Fall Risk Prediction in Parkinson's Disease Using Real-World Inertial Sensor Gait Data

Martin Ullrich, *Student Member, IEEE*, Nils Roth, *Member, IEEE*, Arne Küderle, *Graduate Student Member, IEEE*, Robert Richer, *Student Member, IEEE*, Till Gladow, Heiko Gaßner, Franz Marxreiter, Jochen Klucken, Bjoern M. Eskofier, *Senior Member, IEEE*, and Felix Kluge

***Abstract*—Falls are an eminent risk for older adults and especially patients with neurodegenerative disorders, such as Parkinson's disease (PD). Recent advancements in wearable sensor technology and machine learning may provide a possibility for an individualized prediction of fall risk based on gait recordings from standardized gait tests or from unconstrained real-world scenarios. However, the most effective aggregation of continuous real-world data as well as the potential of unsupervised gait tests recorded over multiple days for fall risk prediction still need to be investigated. Therefore, we present a data set containing real-world gait and unsupervised 4x10-Meter-Walking-Tests of 40 PD patients, continuously recorded with foot-worn inertial sensors over a period of two weeks. In this prospective study, falls were self-reported during a three-month follow-up phase, serving as ground truth for fall risk prediction. The purpose of this study was to compare different data aggregation approaches and machine learning models for the prospective prediction of fall risk using gait parameters derived either from continuous real-world recordings or from unsupervised gait tests. The highest balanced accuracy of 74.0% (sensitivity: 60.0%, specificity: 88.0%) was achieved with a Random Forest Classifier applied to the real-world gait data when aggregating all walking bouts and days of each participant. Our findings suggest that fall risk can be predicted best by merging the entire two-week real-world gait data of a patient, outperforming the prediction using unsupervised gait tests (68.0% balanced accuracy) and contribute to an improved understanding of fall risk prediction.**

***Index Terms*—Classification, gait analysis, home monitoring, inertial measurement unit.**

I. Introduction

FALLING is a life-threatening risk for older adults in general and especially for Parkinson's disease (PD) patients. Different studies revealed that more than sixty percent of PD patients experience one or multiple falls per year [1], [2]. The consequences of falls include a high likelihood of hospitalization [3], for example because of fractures of the lower extremities [4], but also an increased fear of falling which prevents activities of daily life and thus reduces quality of life [5].

Besides general disease severity and fall history, many studies report particularly mobility aspects, like level of physical activity, postural instability, freezing of gait, shuffling gait, or impaired stair ambulation as major risk factors for falls [2], [6], [7], [8], [9], [10]. Until a decade ago, fall risk has mainly been studied using clinical tests, data of supervised standardized gait tests acquired by stationary motion tracking systems, and questionnaires that were analyzed with traditional statistical methods. In general, retrospective and prospective trials have been published, where in the former the ground truth for fall risk was determined based on the participants' history of falls

Manuscript received 1 March 2022; revised 3 August 2022 and 25 September 2022; accepted 13 October 2022. Date of publication 19 October 2022; date of current version 5 January 2023. The work of Heiko Gaßner was supported by the Fraunhofer Internal Programs under Grant Attract 044-602140 and Grant 044-602150. The work of Jochen Klucken was supported by dHealthPD (FNR, Luxembourg). The work of Bjoern M. Eskofier was supported by DFG within the framework of the Heisenberg Professorship Programme under Grant ES 434/8-1. This work was supported in part by the Bavarian Ministry for Economy, Regional Development and Energy through the Medical Valley Award 2017 (FallRiskPD Project) and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grants SFB 1483–Project-ID 442419336 (EmpkinS) and Project-ID 438496663 (Mobility_APP). This work was supported by the Mobilise-D project that has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under Grant Agreement 820820. This JU received support from the European Union's Horizon 2020 Research and Innovation Program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). *(Corresponding author: Martin Ullrich.)*

Martin Ullrich, Nils Roth, Arne Küderle, Robert Richer, Bjoern M. Eskofier, and Felix Kluge are with the Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany (e-mail: martin.ullrich@fau.de; nils.roth@fau.de; arne.kuederle@fau.de; robert.richer@fau.de; bjoern.eskofier@fau.de; felix.kluge@fau.de).

Till Gladow was with the Department of Molecular-Neurology, University Hospital Erlangen, 91054 Erlangen, Germany. He is now with the Medical Valley Digital Health Application Center, 96050 Bamberg, Germany (e-mail: till.gladow@mv-dmac.de).

Heiko Gaßner is with the Department of Molecular-Neurology, University Hospital Erlangen, 91054 Erlangen, Germany, and also with the Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany (e-mail: heiko.gassner@uk-erlangen.de).

Franz Marxreiter is with the Department of Molecular-Neurology, University Hospital Erlangen, 91054 Erlangen, Germany (e-mail: franz.marxreiter@uk-erlangen.de).

Jochen Klucken was with the Medical Valley Digital Health Application Center, 96050 Bamberg, Germany, and also with the Department of Molecular-Neurology, University Hospital Erlangen, 91054 Erlangen, Germany. He is now with the Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg, also with the Luxembourg Institute of Health, 1445 Strassen, Luxembourg, and also with the Centre Hospitalier de Luxembourg, 1210 Luxembourg, Luxembourg (e-mail: jochen.klucken@uni.lu).

Digital Object Identifier 10.1109/JBHI.2022.3215921

and in the latter based on falls recorded after the data collection. Kerr et al. [11] performed numerous neurological and functional tests with 101 PD patients and used a multivariate statistical model to predict falls recorded in a six-months follow-up period with a sensitivity of 78% and specificity of 84%. A combination of three predictors, namely history of falls in the previous year, freezing of gait in the past month and self-selected gait speed, was proposed by Paul et al. [12] to discriminate future fallers and non-fallers among 205 PD patients. Camera-based motion capture studies allowed to get further kinematic evidence for the relation of motion and fall risk. In a study by Cole et al. [13] with PD patients and healthy controls, the future fallers of the PD cohort had an increased stride time variability, walked slower and with shorter strides, and showed a poorer gait stability compared to the fallers from the control cohort. However, all the previously mentioned studies have the drawback that they required clinical expert ratings and partially expensive stationary laboratory equipment.

The technological advancements of inertial measurement units (IMUs), including accelerometers and gyroscopes, over the recent years enabled flexible studies with standardized gait tests in the laboratory. Greene et al. [14] and Silva et al. [15] recorded signals of body-worn IMUs while study participants performed timed up and go (TUG) tests. Features derived from the raw signals and in [14] also spatio-temporal parameters were used to perform a fall risk assessment based on the history of falls of the included older adults. Similar analysis approaches were applied on sensor measurements of straight walking along a path of ten to twenty meters by Latt et al. [16] for retrospective and by Doi et al. [17] for prospective fall risk estimation.

