Detection of Unsupervised Standardized Gait Tests From Real-World Inertial Sensor Data in Parkinson's Disease

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

***Abstract*— Gait tests as part of home monitoring study protocols for patients with movement disorders may provide valuable standardized anchor-points for real-world gait analysis using inertial measurement units (IMUs). However, analyzing unsupervised gait tests relies on reliable test annotations by the patients requiring a potentially error-prone interaction with the recording system. To overcome this limitation, this work presents a novel algorithmic pipeline for the automated detection of unsupervised standardized gait tests from continuous real-world IMU data. In a study with twelve Parkinson's disease patients, we recorded real-world gait data over two weeks using foot-worn IMUs. During continuous daily recordings, the participants performed series of three consecutive 4 × 10-Meters-Walking-Tests (4×10MWTs) at different walking speeds, besides their usual daily-living activities. The algorithm first detected these gait test series using a gait sequence detection algorithm, a peak enhancement pipeline, and subsequence Dynamic Time Warping and then decomposed them into single 4 × 10MWTs based on the walking speed. In the evaluation with 419 available gait test series, the detection reached an F1-score of 88.9% and the decomposition an F1-score of 94.0%. A concurrent validity evaluation revealed very good agreement between spatio-temporal gait parameters derived from manually labelled and automatically detected 4 × 10MWTs. Our algorithm allows to remove the burden of system interaction from the patients and reduces the time for manual data annotation for researchers. The study contributes to an improved automated processing of real-world IMU gait data and enables a simple integration of standardized tests into continuous long-term recordings. This will help to bridge the gap between supervised and unsupervised gait assessment.**

***Index Terms*— Machine learning, activity recognition, accelerometer, gyroscope.**

Manuscript received August 3, 2021; revised September 29, 2021; accepted October 7, 2021. Date of publication October 11, 2021; date of current version October 20, 2021. This work was supported in part by the Bavarian Ministry for Economy, Regional Development and Energy via the Medical Valley Award 2017 (FallRiskPD Project), in part by the Innovative Medicines Initiative 2 Joint Undertaking under Agreement 820820, in part by the European Union's Horizon 2020 Research and Innovation Program, in part by the European Federation of Pharmaceutical Industries and Associations (EFPIA), and in part by the German Research Foundation [Deutsche Forschungsgemeinschaft (DFG)]-SFB 1483, under Project 442419336, EmpkinS. The work of Bjoern M. Eskofier was supported by DFG within the framework of the Heisenberg Professorship Program under Grant ES 434/8-1. *(Corresponding author: Martin Ullrich.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Local Ethics Committee at Friedrich-Alexander-University Erlangen-Nuremberg, Germany, under Application No. 165_18B.

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

Till Gladow is with the Medical Valley Digital Health Application Center, 96047 Bamberg, Germany (e-mail: till.gladow@mv-dmac.de).

Heiko Gaßner and Franz Marxreiter are with the Department of Molecular Neurology, University Hospital Erlangen, 91054 Erlangen, Germany (e-mail: heiko.gassner@uk-erlangen.de; franz.marxreiter@uk-erlangen.de).

Jochen Klucken is with the Department of Digital Medicine, University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg, and also with the Luxembourg Institute of Health, 1445 Strassen, Luxembourg (e-mail: jochen.klucken@uni.lu).

Digital Object Identifier 10.1109/TNSRE.2021.3119390

I. Introduction

PARKINSON'S disease (PD) is characterized by movement impairments in general and pathological gait in particular, including the cardinal symptoms tremor, rigidity, bradykinesia, and postural instability [1]. For a clinical assessment of these symptoms, rating scales are being applied, such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS–UPDRS) [2]. However, those scales lack objective quantitative measurements to evaluate motor symptoms and have a limited inter- and intra-rater reliability [3], [4].

