurate classifiers. Classifier performance, however, is only as good as the data used to train it. Our systematic approach to video annotation [Fig 1B] may have reduced one developmental barrier by providing consistent reference data. The next step may be to collaboratively combine data from multiple sources to develop future FOG classifiers with higher accuracy and lower validation losses. Another priority may be to improve the performance of the ubiquitous wrist device, which may help lower barriers to widespread use in people with PD [33].

The present study has several strengths. Firstly, data from a relatively large sample of 107 participants were analyzed and

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

an algorithmic approach to annotate FOG was developed [Fig 1B]. We compared eight different machine learning classifiers and after 10-fold cross-validation found the accuracy in detecting precise FOG episodes was relatively algorithm-agnostic. Secondly, as anticipated, FOG episodes were less common than other activities and therefore to improve sensitivity we increased the misclassification cost for FOG. Thirdly, the potential for low-power FOG detection was demonstrated by using the 67 features from the single triaxial accelerometer. A small user-friendly device with an accelerometer would require less frequent charging and could help to improve compliance through better comfort and convenience. Finally, the use of 67 accelerometer features only resulted in lower validation loss suggesting generalizability to other cohorts.

The current findings also need to be placed in the context of limitations. Firstly, our data were collected from a complex walking and mobility task designed to induce FOG, which may have reduced the variation of FOG recorded. Although it included sitting and transitions, it did not include all activities encountered in home-based environments. Secondly, despite our 10-fold validation, it is likely our method fits our sample better than others and will require further independent validation. Future work should aim to (i) collaboratively collect substantially more FOG data from more people with PD; (ii) establish a broad international consensus on the beginning and end of each annotated FOG episode; and (iii) include both FOG-inducing complex mobility tasks and daily activity FOG data for increased ecological validity.

V. Conclusion and Significance

Using adaptive data windows and wavelet analyses, we found that a waist-worn device using only accelerometer data provided a good balance between sensitivity to detect FOG and solution stability. The flowchart approach to FOG annotation also provided consistent reference data. Automatic FOG detection has clinical implications to improve PD care by reducing reliance on subjective in-person FOG reporting and time-consuming visual inspection. Remote objective FOG assessments could help inform and optimise intervention strategies. These findings may contribute to improved device-based FOG detection and future assistive technologies capable of “breaking” FOG episodes in real time.

VI. Acknowledgement

This study was supported by an Australian Government Research Training Program (RTP) Scholarship, Parkinson's UK, the British Geriatrics Society and North Bristol NHS Trust. Novartis supplied the rivastigmine and placebo medication free of charge but had no other input into the study or preparation of this manuscript.

Ethics approval for this study was granted by the South West Research Ethics Committee, UK in September 2011, and by the Medicines and Healthcare Regulatory Agency (MHRA) in June 2012. Informed consent was obtained from all participants prior to their participation in the study. We confirm that we have read the Journal’s position on issues involved in ethical publication and affirm that this work is consistent with those guidelines.

LLYC received support from the Australian Government Research Training Program (RTP) Scholarship. The CHIEF-PD trial was funded by Parkinson's UK for EH. EH also received funding from the British Geriatrics Society and North Bristol NHS Trust. Novartis supplied medication for the study.

VII. References

[1] W. Dauer and S. Przedborski, "Parkinson's disease: mechanisms and models," (in eng), *Neuron,* vol. 39, no. 6, pp. 889-909, Sep 11 2003, doi: 10.1016/s0896-6273(03)00568-3.

[2] S. Perez-Lloret *et al.*, "Prevalence, determinants, and effect on quality of life of freezing of gait in Parkinson disease," (in eng), *JAMA Neurol,* vol. 71, no. 7, pp. 884-90, Jul 1 2014, doi: 10.1001/jamaneurol.2014.753.

[3] J. G. Nutt, B. R. Bloem, N. Giladi, M. Hallett, F. B. Horak, and A. Nieuwboer, "Freezing of gait: moving forward on a mysterious clinical phenomenon," (in eng), *Lancet Neurol,* vol. 10, no. 8, pp. 734-44, Aug 2011, doi: 10.1016/s1474-4422(11)70143-0.

[4] N. Giladi and J. M. Hausdorff, "The role of mental function in the pathogenesis of freezing of gait in Parkinson's disease," (in eng), *J Neurol Sci,* vol. 248, no. 1-2, pp. 173-6, Oct 25 2006, doi: 10.1016/j.jns.2006.05.015.