Standardized gait tests in the laboratory have certain limitations, including the lack of longitudinal data from the daily life of the patients who are often experiencing fluctuating symptom severity which cannot be captured during short and seldom clinical visits [18]. Therefore, researchers are currently transferring IMU-based gait analysis into the real world [19] and use the derived outcomes for fall risk estimation. Weiss et al. [20] recorded accelerometer data from 107 PD patients over three days with a sensor worn at the lower back. Raw signal frequency analysis revealed a wider band of the dominant frequency for patients with a fall history. The width of the dominant frequency band was also related to the time for the first fall in the year after the recordings based on Cox regression analysis. Furthermore, Din et al. [21] defined macro (e.g., number of walking bouts per day) and micro (e.g., step length) gait parameters and investigated their association to falls in older adults and PD patients using free-living sensor recordings and statistical significance tests [22]. They concluded that the information added by real-world monitoring enhance group differences between people with and without fall history as they capture environmental challenges in the daily life. However, an individual fall risk was neither determined in the study by Weiss et al. nor in the one by Din et al.

With respect to the large amount of gait data collected in sensor-based studies in laboratories and the real world, the use of machine learning methods has been shown to be helpful for an accurate prediction of fall risk in older adult cohorts incorporating multivariate data. One of the first studies applying machine learning was presented by Howcroft et al. [23] who investigated prospective fall risk predictions by neural networks, Naive Bayes classifiers, and Support Vector Machines. Their data set contained gait data from straight walking performed by older adults in a laboratory, recorded with pressure insoles at both feet and multiple IMUs, from which they derived several time series features and gait parameters. The best performing model was a neural network with an accuracy of 57%. Meyer et al. [24] trained neural networks with raw accelerometer signals of single strides recorded with Multiple Sclerosis patients in a laboratory hallway during $1\,\mathrm{min}$ of straight walking and reported an accuracy of up to $86\,\%$. Both publications [23] and [24] were limited to gait recorded in the lab and did not focus on a cohort of PD patients.

Three deep learning approaches were compared by Aicha et al. [25] for a prediction of future falls of older adults during a six-month follow-up period based on one-week accelerometer data. The models were trained on the raw signals of ten-second windows and achieved an area under the receiver operating characteristic curve (AUC) of 0.75. However, using a fixed window length of ten seconds does not use the potential of macro gait parameters or of aggregated information from single walking bouts or entire days as introduced by Din et al. [22]. Tunca et al. [26] followed up on the work presented in [25] by applying deep learning not only on the raw signals but also on sequences of spatio-temporal gait parameters. However, their data set contained only recordings of straight walking in the laboratory and their ground truth was the history of falls of the 76 older adults in their study cohort. The authors of [26] explained their increased accuracy of 92.1% in comparison to the results presented in Aicha et al. [25] by the advantageous domain knowledge contained in the gait parameters.

The literature review underlines that a differentiation of PD fallers and non-fallers is possible on group level by means of traditional statistical analysis of sensor-based data, and that there is potential for an individualized prediction of fall risk using machine learning. However, there are still several research gaps in the current state of the art as highlighted in the previous overview. To the best of our knowledge, there has been no study that attempted to predict fall risk with a pure PD cohort in a prospective manner using machine learning algorithms with spatio-temporal gait parameters obtained in the real world. Furthermore, the benefits of using macro and micro parameters in real-world studies, as well as the dependency of their temporal aggregation on the accuracy of fall risk prediction remain unanswered questions which should be investigated.

Another open topic is the comparison between gait parameters from standardized controlled settings and unconstrained real-world observations. For stroke patients, Punt et al. [27] reported that supervised camera-based measurements of treadmill walking and unsupervised real-world gait analysis with wearable sensors over seven days have similar capabilities for predicting fall risk. However, in their study, the real-world gait data was also only analyzed with a fixed window length of eight seconds. Recently, Gaßner et al. [28] presented a pilot study showing a high correlation of gait parameters recorded during repeated

4x10-Meter-Walking-Test (*4x10MWT*) in the hospital and in the home of PD patients under supervision of a clinical assessor. These tests performed at home might be a valuable link to bridge the gap between recordings in a standardized and in a real-world context. As *4x10MWT*s can be performed by patients without supervision and automatically detected from continuous real-world recordings [29], the potential of fall risk prediction using daily unsupervised gait tests should be investigated and compared to real-world gait.

Hence, the goal of this paper is to fill parts of the previously mentioned gaps in the literature, where our specific contributions are as follows:

1) We present a data set of 40 PD patients who were equipped with foot-worn IMUs for two weeks and recorded unconstrained real-world gait and multiple unsupervised *4x10MWT*s per day. Falls were recorded prospectively in a diary over three months after the sensor recordings.
2) We predict fall risk using different machine learning algorithms and compare the results of real-world and *4x10MWT* gait data.
3) We explore different approaches for the aggregation of sensor-derived parameters from single walking bouts or *4x10MWT*s over daily averages to participant-wise aggregates.

To the best of our knowledge, there is no other study in the literature comparing data from unsupervised standardized gait tests and unconstrained real-world recordings over multiple consecutive days for prospective fall risk estimation. Therefore, this study is closing a gap on the way of bringing sensor-based gait analysis from the laboratory into the real world. Identifying appropriate methods and data aggregation levels for using gait outcomes to predict fall risk has huge potential for revolutionizing healthcare and improving quality of life of patients with mobility impairments.

II. Methods

A. Data Acquisition

The data set used for this work consisted of 40 patients with idiopathic PD who participated in the *FallRiskPD* study (DRKS-ID: DRKS00015085) between March 2019 and June 2021. The University Hospital Erlangen, the Hospital Rummelsberg, and the Ernst von Bergmann Hospital Potsdam recruited the participants in the multicentric study and the data set is publicly available [30]. Participants had to fulfill the following inclusion criteria: Diagnosis of Parkinson's disease according to the guideline of the German Society for Neurology with a Hoehn and Yahr stage of I-III [31], being able to walk 4x10 meters without support, predominantly walking without walking aids in their daily life, and being able to read and understand the instructions for operating the sensor-based gait analysis system on their own during the home-monitoring phase. The following exclusion criteria were applied: A maximum walking distance of less than 100 m, decompensated cardiopulmonary limitations, and other pronounced musculoskeletal disorders that severely limit the ability to walk. The study was approved by the local ethics committee Re-No. 165_18B (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany) and the Landesärtzekammer Brandenburg (AS 25(bB)/2019). All participants gave written, informed consent, prior to the data collection.

Gait recordings were acquired with the *Mobile GaitLab* (Portabiles HealthCare Technologies GmbH, Erlangen, Germany). This real-world gait analysis system included two lightweight foot-worn IMUs, attached to the instep of the left and the right shoe, to avoid influencing the natural walking pattern of the participants. The IMUs recorded 3-d acceleration (range ± 16 g) and 3-d angular rate (range $\pm 2000^\circ/\text{s}$) with a sampling rate of 102.4 Hz. Data of the left and right sensor were synchronized and thus processed on a common time axis [32].