Sensor-based gait analysis using wearable inertial measurement units (IMUs) has increasingly been used in clinical settings [5] and long-term monitoring in the real world [6] to provide complementary objective information on movement impairments. The accuracy of sensor-based gait parameter estimations has been validated for patients with movement impairments [7], [8]. Therefore, IMUs provide a comprehensive impression of the patient's condition by means of continuous digital measures such as walking speed or stride length [9], [10]. Several studies have demonstrated that long-term gait recordings have the potential to support monitoring of disease progression and symptoms, for example for the assessment of risk of falling [11]–[13].

Despite the advancements of technology, the translation of the data collected with wearable sensors into a better clinical understanding of the disease and enhanced care for patients remains challenging [14]. The use of standardized protocols

and validation procedures has been recommended to bridge the gap between assessments in standardized clinical and unsupervised real-world settings, to achieve better comparability across studies [15]. Gassner *et al.* suggested specifically the integration of standardized gait tests into home-monitoring study protocols [16]. The standardized tests performed at home may provide anchor-points during continuous recordings which represent the established standardized measurements in clinical settings.

One example for transferring a standardized gait test into the home environment is the instrumented Timed Up and Go test (iTUG) [17]. It includes seven IMUs and an algorithm for automatic detection, separation, and analysis of different components of the TUG, such as sit-to-stand transitions, turning, and walking. The feasibility of performing this test at home was investigated by Zampieri *et al.*, who asessed two iTUGs performed within 24 hours, once in the clinic and once in the patient's home environment [18]. During both tests the patients were supervised and assisted by a researcher.

The 4 × 10-Meters-Walking-Test (4 × 10*MWT*) is another standardized gait test. Spatio-temporal gait parameters, that were derived from with foot-worn IMUs, provided clinically meaning results for PD in a cross-sectional analysis [19]. In a recent pilot study, Gassner *et al.* found a high correlation between gait parameters recorded during supervised 4 × 10*MWT*s in the hospital and at home [16]. The PD patients were instructed by a clinical assessor in both scenarios. Contextual factors like the medication or the time of day were controlled to be similar between the measurements. Gassner *et al.* concluded that unsupervised standardized gait tests should be investigated in future studies as a stepping stone to an assessment of real-world gait.

However, the relation between gait parameters measured in unsupervised gait tests and those measured under clinical supervision remains unclear. One reason why this has not been investigated in detail so far, are the logistical and usability challenges for patients if they need to self-administer these tests [18]. The acquisition of unsupervised gait test data requires an interaction of the patients with the recording system. They need to start and stop the measurement or set time stamp annotations during continuous recordings, for example using a mobile phone app. However, for PD patients, the adherence to study protocols in home monitoring studies including wearable sensors and mobile apps is challenging [20]. Elderly, motor-impaired people experience difficulties in handling mobile touch devices [21] which can cause incorrect or missing time point annotations that require visual inspection and correction of the labels by an investigator. Thus, it would be desirable to perform a continuous whole-day recording and automatically detect performed gait tests during the data processing without any manual annotations by participants or researchers. This way participants could specifically focus on the actual execution of the gait tests and adhere to the study protocol.

Therefore, the purpose of this work was to develop and evaluate an algorithm for the automated detection of standardized gait tests from real-world IMU data, where we focus on series of repeated 4 × 10*MWT*s. To the best of the authors' knowledge, this is the first work providing data and algorithms for fully automated processing of IMU recordings including unsupervised daily-living activities and standardized gait tests. Hence, we also make the very first attempt of an automated detection of standardized gait tests from real-world gait sensor data. Due to the expected heterogeneity of the real-world recordings, data-driven and machine learning methods are most suitable for this task. After successful preliminary experiments with template matching using subsequence Dynamic Time Warping (sDTW) on recordings in the laboratory [22], we also consider sDTW as the core of the approach in this work. Specifically, sDTW allows template matching including parts of the template being stretched or squeezed to achieve an optimal fit [23] and has been successfully applied in gait analysis (e.g. for stride segmentation [24], [25]).