[5] O. Moore, C. Peretz, and N. Giladi, "Freezing of gait affects quality of life of peoples with Parkinson's disease beyond its relationships with mobility and gait," (in eng), *Mov Disord,* vol. 22, no. 15, pp. 2192-5, Nov 15 2007, doi: 10.1002/mds.21659.

[6] B. R. Bloem, J. M. Hausdorff, J. E. Visser, and N. Giladi, "Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena," (in eng), *Mov Disord,* vol. 19, no. 8, pp. 871-84, Aug 2004, doi: 10.1002/mds.20115.

[7] P. J. Koehler, J. Nonnekes, and B. R. Bloem, "Freezing of gait before the introduction of levodopa," (in eng), *Lancet Neurol,* vol. 20, no. 2, p. 97, Feb 2021, doi: 10.1016/s1474-4422(19)30091-2.

[8] A. Stefani *et al.*, "Mechanisms of action underlying the efficacy of deep brain stimulation of the subthalamic nucleus in Parkinson's disease: central role of disease severity," (in eng), *Eur J Neurosci,* vol. 49, no. 6, pp. 805-816, Mar 2019, doi: 10.1111/ejn.14088.

[9] M. Mancini, B. R. Bloem, F. B. Horak, S. J. G. Lewis, A. Nieuwboer, and J. Nonnekes, "Clinical and methodological challenges for assessing freezing of gait: Future perspectives," (in eng), *Mov Disord,* vol. 34, no. 6, pp. 783-790, Jun 2019, doi: 10.1002/mds.27709.

[10] A. Nieuwboer *et al.*, "Reliability of the new freezing of gait questionnaire: agreement between patients with Parkinson's disease and their carers," (in eng), *Gait Posture,* vol. 30, no. 4, pp. 459-63, Nov 2009, doi: 10.1016/j.gaitpost.2009.07.108.

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3407059

[11] M. Gilat, "How to Annotate Freezing of Gait from Video: A Standardized Method Using Open-Source Software," (in eng), *J Parkinsons Dis*, vol. 9, no. 4, pp. 821-824, 2019, doi: 10.3233/jpd-191700.

[12] Y. Kondo *et al.*, "Measurement Accuracy of Freezing of Gait Scoring Based on Videos," (in eng), *Front Hum Neurosci*, vol. 16, p. 828355, 2022, doi: 10.3389/fnhum.2022.828355.

[13] N. Giladi *et al.*, "Validation of the freezing of gait questionnaire in patients with Parkinson's disease," (in eng), *Mov Disord*, vol. 24, no. 5, pp. 655-61, Apr 15 2009, doi: 10.1002/mds.21745.

[14] J. Nonnekes, A. H. Snijders, J. G. Nutt, G. Deuschl, N. Giladi, and B. R. Bloem, "Freezing of gait: a practical approach to management," (in eng), *Lancet Neurol*, vol. 14, no. 7, pp. 768-78, Jul 2015, doi: 10.1016/s1474-4422(15)00041-1.

[15] F. Hulzinga *et al.*, "The New Freezing of Gait Questionnaire: Unsuitable as an Outcome in Clinical Trials?," (in eng), *Mov Disord Clin Pract*, vol. 7, no. 2, pp. 199-205, Feb 2020, doi: 10.1002/mdc3.12893.

[16] S. Pardoel, J. Kofman, J. Nantel, and E. D. Lemaire, "Wearable-Sensor-based Detection and Prediction of Freezing of Gait in Parkinson's Disease: A Review," (in eng), *Sensors (Basel)*, vol. 19, no. 23, Nov 24 2019, doi: 10.3390/s19235141.

[17] M. Mancini *et al.*, "Measuring freezing of gait during daily-life: an open-source, wearable sensors approach," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, p. 1, 2021/01/04 2021, doi: 10.1186/s12984-020-00774-3.

[18] D. Rodríguez-Martín *et al.*, "Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer," (in eng), *PLoS One*, vol. 12, no. 2, p. e0171764, 2017, doi: 10.1371/journal.pone.0171764.

[19] Pérez-López, *et al.*, "Comparison of the Results of a Parkinson's Holter Monitor With Patient Diaries, in Real Conditions of Use: A Sub-analysis of the MoMoPa-EC Clinical Trial." Frontiers in Neurology, 2022. 13.

[20] T. Reches *et al.*, "Using Wearable Sensors and Machine Learning to Automatically Detect Freezing of Gait during a FOG-Provoking Test," *Sensors*, vol. 20, no. 16, p. 4474, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/16/4474.