In a pre-visit at the hospital, the Unified Parkinson's Disease Rating Scale (UPDRS)-III score [33] and the Hoehn and Yahr stage [31] were determined by a clinician, while patients were in ON state, following their regular medication plan. Furthermore, the recording system and the study protocol were thoroughly explained to the participants including a detailed interactive demonstration and a written manual. Afterwards, the participants wore the system over two weeks during their wake time indoors and outdoors, following their activities of daily living to record real-world gait. No restrictions were made regarding any activities the participants could perform. In order to clarify and sort out any technical difficulties that could occur when the participants operated the recording system at home, regular phone calls during the first few days of the monitoring period were performed by a study nurse. Additionally, the protocol included unsupervised *4x10MWT*s at preferred walking speed in the morning after the start of the recording, around noon, and in the evening before stopping the recording, as described in our previous work [29]. The participants used an adapted version of the *PatientConcept* (NeuroSys GmbH, Ulm, Germany) smartphone application for annotating the *4x10MWT* start and end times.

After the two-week recording period the participants filled a paper-based daily fall diary during a follow-up phase of three months without sensor recordings. A fall was defined as an event where the patient landed unintentionally on the floor, ground, or on a lower level as suggested by Lamb et al. [34]. Similar as in previous work [23], [25], participants who were experiencing any falls during the follow-up phase were assigned to the *high* fall risk group and those who did not fall to the *low* fall risk group for model training.

For the analysis, four participants were excluded as they did not fulfill the criteria of at least 5 hours wear time on at least 6 days during the recording period. One more participant was excluded as there were no *4x10MWT*s performed in any of the gait recordings. The demographic information for the remaining 35 participants are presented in Table I.

B. Gait Parameters

A general overview of the algorithmic approach of this work is given in Fig. 1. Sensor recordings were processed with the goal to derive spatio-temporal gait parameters from real-world gait and

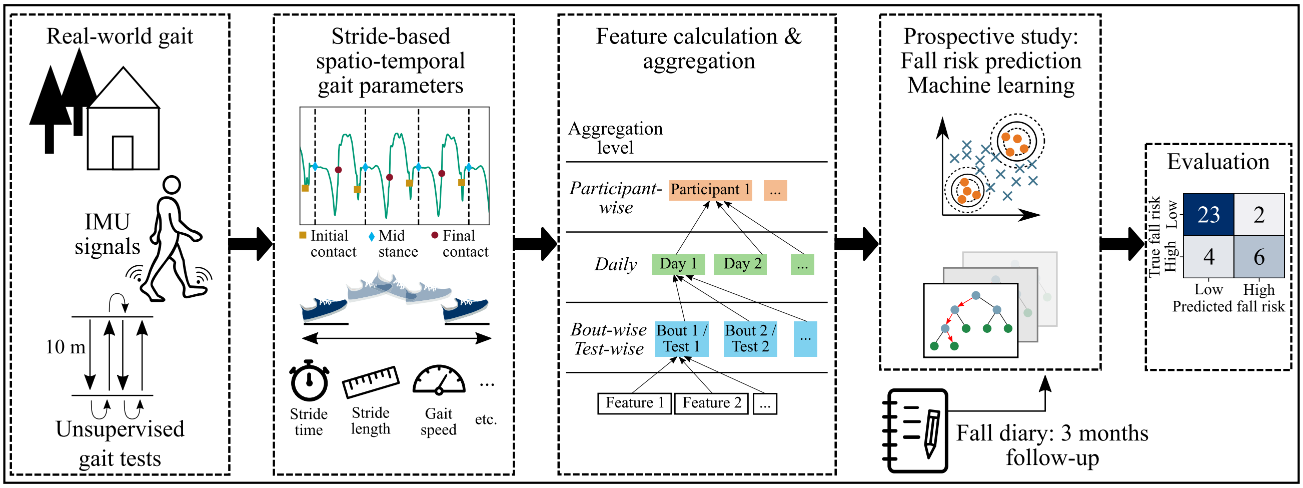

Fig. 1. Overview of the algorithmic approach from IMU raw data of real-world gait and *4x10MWT*s to fall risk prediction.

TABLE I
PATIENT CHARACTERISTICS (N = 35) AFTER EXCLUDING 5 PARTICIPANTS DUE TO TECHNICAL ISSUES (SECTION II-A). PARAMETERS ARE GIVEN AS MEAN ± STANDARD DEVIATION OR COUNTS IN CASE OF SEX. SIGNIFICANT p-VALUES ($\alpha = 0.05$) OF PAIRWISE T-TESTS ARE HIGHLIGHTED IN BOLD

Characteristic	Non-Faller (N = 25)	Prospective Faller (N = 10)	p
Sex (m / f)	20 / 5	6 / 4	0.292
Age [years]	63.6 ± 8.4	64.5 ± 8.2	0.785
Height [cm]	175.3 ± 8.0	173.1 ± 10.0	0.547
Weight [kg]	79.4 ± 12.8	68.6 ± 14.8	0.061
BMI [kg/m^2]	25.9 ± 3.9	22.9 ± 5.2	0.136
UPDRS-III	12.9 ± 6.2	21.7 ± 9.1	**0.016**
Hoehn & Yahr	2.3 ± 0.5	2.9 ± 0.5	**0.013**

the performed gait tests, respectively. With the term real-world gait we refer to free-living, unsupervised and unconstrained walking that could take place inside or outside the home of the participants and is distinguished from standardized tests, like the *4x10MWT*, as defined by Kluge et al. [19]. To identify macroscopic sequences of real-world gait from the raw sensor signals, an algorithm based on harmonic frequency patterns, as presented in our previous work, was applied [35]. Start and end time points of the *4x10MWT*s were manually labeled by two trained human annotators who visually inspected the raw data with assistance of the participants' annotations as previously described [29].

Real-world gait sequences and gait tests were further processed with the same pipeline. First, single strides were identified with a stride segmentation approach using subsequence Dynamic Time Warping as presented by Barth et al. [36]. Gait events, namely initial contact (IC), mid stance (MS), and final contact (FC), and in total 8 spatio-temporal micro gait parameters (stride time, stance time, swing time, stride length, gait speed, IC foot angle, FC foot angle, maximum foot lift) were then calculated using the gait event detection and de-drifted double integration approach of Rampp et al. [37].

The parameterized strides of the real-world gait part were then systematically filtered and assembled to walking bouts using the following criteria. A stride could have a duration between 0.5 and 2 s and a length between 0.3 and 2 m, where the maximum turning angle was set to 45°. These criteria allowed for a wide range around values that are typically measured for PD [38], [39]. Strides who passed these inclusion criteria were assembled to walking bouts, where bouts had a minimum number of four strides and ended after a maximum break of 3 s [19]. Pairwise group differences of the 8 gait parameters measured from participants with *high* and *low* fall risk in real-world gait and gait tests were assessed using Mann-Whitney-U-tests with a significance level of 0.05. A Bonferroni correction for multiple comparisons ($n = 8$) was applied to the p-values.

C. Data Aggregation

According to the goal of this study to compare the predictive power of different aggregation levels of real-world gait and gait tests for predicting fall risk, different feature sets were created and will be described in the following.