The presented method removes the burden of interaction with the recording system from the patients and reduces manual label inspection and correction work for researchers. Hence, our work significantly contributes to an improved automated processing of real-world gait recordings and enables better comparability across studies by simplifying the implementation of gait tests in real-world monitoring protocols.

TABLE I
PATIENT CHARACTERISTICS (N = 12). PARAMETERS ARE GIVEN AS MEAN ± STANDARD DEVIATION OR COUNTS IN CASE OF THE SEX

Characteristic	
Sex (m / f)	10 / 2
Age [years]	64.9 ± 9.3
Height [cm]	173.2 ± 8.1
Weight [kg]	77.9 ± 13.8
UPDRS-III	17.5 ± 6.6
Hoehn & Yahr	2.6 ± 0.4

II. METHODS AND PROCEDURES

A. Data Acquisition

A data set including twelve patients with idiopathic PD who completed the *FallRiskPD* study (DRKS-ID: DRKS00015085) was used for the development and evaluation of the presented algorithm (Table I). The participants were recruited by the University Hospital Erlangen, the Hospital Rummelsberg, and the Ernst von Bergmann Hospital Potsdam. The study was approved by the local ethics committee Re-No. 165_18B (Friedrich-Alexander-University Erlangen-Nuremberg, Germany). All participants gave written, informed consent, prior to the data collection.

Gait recordings were acquired over two weeks using the *Mobile GaitLab* (Portabiles HealthCare Technologies GmbH, Erlangen, Germany), a wearable sensor system for real-world gait analysis. The system consisted of two IMUs including a 3-d accelerometer (range ±16g), and a 3-d gyroscope (range ±2000°/s). The IMUs were synchronized via wireless continuous synchronization as described in [26]. Data was recorded at a sampling rate of 102.4 Hz. One sensor was mounted on the instep of each shoe and the same shoe model was used for all participants in their respective shoe size. The participants wore the sensor system during their wake time while pursuing their activities of daily living indoors and outdoors. Single full-day recordings were transferred from the sensors to a smartphone during night-time and will be called *daily recording* in the following. In total, the data set contained 151 *daily recordings*.