[21] D. Sweeney, L. R. Quinlan, P. Browne, M. Richardson, P. Meskell, and G. ÓLaighin, "A Technological Review of Wearable Cueing Devices Addressing Freezing of Gait in Parkinson's Disease," (in eng), *Sensors (Basel)*, vol. 19, no. 6, Mar 13 2019, doi: 10.3390/s19061277.

[22] L. Rosenthal, D. Sweeney, A. L. Cunnington, L. R. Quinlan, and G. ÓLaighin, "Sensory Electrical Stimulation Cueing May Reduce Freezing of Gait Episodes in Parkinson's Disease," (in eng), *J Healthc Eng*, vol. 2018, p. 4684925, 2018, doi: 10.1155/2018/4684925.

[23] M. Mancini, K. Smulders, R. G. Cohen, F. B. Horak, N. Giladi, and J. G. Nutt, "The clinical significance of freezing while turning in Parkinson's disease," (in eng), *Neuroscience*, vol. 343, pp. 222-228, Feb 20 2017, doi: 10.1016/j.neuroscience.2016.11.045.

[24] S. T. Moore *et al.*, "Autonomous identification of freezing of gait in Parkinson's disease from lower-body segmental accelerometry," (in eng), *J Neuroeng Rehabil*, vol. 10, p. 19, Feb 13 2013, doi: 10.1186/1743-0003-10-19.

[25] S. Mazilu, U. Blanke, A. Calatroni, E. Gazit, J. M. Hausdorff, and G. Tröster, "The role of wrist-mounted inertial sensors in detecting gait freeze episodes in Parkinson's disease," *Pervasive and Mobile Computing*, vol. 33, pp. 1-16, 2016/12/01/ 2016, doi: https://doi.org/10.1016/j.pmcj.2015.12.007.

[26] A. Weiss, T. Herman, N. Giladi, and J. M. Hausdorff, "Objective assessment of fall risk in Parkinson's disease using a body-fixed sensor worn for 3 days," (in eng), *PLoS One*, vol. 9, no. 5, p. e96675, 2014, doi: 10.1371/journal.pone.0096675.

[27] E. Heremans, A. Nieuwboer, and S. Vercruysse, "Freezing of gait in Parkinson's disease: where are we now? " Curr Neurol Neurosci Rep 13, 350, doi:10.1007/s11910-013-0350-7 (2013).

[28] E. J. Henderson, S. R. Lord, J. C. Close, A. D. Lawrence, A. Whone, and Y. Ben-Shlomo, "The ReSPonD trial--rivastigmine to stabilise gait in Parkinson's disease a phase II, randomised, double blind, placebo controlled trial to evaluate the effect of rivastigmine on gait in patients with Parkinson's disease who have fallen," (in eng), *BMC Neurol*, vol. 13, p. 188, Dec 3 2013, doi: 10.1186/1471-2377-13-188.

[29] W. Zijlstra and A. L. Hof, "Assessment of spatio-temporal gait parameters from trunk accelerations during human walking," *Gait & Posture*, vol. 18, no. 2, pp. 1-10, 2003/10/01/ 2003, doi: https://doi.org/10.1016/S0966-6362(02)00190-X.

[30] H. Zach *et al.*, "Identifying freezing of gait in Parkinson's disease during freezing provoking tasks using waist-mounted accelerometry," (in eng), *Parkinsonism Relat Disord*, vol. 21, no. 11, pp. 1362-6, Nov 2015, doi: 10.1016/j.parkreldis.2015.09.051.

[31] M. Bächlin, M. Plotnik, D. Roggen, N. Giladi, J. M. Hausdorff, and G. Tröster, "A wearable system to assist walking of Parkinson s disease patients," (in eng), *Methods Inf Med*, vol. 49, no. 1, pp. 88-95, 2010, doi: 10.3414/me09-02-0003.

[32] S. T. Moore, H. G. MacDougall, and W. G. Ondo, "Ambulatory monitoring of freezing of gait in Parkinson's disease," (in eng), *J Neurosci Methods*, vol. 167, no. 2, pp. 340-8, Jan 30 2008, doi: 10.1016/j.jneumeth.2007.08.023.

[33] C. Virbel-Fleischman, Y. Rétory, S. Hardy, C. Huiban, J. C. Corvol, and D. Grabli, "Body-Worn Sensors for Parkinson's disease: A qualitative approach with patients and healthcare professionals," (in eng), *PLoS One*, vol. 17, no. 5, p. e0265438, 2022, doi: 10.1371/journal.pone.0265438.

Authorized licensed use limited to: UNIVERSITY OF BRISTOL. Downloaded on July 19,2024 at 09:18:19 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.