1) Real-World Gait: Three different feature sets were created from the real-world gait data based on the three different aggregation levels *bout-wise*, *daily* and *participant-wise*.

Bout-wise: Given the parameterized strides of single walking bouts, a feature vector with spatio-temporal gait parameters was created, similar as described by Tunca et al. [26]. For each walking bout, the respective micro gait parameter values of the single strides were aggregated by computing the mean, standard deviation (SD) and coefficient of variation (CV). Additionally, the asymmetry between left and right foot was determined for stance time, swing time, IC angle, FC angle, and maximum foot lift by calculating the absolute differences between the mean values from left and right strides. By adding the number of strides and the duration of the walking bout, a feature vector of in total 31 features was available for each walking bout. To compensate differences of the number of walking bouts between participants, and keep the time for training the machine learning models within a reasonable limit, a random sub-sample of 100 walking bouts per participant was used for the classification.

Daily: For the *daily* aggregation, the micro parameters explained in the *bout-wise* feature set were computed over the strides of all walking bouts of a day. Additionally, we calculated macro parameters including the number of strides and walking bouts in total, as well as the overall time spent walking and the average and maximum walking bout duration. The gait speed capacity during the day was derived by determining a lower (5th percentile) and upper (95th percentile) boundary of the average gait speed of all walking bouts, as well as the difference between those values. In total, 37 features were available in the *daily* aggregation approach.

Participant-wise: In order to merge all data of a participant into a single *participant-wise* feature vector, the *daily* features were aggregated by calculating the mean, SD, maximum (Max.), and difference between maximum and minimum (Max.-Min.) over all *daily* features. This resulted in 148 features per participant (37 features from *daily* × 4 summary measures) for the *participant-wise* feature set.

2) Gait Tests: For the *4x10MWT*s a similar aggregation strategy as for the real-world gait data was applied. We created three different feature sets based on the three aggregation levels *test-wise*, *daily*, and *participant-wise*.

Test-wise: The same micro gait parameters as for the *bout-wise* approach were calculated on every single *4x10MWT*, except for the number of strides and the duration, as those were limited by the *4x10MWT* protocol. Hence, 29 features were available for the *test-wise* feature set.

Daily: The same 29 features as for the *test-wise* feature set were used, but aggregated from all strides of the *4x10MWT*s available on a day for the *daily* feature set.

Participant-wise: Similar to the real-world *participant-wise* feature set, we aggregated the *daily* gait test features by computing the mean, SD, Max., and Max.-Min. over the single days for each participant such that only one multi-dimensional data point per patient was determined. This resulted in one feature vector with 116 features for each participant (29 features from *daily* × 4 summary measures).

D. Classification Experiments

We performed equivalent classification experiments with all six described feature sets described above. The ground truth labels *high* and *low* fall risk were determined by the results of the fall diaries, as explained above. Different combinations of feature selection and classification methods were compared using the BioPsyKit (version 0.3.6) [40] and scikit-learn (version 0.24.2) [41] libraries and were implemented in Python 3.7. The goal was to find the best combination of methods for each feature set with respect to the classification performance. As feature selection methods we used Select-K-Best based on the ANOVA F-value and Recursive Feature Elimination (RFE) using a linear Support Vector Machine (SVM) as estimator [41]. The optimal number of features to select was treated as a hyperparameter which was optimized as described below. For the classification we compared the following methods: Naive Bayes (NB), SVM with linear and radial basis function kernels (SVM-lin, SVM-rbf), Random Forest (RF), and Gradient Boosting Classification (GB).

TABLE II
GRID SEARCH SPACE FOR THE INDIVIDUAL CLASSIFIERS: SVM WITH LINEAR AND RADIAL BASIS FUNCTION KERNELS (SVM-LIN, SVM-RBF), RANDOM FOREST (RF), AND GRADIENT BOOSTING (GB). THE NAIVE BAYES CLASSIFIER DOES NOT HAVE ANY PARAMETERS THAT NEEDED TO BE OPTIMIZED. THE PARAMETER NAMES ARE GIVEN AS SPECIFIED BY SCIKIT-LEARN [41]

		Values
Classifier	Parameter	Values
SVM-lin	`C`	{0.1, 1, 10, ..., 10000}

SVM-rbf	`C`
	`gamma`
		{0.1, 1, 10, ..., 10000}
		{1, 0.1, 0.01, ..., 0.0001}

RF	`n_estimators`
	`max_depth`
	`min_samples_split`
	`min_samples_leaf`
	`bootstrap`
		{200, 400, 600, ..., 2000}
		{10, 20, 30, ..., 110, None}
		{2, 5, 10}
		{1, 2, 4}
		{True, False}

GB	`n_estimators`
	`max_depth`
	`min_samples_split`
	`min_samples_leaf`
		{200, 400, 600, ..., 2000}
		{10, 20, 30, ..., 110, None}
		{2, 5, 10}
		{1, 2, 4}

TABLE III
THE NUMBER OF FEATURES n_{FEATURES}, THE NUMBER OF SAMPLES n_{SAMPLES} AND THE GRID SEARCH SPACE FOR `k` AND `n_features_to_select` IN THE FEATURE SELECTION WITH SELECT-K-BEST AND RECURSIVE FEATURE ELIMINATION [41] IN THE **REAL-WORLD GAIT** EXPERIMENTS

	n_{features}	n_{samples}	`k` / `n_features_to_select`
Participant-wise	148	35	{2, 3, 4, ..., 34, all}
Daily	37	385	{2, 3, 5, 8, 13, 21, 34, all}
Bout-wise	31	55965	{2, 3, 5, 8, 13, 21, all}

The different approaches were evaluated using leave-one-participant-out cross-validation (LOPO-CV). Hence, for each combination of feature selection method and classifier, 35 models were trained using the data of 34 participants and tested on one left-out participant. For each participant, one ground truth label (*high* or *low* fall risk) was available. Therefore, the label of each participant was mapped to all their samples for the lower levels of aggregation (*bout-wise*, *test-wise*, *daily*) for training. For these aggregation levels, a majority vote on the predicted labels determined the final prediction of a participant belonging to the *high* or *low* fall risk class.

All features were standardized by subtracting the mean and scaling to unit variance within the cross-validation. Within each fold of the LOPO-CV, hyperparameters of the feature selection and classifier were optimized using a stratified five-fold cross validation with grid search for NB, SVM-lin, and SVM-rbf and randomized search with ten iterations for RF and GB. The range for the different parameters of the classifiers are given in Table II.