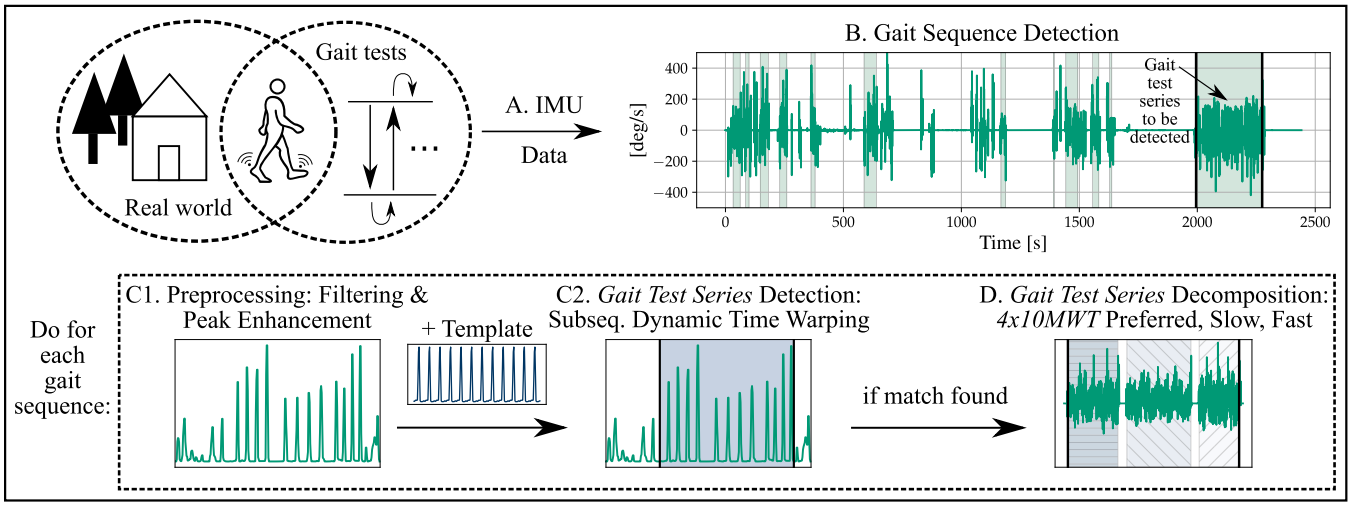

Fig. 1. Pipeline overview: Four main processing steps were required to segment 4 × 10MWTs from real-world recordings.

In addition to activities of daily living, the participants were asked to perform unsupervised standardized gait tests three times a day: in the morning after the start of the *daily recording*, at around noon, and in the evening before stopping the recording. These tests comprised a set of three 4×10*MWT*s performed subsequently with short resting periods in between. One 4×10*MWT* consisted of walking along a straight walkway of 10 m and performing a 180° turn, followed by returning the walkway and another 180° turn. This procedure was then repeated one more time to complete four 10 *m* passes. The three 4 × 10*MWT*s were executed in different speed levels: Deliberately slow (*Slow*), preferred self-selected (*Preferred*), and comparably fast (*Fast*). No particular order was specified for the three walking speeds. The sets of three consecutive tests will be referred to as *gait test series* in the following.

According to the study protocol, the following assumptions were made:

1) Within every 4 × 10*MWT*, a 180° turn was performed after each 10-meter walk.
2) A 180° turn was performed after each 4 × 10*MWT*, resulting in a total of twelve turns in a *gait test series*.
3) Resting periods could occur between successive 4 × 10*MWT*s, but were shorter than 15 s.

There might not always have been a path of exactly 10 m available for the test due to different spatial conditions depending on the participant's environment at the time of test execution. Therefore, the participants were instructed to always choose a straight path with a length as close to 10 m as possible for the test execution. For manual annotations of the start and end times of the *gait test series*, the participants used an adapted version of the *PatientConcept* (NeuroSys GmbH, Ulm, Germany) smartphone application.

B. Gait Sequence Detection

All IMU raw signals were calibrated to yield physically meaningful units using the method of Ferraris *et al.* [27]. The superior-inferior sensor axis was aligned to gravity using static signal windows. To detect the 4 × 10*MWT*s in the *daily recordings*, the sensor data were analyzed in four processing steps, namely gait sequence detection, preprocessing, *gait test series* detection, and *gait test series* decomposition (Fig. 1).

To increase the pipeline efficiency by only considering those parts of the recorded data containing gait, a gait sequence detection was applied first. For this purpose, we used an algorithm proposed in our recent work, where the frequency spectrum of the gyroscope signal of the medio-lateral axis of both foot-worn sensors is investigated regarding harmonic frequency patterns [28]. As resting periods shorter than 15 s could occur between single 4 × 10*MWT*s, adjacent gait sequences, that were not more than 15 s apart, were concatenated. This ensured a *gait test series* to be entirely included in a single gait sequence.

C. Gait Test Series Detection

The detected gait sequences were further investigated regarding the occurrence of *gait test series*. Characteristic features of 4 × 10*MWT*s, differentiating them from daily-living movements, were found to be the regular pattern of turns followed by straight walking periods. The duration of the walks could differ between *gait test series*. Therefore, the turns were determined to be the most valuable feature, represented by peaks in the superior-inferior gyroscope axis signal (gyr_{si}).

1) Preprocessing: To enhance the turning-related peaks, an adapted version of our previously introduced preprocessing pipeline for the detection of gait tests from IMU recordings in the laboratory environment [22] was applied to the gyr_{si} signal (Fig. 2). A 4th order Butterworth low-pass filter with a cutoff frequency of 0.5 Hz was used to only retain the turning-related signal features with frequencies beneath the straight walking human locomotion band [11]. Afterwards, the signal was smoothed using a median filter with a window size of 2 s. Finally, the resulting signal was squared to enhance the turning peaks and to obtain only positive values regardless of the direction of rotation. A uniform scaling of the signal values to the range of [0, 1] was achieved by min-max normalization per gait sequence (Fig. 2).

Subsequently, the signal was down-sampled by factor 50 to 2.048 Hz using linear interpolation resulting in the final preprocessed signal gyr_{pre}. Considering the previously applied

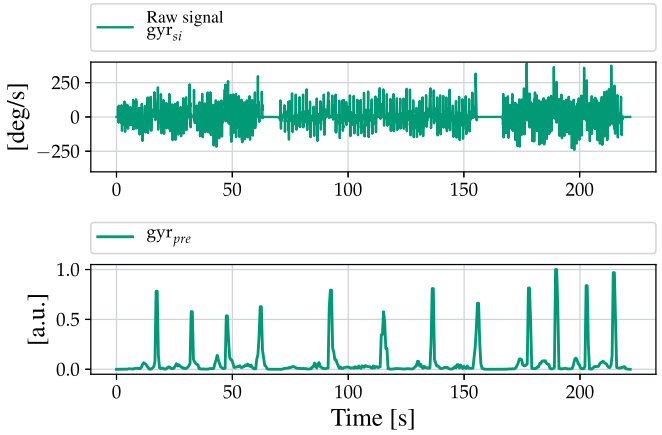

Fig. 2. Top: The raw gyr_{si} signal of a gait sequence containing a *gait test series*. Bottom: The resulting signal gyr_{pre} after the preprocessing pipeline including low-pass filtering, smoothing, squaring, normalization, and downsampling. This signal will be the input for the template matching.

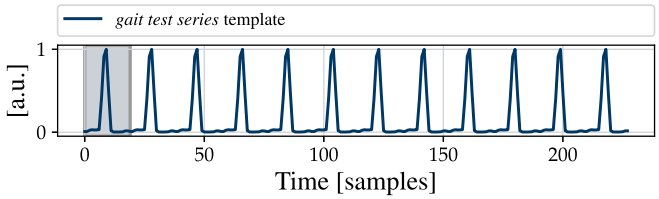

Fig. 3. The template used for the sDTW-based *gait test series* detection. The basic turning peak shaded in grey was concatenated twelve times to represent the expected number of turns in a *gait test series*.

Butterworth low-pass filter, the filtered signal contained negligible energy in frequencies above 1 Hz. Therefore, the reduced sampling rate was sufficient for storing all necessary information and helped to decrease the computational effort of the following template matching by factor 50.

2) Template Matching: After preprocessing, the characteristic turning peak pattern in gyr_{pre} was detected using template matching. As the $4 \times 10MWT$ duration varied due to the different self-selected velocities, and the path length that could deviate from 10 m, we used sDTW for this work [23].

a) Template generation: Data of few participants was sufficient to create a representative template for gait test detection [22]. Therefore, one $4 \times 10MWT$ each from three different participants was chosen randomly. The respective gyr_{si} signals were preprocessed according to the procedure explained above and averaged sample-wise after resampling to their average duration. We selected a single 180° turning peak from the resulting signal as the basic component and concatenated it twelve times subsequently for the final *gait test series* template T (Fig. 3).

b) Subsequence dynamic time warping: In summary, a distance matrix D was calculated, where each matrix entry $D(m, n)$ contained the Euclidean distance between the respective values of the template T and the preprocessed signal gyr_{pre}, as described in [24]:

$$D(m,n) = \sqrt{(T[m] - gyr_{pre}[n])^2} \tag{1}$$

The overall cost for warping T completely to gyr_{pre} in the cheapest way, was then computed using an accumulated cost matrix C based on the implementation of `subsequence_cost_matrix` by tslearn [29]. The upper