For the hyperparameter of the number of features to be selected by the feature selection, we used a grid starting from 2 up to all features, where the maximum number of features should not exceed the total number of samples in the respective training data set (Tables III and IV). In case of the highest aggregation

TABLE IV
THE NUMBER OF FEATURES n_{FEATURES}, THE NUMBER OF SAMPLES n_{SAMPLES} AND THE GRID SEARCH SPACE FOR k AND n_features_to_select IN THE FEATURE SELECTION WITH SELECT-K-BEST AND RECURSIVE FEATURE ELIMINATION [41] IN THE **GAIT TEST** EXPERIMENTS

	n_{features}	n_{samples}	k / n_features_to_select
Participant-wise	116	35	{2, 3, 4, ..., 34, all}
Daily	29	422	{2, 3, 5, 8, 13, 21, all}
Test-wise	29	1059	{2, 3, 5, 8, 13, 21, all}

level (*participant-wise*) with only 34 samples in the training set, a dense grid with all possible integer values between 2 and 34 was inspected for the real-world and the gait test data. The other aggregation levels contained considerably more samples, thus, they required a significantly higher model training time when using the dense grid. Therefore, we used Fibonacci numbers for the possible number of selected features to keep training times at a reasonable level. The Fibonacci sequence has a higher density in the lower value range, thus preferring a lower number of features to be selected, which is beneficial with respect to the curse of dimensionality [42]. Finally, the option of avoiding the feature selection and thus using all available features was included.

The balanced accuracy, which is the mean of sensitivity and specificity, was used as the target metric for the parameter optimization, to avoid a bias because of an unbalanced class distribution in the data set. Using the trained models, the predictions were performed on the respective test data of the corresponding LOPO-CV folds. Based on the resulting predicted labels and the true labels of the outer cross-validation we derived a confusion matrix for each combination of feature selector and classifier, from which metrics like balanced accuracy, sensitivity, and specificity were calculated.

III. RESULTS

The data set contained sensor recordings from 422 days, with a total number of 1059 *4x10MWT*s, 55965 real-world walking bouts, and 1444985 parameterized strides. The range of recorded days per patient was between a minimum of 6 and a maximum of 16. Between 7 and 43 *4x10MWT*s were available for each participant, where on average 2.51 were performed per day. The total number of walking bouts ranged from 371 to 4525 over all patients. Within the three-months follow-up period, 10 of the 35 participants reported at least one fall, where 5 participants experienced 1 fall and the other 5 participants between 2 and up to 7 falls. Reviewing the fall diaries revealed that falls occurred mainly while getting up from a chair, walking on uneven terrain (e.g. in the forest) or because of balance loss during housework or gardening.

From the different feature sets, real-world gait with the *participant-wise* aggregation resulted in the highest classification performance with a balanced accuracy of 74.0% (sensitivity: 60.0%, specificity: 88.0%) in the LOPO-CV (Table V). This best result was achieved with the combination of the Select-K-Best feature selector and the RF classifier. The *participant-wise* aggregation also provided the highest balanced accuracy of 68.0% (sensitivity: 40.0%, specificity: 96.0%) for the gait test data in the LOPO-CV (Table VI). The confusion matrices for the best combination of feature selector and classifier for each experimental condition show the distribution of correct and false predictions on the two classes (Fig. 2). In the *participant-wise* aggregation, 11 features were selected by the Select-K-Best feature selector in more than 30 folds of the LOPO-CV, including summary and variability measures of gait parameters (Table VII). The real-world measurements show decreased gait speed and stride length compared to the gait test results, where the participants with *high* fall risk have significantly lower values in these and most of the other parameters (Tables VIII and IX).

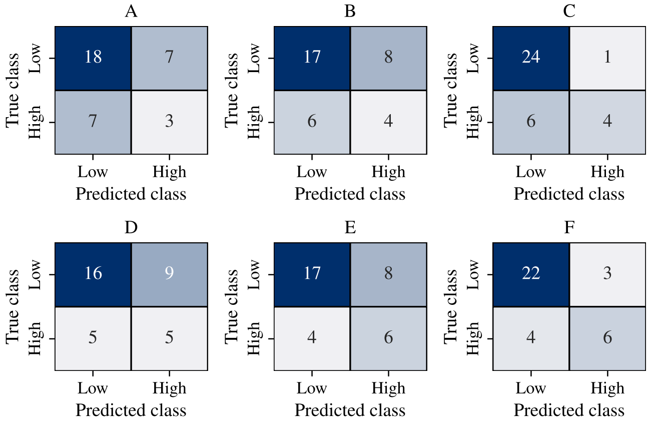

Fig. 2. Confusion matrices of the leave-one-participant-out cross-validation for the best feature selector-classifier combinations of the six experimental conditions (Tables V and VI). Top: **Gait tests** aggregated test-wise (a), daily (b), and participant-wise (c). Bottom: **Real-world gait** aggregated bout-wise (d), daily (e), and participant-wise (f). The numbers in the cells correspond to the respective number of classified participants.

IV. DISCUSSION

In this paper we predicted the future fall risk of PD patients using unsupervised IMU-based gait measurements and different machine learning approaches. We compared features from recordings of unconstrained real-world gait with unsupervised 4x10-Meter-Walking-Tests (*4x10MWT*) by means of a data set with 35 PD patients over two weeks. A specific focus was set on different data aggregation strategies for the sensor data of single walking bouts or *4x10MWT*s, whole days, and the entire data of a participant. Our results showed that real-world gait data that were aggregated on a *participant-wise* level, achieved the highest balanced accuracy in the prediction of fall risk.

A. Fall Risk Prediction

The data set used in this study was recorded over multiple days, allowing to capture the gait and mobility capacity of PD patients in their daily life. The patient characteristics showed significant differences between future fallers and non-fallers for

TABLE V
Classification Results of the **Real-World Gait** Based on the Leave-One-Participant-Out Cross-Validation in [%]. For Each Aggregation Level, the Feature Selector-Classifier Combination Achieving the Best Classification Performance is Shown

	Feature Selector	Classifier	Balanced Accuracy	Sensitivity	Specificity
Participant-wise	Select-K-Best	RF	74.0	60.0	88.0
Daily	Select-K-Best	SVM-rbf	64.0	60.0	68.0
Bout-wise	RFE	SVM-lin	57.0	50.0	64.0

TABLE VI
Classification Results of the **Gait Tests** Based on the Leave-One-Participant-Out Cross-Validation in [%]. For Each Aggregation Level, the Feature Selector-Classifier Combination Achieving the Best Classification Performance is Shown

	Feature Selector	Classifier	Balanced Accuracy	Sensitivity	Specificity
Participant-wise	Select-K-Best	GB	68.0	40.0	96.0
Daily	RFE	SVM-rbf	54.0	40.0	68.0
Test-wise	Select-K-Best	SVM-lin	51.0	30.0	72.0

TABLE VII
The List of Features That Were Selected by Select-K-Best [41] in at Least 30 of the 35 Folds of the Leave-One-Participant-Out Cross-Validation for the Real-World *Participant-Wise* Experiment. Outer Summary Measures in Bold Font Correspond to the *Patient-Wise* Aggregation of the Inner Features (Italic Font) That Were Computed Over All Walking Bouts per Day

Feature	Count
Max. (*CV* (Stride length))	35
SD (*Mean* (Swing time))	35
Max.-Min. (*Mean* (Swing time))	35
Max. (*CV* (Max. foot lift))	33
SD (*CV* (Max. foot lift))	33
Max. (*Mean* (IC angle))	33
Mean (*CV* (Stride length))	33
Max. (*CV* (Swing time))	32
Max.-Min. (*CV* (Max. foot lift))	32
SD (*CV* (Swing time))	30
Mean (*CV* (Swing time))	30

TABLE VIII
Spatio-Temporal Gait Parameters of **Real-World Gait** Walking Bouts of Participants With *High* ($n = 12396$ Walking Bouts) and *Low* ($n = 43569$ Walking Bouts) Fall Risk. Values of Gait Parameters are Given as Median (Inter-Quartile Range). Bonferroni-Corrected p Values are Given for Pairwise Mann-Whitney-U-Tests Between the Fall Risk Groups

	High fall risk	*Low* fall risk	p
Gait speed [m/s]	0.66 (0.34)	0.77 (0.38)	< 0.001
Stride length [m]	0.79 (0.44)	0.95 (0.37)	< 0.001
Stride time [s]	1.21 (0.31)	1.22 (0.25)	< 0.001
Swing time [s]	0.40 (0.11)	0.41 (0.07)	< 0.001
Stance time [s]	0.81 (0.24)	0.81 (0.19)	0.01
IC angle [°]	-4.30 (11.15)	-9.64 (10.21)	< 0.001
FC angle [°]	48.62 (17.20)	50.92 (16.06)	< 0.001
Max. foot lift [m]	0.05 (0.04)	0.06 (0.03)	< 0.001

disease-specific assessments, with higher values for the UPRDS-III and Hoehn & Yahr scales (Table I). Previous studies [11], [43] also reported such differences and partially used these scores for fall risk prediction. Still, the strength of our study is the identification mobility impairments related to fall risk using solely sensor-derived gait parameters that do not require clinical expert ratings and give an objective reflection of the patients' daily live.

TABLE IX
Spatio-Temporal Gait Parameters of the **Gait Tests** of Participants With *High* ($n = 289$ Gait Tests) and *Low* ($n = 770$ Gait Tests) Fall Risk. Values of Gait Parameters are Given as Median (Inter-Quartile Range). Bonferroni-Corrected p Values are Given for Pairwise Mann-Whitney-U-Tests Between the Fall Risk Groups

	High fall risk	*Low* fall risk	p
Gait speed [m/s]	0.97 (0.33)	1.00 (0.31)	< 0.001
Stride length [m]	1.06 (0.32)	1.18 (0.28)	< 0.001
Stride time [s]	1.14 (0.13)	1.15 (0.17)	0.30
Swing time [s]	0.40 (0.06)	0.41 (0.05)	0.01
Stance time [s]	0.73 (0.11)	0.74 (0.12)	0.56
IC angle [°]	-11.22 (11.65)	-15.52 (10.08)	< 0.001
FC angle [°]	56.43 (11.12)	59.64 (12.54)	0.42
Max. foot lift [m]	0.06 (0.03)	0.07 (0.02)	< 0.001

The participants were instructed to perform three unsupervised *4x10MWT*s every day, that served as a second data source besides real-world gait, with conditions that are closely related to the well-known laboratory-based gait tests [28]. Thus, observations of various spatio-temporal gait parameters were available over multiple days points from standardized and real-world conditions (Tables VIII and IX). An additional strength of our data set is the prospective information of actual falls. Retrospective [14], [16], [26] and prospective [17], [25], [27] approaches have previously been used to determine fall risk, where the prospective prediction is considered to be more accurate [26].

In earlier studies, traditional statistical significance tests have been applied to determine group differences for fallers and non-fallers based on gait parameters measured in the laboratory [14], [16] or in the real world [20], [22]. Recent publications have shown the predictive power of machine learning, allowing an individual fall risk prediction using complex multidimensional data available from a long-term monitoring study [25], [27].

In contrast to [25] and [26], we did not consider deep learning approaches for this study as we aimed to use the same pool of methods for all tested data aggregation levels. Given the cohort size of 35 participants, the number of samples is small compared to the generally recommended sample size for artificial neural networks [44]. Therefore, we compared established traditional machine learning methods.

The highest classification performance in our study was achieved using the real-world gait data aggregated on the *participant-wise* level with a balanced accuracy of 74.0% with a sensitivity of 60.0% and a specificity of 88.0%. These results underline the finding of previous work, that a prospective prediction of fall risk is possible using machine learning approaches and IMU-based gait recordings. A fair comparison of our results with the related work is challenging as there are large differences in cohorts, study protocols, classification approaches and reported performance metrics. In the study of Punt et al. [27], gait recordings of stroke patients in the real-world allowed a classification of future fallers with a sensitivity of 85%. The study of Howcroft et al. [23] used only laboratory data and reported an accuracy of 57% for their best performing neural network model for prospective fall risk prediction. Aicha et al. [25] only reported the AUC values for their deep learning experiments for prospective fall risk prediction with real-world raw sensor data, reaching a value of 0.75. In the recent paper of Tunca et al. [26], an accuracy value of 89% was achieved for retrospective fall risk classification using deep learning and gait recorded from straight walking in a supervised setting and a cohort of 76 subjects. Despite the fact that the mentioned studies all used different study protocols and cohorts, the classification performance of the best configuration in our study can in general be considered comparable to other state-of-the-art approaches. Nevertheless, the sensitivity of 60.0% is relatively low compared to the strong specificity of 88.0% and hence there is room for improvement regarding further optimization of feature engineering and model training. Still, for this work the focus was set on the investigation and comparison of the potential of the presented methods for data aggregation and data acquisition during gait tests and in the real world. Therefore, targeting a maximum balanced accuracy was a reasonable choice at this stage.

B. Comparison of Aggregation Levels

One goal of this study was the investigation of different aggregation levels given the long-term monitoring data. In related work, signals have either been analyzed with a fixed window size [25], [27], or by accumulating or averaging information over the entire recording duration [20], [22].

As we had gait parameters derived from real-world walking bouts and unsupervised *4x10MWT*s available over multiple days, we were able to compare features of single bouts or tests, daily aggregates, or participant-wise aggregates. The results indicate that for real-world gait and gait tests, the classification performance increased with the level of aggregation and the *participant-wise* data yields the highest balanced accuracy of 74.0% for real-world gait and 68.0% for the gait tests (Tables V and VI).

A possible explanation can be given by the relation of the granularity of features and labels. The ground truth label is only available on the participant level and there are no detailed fall risk labels for single walking bouts, *4x10MWT*s, or days. Therefore, similar to previous studies [25], [26], a mapping of the label to the single samples for model training and a majority vote of the predictions for testing had to be performed to achieve a single prediction per participant in all aggregation levels. According to our results, the fall risk is represented best by incorporating all available data of a participant. The data on the finer granularity levels is most likely inhering a higher variance with respect to the class label and can thus be considered more noisy. The results suggest that fall risk is rather a global item of information that can be identified most reliably using all available data at once. The high importance of SD and Max.-Min.-features, representing the overall inter-day variability, underlines the finding that fall risk prediction requires summarizing all available recordings of a participant (Table VII). The frequently selected CV-related features (e.g. of stride length or swing time) hint towards the influence of gait variability to fall risk, as also found in [20], [22].