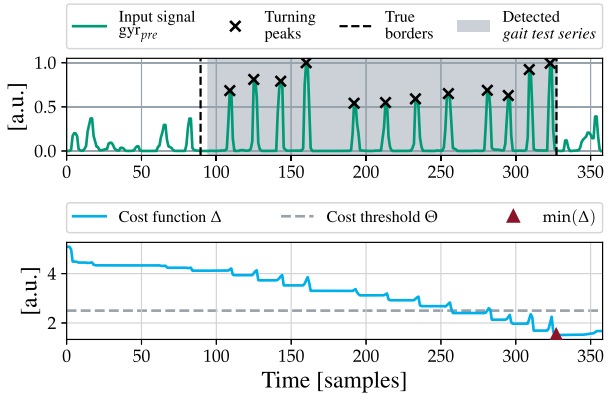

Fig. 4. Top: Example for the *gait test series* detection given the input signal gyr_{pre} with the expected number of twelve turning peaks. Bottom: Using sDTW resulted in the cost function Δ. The local minima of the cost function below Θ were determined (only one in the example), and the *gait test series* borders were calculated from the corresponding warping path.

row of C contained the accumulated costs for warping the complete template T to gyr_{pre} and can be interpreted as a discrete cost function Δ (Fig. 4). Local minima of Δ below a cost threshold Θ indicated end points of matches between T and gyr_{pre} and hence *gait test series* candidates (gyr_{cand}).

To determine the beginning of the *gait test series* candidates, the warping paths starting from the minima were reconstructed from the cost matrix C, where the beginning of the warping path corresponded to the beginning of the match [24]. In case Δ never reached values below the threshold Θ, the algorithm rejected the gait sequence from containing a *gait test series*. The sDTW approach described above was performed separately using the signals from the two sensors and the results were merged in an additive manner.

3) Postprocessing: Due to turnings directly before and after the *gait test series*, there could be several local minima in the cost function and cause overlapping matches. The postprocessing included further checks for each *gait test series* candidate. Two requirements needed to be fulfilled for the gyr_{cand} signals of both sensors to be confirmed as a *gait test series*:

1) gyr_{cand} had an adequate duration between the thresholds for the minimum (κ_{min}) and maximum (κ_{max}) allowed duration of a *gait test series*.
2) gyr_{cand} had a number of turning-related peaks in the range of the minimum and maximum allowed number of peaks of a *gait test series*, $[\eta_{min}, \eta_{max}]$ (Fig. 4).

To determine the number of turning-related signal peaks, the `find_peaks` function by SciPy [30] was applied to the gyr_{cand} signal. The values for κ_{min} and κ_{max} were calculated based on the durations of the gait tests in the data set, whereas η_{min} and η_{max} were determined data driven using a grid search (see III. Evaluation Study). For remaining overlapping matches after these postprocessing steps, the match with the lowest costs according to the cost function was selected.

D. Gait Test Series Decomposition

In the final step of the pipeline, the individual test borders were determined and the speed levels assigned to each $4 \times 10MWT$ in a *gait test series*. To calculate the gait speed

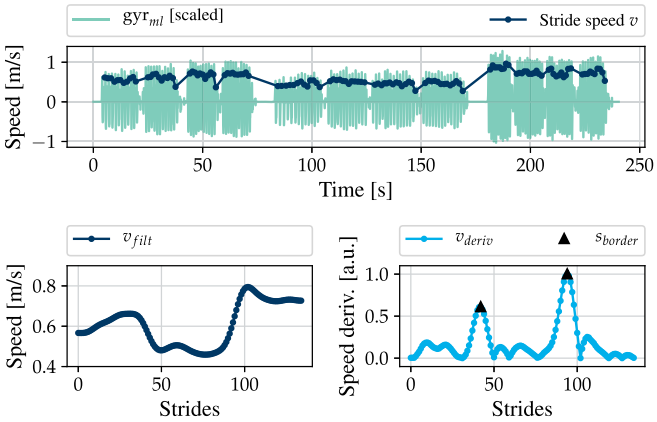

Fig. 5. Edge detection for gait test border calculation. Top: The stride speed v over one *gait test series*. Bottom left: Smoothing of v resulted in the filtered gait speed function v_{filt}. Bottom right: Derivation, absolute value calculation and normalization yielded the function v_{deriv}. Note: v_{filt} and v_{deriv} are defined as functions over strides, hence the two peaks in v_{deriv} represent the test border strides s_{border}.