C. Comparison Real-World vs. Gait Tests

The second main objective of the study was the comparison between prediction of fall risk using features derived from real-world gait and *4x10MWT*s. Both data were recorded by the same participants during the same time period in continuous recordings, so the general circumstances of the two data sources can be considered similar.

The prediction based on real-world gait data yielded higher performance values than the gait tests in all corresponding aggregation levels, for example 74.0% vs. 68.0% balanced accuracy in the *participant-wise* aggregation (Tables V and VI). As discussed in the literature, real-world gait data, especially when aggregated over one or multiple days, have the advantage to cover not only micro but also macro information of gait like the number and duration of walking bouts [21]. In addition to the spatio-temporal gait parameters that are measured in the *4x10MWT*s, these macro measures enclose higher level information about the behavior and volume of activity which have been reported to be influential to fall risk [9].

To the best of our knowledge, this is the first study providing real-world gait and gait tests continuously recorded over multiple consecutive days. In the study of Punt et al. [27], a comparison of fall risk assessment with a gait test in the laboratory and real-world gait recorded over one week was performed on a cohort of stroke patients. Similar to our study, the authors of [27] came to the conclusion that the advantages of long-term real-world data compensate potential difficulties and insecurities created by the unsupervised character of the data.

Even though unsupervised *4x10MWT*s did not prove to be a reliable source of data for fall risk prediction in this study, they still provide an efficient and helpful tool for long-term patient monitoring [28]. The execution of the *4x10MWT* only takes small effort in terms of time and required space and the data can be automatically processed [29].

D. Limitations

The unsupervised character of the study is causing several limitations. Even though the ability of operating the recording system and following the study protocol was considered in the inclusion criteria, not all participants provided the same amount of data. There are differences regarding the number of recordings and the number of gait test executions per day. Reasons for missing recordings may relate to technical and usability difficulties of the participants when handling the recording, while missing gait tests could be caused by the study protocol's additional burden onto the daily challenges of the PD patients. We limited the influence of these differences by setting a minimum wear time and performing a majority vote for the predicted fall risk label in the lower aggregation levels. As discussed in our previous work, exact adherence of the patients to the protocol of the standardized gait tests can not be ensured [29]. Nevertheless, the unsupervised tests provide an additional source of gait data that is to a high degree standardized in contrast to the unconstrained real-world walking. Moreover, the exact medication intake times during the home monitoring phase were not tracked. Hence, the impact of ON/OFF periods on the gait parameters and specifically on fall risk prediction is not known and should be investigated in future work. In addition, the influence emotional and cognitive functions of the patients on the gait performance was not considered.

Furthermore, the cohort of 35 participants with 10 fallers is comparably small, and an extension of the data set would most likely help to improve the fall risk prediction with more reliable classification models. By using the balanced accuracy as the main performance metric, the imbalance between participants with high and low fall risk was compensated. In addition, the selection of classifiers was limited to the presented five methods, where further approaches from the literature could be employed. Nevertheless, the used methods, representing probabilistic, maximum margin, and ensemble classifiers, give a good impression of the potential of fall risk prediction with the given data set and aggregation strategies.

V. Conclusion and Outlook

In this study, we investigated different machine learning-based approaches for fall risk prediction in PD. We compared different levels of aggregation of the long-term IMU-derived gait parameters including real-world gait data and unsupervised 4x10-Meter-Walking-Test recordings. The highest aggregation level provided the best classification performance for the real-world (74.0% balanced accuracy) and the gait test data (68.0% balanced accuracy). These values are comparable to the accuracy of fall risk prediction presented in previous studies.

In future research, experiments with gradually increased number of included daily recordings could be performed to acquire a better understanding for the number of days that is actually required to be aggregated for a reliable fall risk prediction. These experiments were not possible with our data set, where the available data quantity between participants varied most likely due to user errors when operating the recording system. For applications of machine learning-based fall risk prediction in clinical practice, further research should also investigate the explainability of classification pipelines to enable clinicians making conclusive therapy decisions using the prediction outcomes.

Furthermore, there is currently no possibility for entirely fair comparisons between different fall risk prediction approaches in the literature from an algorithm perspective. Thus, the goal should be to establish a specific benchmark data set and agree on performance metrics or apply a unified study protocol for future real-world studies. The latter would also allow the fusion of different data sets and ultimately enable the efficient application of deep learning techniques.

Our study is the first one applying data from the real world and unsupervised *4x10MWT*s recorded over multiple days for fall risk prediction. The cohort we investigated consisted only of PD patients, but the general findings, especially regarding the data aggregation is certainly also valuable for any future long-term real-world gait monitoring study. Therefore, we are closing an open research gap with the goal of advancing sensor-based gait analysis on the way from the laboratory to the real world. In conclusion, we believe that our study makes an important contribution to the field of real-world gait analysis and will ultimately help to avoid falls through an improved preventive treatment of the patients and an early fall risk prediction.

Acknowledgment

The authors would like to thank M. Südmeyer, A. Amouzandeh, M. Graap, M. Winterholler, S. Zimmet, R. Zagel, T. Reichhardt and T. Greinwalder for their effort in recording the data set as well as all participants of the study for their contributions. Content in this publication reflects the authors' view and neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained herein.

References

[1] B. Wood et al., "Incidence and prediction of falls in Parkinson's disease: A prospective multidisciplinary study," *J. Neurol. Neurosurgery Psychiatry*, vol. 72, no. 6, pp. 721–725, 2002.

[2] N. E. Allen et al., "Recurrent falls in Parkinson's disease: A systematic review," *Parkinson's Dis.*, vol. 2013, 2013, Art. no. 906274.

[3] S. Paul et al., "Fall-related hospitalization in people with Parkinson's disease," *Eur. J. Neurol.*, vol. 24, no. 3, pp. 523–529, 2017.

[4] N. Mühlenfeld et al., "Fractures in Parkinson's disease: Injury patterns, hospitalization, and therapeutic aspects," *Eur. J. Trauma Emerg. Surg.*, vol. 47, no. 2, pp. 573–580, 2021.

[5] H. Brozova et al., "Fear of falling has greater influence than other aspects of gait disorders on quality of life in patients with Parkinson's disease," *Neuroendocrinol. Lett.*, vol. 30, no. 4, pp. 453–457, 2009.

[6] M. A. van der Marck et al., "Consensus-based clinical practice recommendations for the examination and management of falls in patients with Parkinson's disease," *Parkinsonism Related Disord.*, vol. 20, no. 4, pp. 360–369, 2014.

[7] S. A. Parashos et al., "What predicts falls in Parkinson disease? Observations from the Parkinson's foundation registry," *Neurol. Clin. Pract.*, vol. 8, no. 3, pp. 214–222, 2018.

[8] S. Lord et al., "Predicting first fall in newly diagnosed Parkinson's disease: Insights from a fall-naïve cohort," *Movement Disord.*, vol. 31, no. 12, pp. 1829–1836, 2016.

[9] A. Fasano et al., "Falls in Parkinson's disease: A complex and evolving picture," *Movement Disord.*, vol. 32, no. 11, pp. 1524–1536, 2017.

[10] N. Roth et al., "Real-world stair ambulation characteristics differ between prospective fallers and non-fallers in Parkinson's disease," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 9, pp. 4733–4742, Sep. 2022.

[11] G. K. Kerr et al., "Predictors of future falls in Parkinson disease," *Neurology*, vol. 75, no. 2, pp. 116–124, 2010.

[12] S. S. Paul et al., “Three simple clinical tests to accurately predict falls in people with Parkinson’s disease,” *Movement Disord.*, vol. 28, no. 5, pp. 655–662, 2013.

[13] M. H. Cole et al., “Falls in Parkinson’s disease: Kinematic evidence for impaired head and trunk control,” *Movement Disord.*, vol. 25, no. 14, pp. 2369–2378, 2010.

[14] B. R. Greene, A. O’Donovan, R. Romero-Ortuno, L. Cogan, C. N. Scanaill, and R. A. Kenny, “Quantitative falls risk assessment using the timed up and go test,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 12, pp. 2918–2926, Dec. 2010.

[15] J. Silva and I. Sousa, “Instrumented timed up and go: Fall risk assessment based on inertial wearable sensors,” in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, 2016, pp. 1–6.

[16] M. D. Latt et al., “Acceleration patterns of the head and pelvis during gait in older people with Parkinson’s disease: A comparison of fallers and nonfallers,” *J. Gerontol. Ser. A: Biomed. Sci. Med. Sci.*, vol. 64, no. 6, pp. 700–706, 2009.

[17] T. Doi et al., “The harmonic ratio of trunk acceleration predicts falling among older people: Results of a 1-year prospective study,” *J. Neuroeng. Rehabil.*, vol. 10, no. 1, pp. 1–6, 2013.

[18] M. Sica et al., “Continuous home monitoring of Parkinson’s disease using inertial sensors: A systematic review,” *PLoS One*, vol. 16, no. 2, 2021, Art. no. e0246528.

[19] F. Kluge et al., “Consensus based framework for digital mobility monitoring,” *PLoS One*, vol. 16, no. 8, 2021, Art. no. e0256541.

[20] A. Weiss et al., “Objective assessment of fall risk in Parkinson’s disease using a body-fixed sensor worn for 3 days,” *PLoS One*, vol. 9, no. 5, 2014, Art. no. e96675.

[21] S. D. Din et al., “Free-living monitoring of Parkinson’s disease: Lessons from the field,” *Movement Disord.*, vol. 31, no. 9, pp. 1293–1313, 2016.

[22] S. D. Din et al., “Analysis of free-living gait in older adults with and without Parkinson’s disease and with and without a history of falls: Identifying generic and disease-specific characteristics,” *J. Gerontol.: Ser. A*, vol. 74, no. 4, pp. 500–506, 2019.

[23] J. Howcroft, J. Kofman, and E. D. Lemaire, “Prospective fall-risk prediction models for older adults based on wearable sensors,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1812–1820, Oct. 2017.

[24] B. M. Meyer et al., “Wearables and deep learning classify fall risk from gait in multiple sclerosis,” *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1824–1831, May 2021.

[25] A. N. Aicha et al., “Deep learning to predict falls in older adults based on daily-life trunk accelerometry,” *Sensors*, vol. 18, no. 5, 2018, Art. no. 1654.

[26] C. Tunca, G. Salur, and C. Ersoy, “Deep learning for fall risk assessment with inertial sensors: Utilizing domain knowledge in spatio-temporal gait parameters,” *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1994–2005, Jul. 2020.

[27] M. Punt et al., “Do clinical assessments, steady-state or daily-life gait characteristics predict falls in ambulatory chronic stroke survivors?,” *J. Rehabil. Med.*, vol. 49, no. 5, pp. 402–409, 2017.

[28] H. Gaßner et al., “Clinical relevance of standardized mobile gait tests. Reliability analysis between gait recordings at hospital and home in Parkinson’s disease: A pilot study,” *J. Parkinson’s Dis.*, vol. 10, pp. 1763–1773, 2020.

[29] M. Ullrich et al., “Detection of unsupervised standardized gait tests from real-world inertial sensor data in Parkinson’s disease,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2103–2111, 2021.

[30] M. Ullrich et al., “FallRiskPD dataset,” 2022. [Online]. Available: osf.io/h6apq

[31] M. M. Hoehn et al., “Parkinsonism: Onset, progression, and mortality,” *Neurology*, vol. 50, no. 2, pp. 318–318, 1998.

[32] N. Roth et al., “Hidden Markov model based stride segmentation on unsupervised free-living gait data in Parkinson’s disease patients,” *J. Neuroeng. Rehabil.*, vol. 18, no. 1, pp. 1–15, 2021.

[33] C. G. Goetz et al., “Movement disorder society-sponsored revision of the unified Parkinson’s disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric assessment,” *Movement Disord.*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[34] S. E. Lamb et al., “Development of a common outcome data set for fall injury prevention trials: The prevention of falls network Europe consensus,” *J. Amer. Geriatr. Soc.*, vol. 53, no. 9, pp. 1618–1622, 2005.

[35] M. Ullrich et al., “Detection of gait from continuous inertial sensor data using harmonic frequencies,” *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1869–1878, Jul. 2020.

[36] J. Barth et al., “Stride segmentation during free walk movements using multi-dimensional subsequence dynamic time warping on inertial sensor data,” *Sensors*, vol. 15, no. 3, pp. 6419–6440, 2015.

[37] A. Rampp, J. Barth, S. Schülein, K.-G. Gaßmann, J. Klucken, and B. M. Eskofier, “Inertial sensor-based stride parameter calculation from gait sequences in geriatric patients,” *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1089–1097, Apr. 2015.

[38] J. C. Schlachetzki et al., “Wearable sensors objectively measure gait parameters in Parkinson’s disease,” *PLoS One*, vol. 12, no. 10, 2017, Art. no. e0183989.

[39] S. D. Din et al., “Free-living gait characteristics in ageing and Parkinson’s disease: Impact of environment and ambulatory bout length,” *J. Neuroeng. Rehabil.*, vol. 13, no. 1, pp. 1–12, 2016.

[40] R. Richer et al., “BioPsyKit: A Python package for the analysis of biopsychological data,” *J. Open Source Softw.*, vol. 6, no. 66, 2021, Art. no. 3702.

[41] F. Pedregosa et al., “Scikit-learn: Machine learning in python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[43] B. R. Bloem et al., “Prospective assessment of falls in Parkinson’s disease,” *J. Neurol.*, vol. 248, no. 11, pp. 950–958, 2001.

[44] A. Alwosheel et al., “Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis,” *J. Choice Modelling*, vol. 28, pp. 167–182, 2018.