per stride, first the single strides were segmented from the *gait test series* signal using the algorithm of Barth *et al.* [24]. Spatio-temporal gait parameters were then computed using the gait event detection algorithm and double integration approach of Rampp *et al.* [8]. The stride speed as the quotient of stride length and stride time given was described as a function v over the single strides s_i in the *gait test series* with a speed value $v[s_i]$ in $[m/s]$ for every stride (Fig. 5).

The following steps were applied on v to separate the three tests. To smoothen the gait speed variability within the individual $4 \times 10MWT$s, we employed a Gaussian filter kernel resulting in the filtered stride speed signal v_{filt} (Fig. 5). The optimal width w of the Gaussian bell was determined in a grid search (see III. Evaluation Study). We then computed the first derivative of v_{filt} where the extrema of the resulting v_{deriv} indicated the points of strong changes in the gait speed. These were related to the changes in the speed level between the three gait tests. As the type of the extrema was irrelevant, the absolute values of v_{deriv} were determined and the function was then min-max normalized. The two peaks with maximum height were determined using the SciPy function `find_peaks` [30]. The peaks represented the strides s_{border} that constituted the borders between the three $4 \times 10MWT$s in the *gait test series* (Fig. 5).

Finally, to assign the speed labels to the respective $4 \times 10MWT$s, the medians of the stride speed values in each of the three determined sequences were computed. The gait test speed levels were assigned to all respective strides of the segmented tests in the order *Slow*, *Preferred*, and *Fast* after sorting the median values from low to high. Hence, the borders between the individual $4 \times 10MWT$s were determined by the strides following a change of the speed level (Fig. 6).

III. Evaluation Study

A. Manual Gait Test Annotations

To provide ground truth information for the evaluation, all $4 \times 10MWT$s in the data set were manually labeled with start and end time stamp by two trained human annotators after

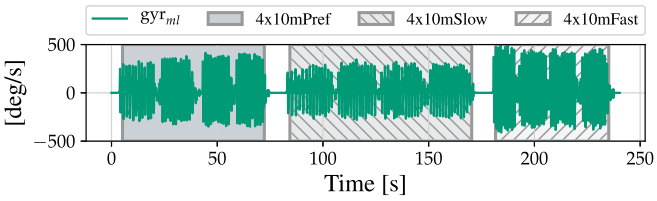

Fig. 6. The final $4 \times 10MWT$ borders and assignment of speed labels to each test.

visual inspection of the raw IMU data. The annotations set by the participants with the smartphone application served as assistance, where the sensor signals in the area of ± 1000 s around each annotation were visually inspected. The ground truth *gait test series* started with the beginning of the first $4 \times 10MWT$ and ended with the end of the last of three consecutive $4 \times 10MWT$s.

B. Gait Test Series Detection

1) Evaluation: The evaluation of the *gait test series* detection was performed participant-wise. The evaluation metrics (i.e. precision, recall, and F1-score) were determined for every participant by summing up the true positives, false positives, and true negatives over their *daily recordings*. The performance of the *gait test series* detection was investigated on two different levels of granularity. On the coarse level (area-based evaluation), we asserted whether a gait sequence either contained a *gait test series* or not. If at least one sample of a detected *gait test series* was overlapping with a ground truth *gait test series*, the detected *gait test series* was considered as a true positive. Every detected *gait test series* not fulfilling this condition was treated as false positive.

On a finer level of evaluation, we compared detected and ground truth *gait test series* based on single strides (stride-based evaluation). Two lists of strides were created by running the stride segmentation algorithm by Barth *et al.* [24] on the detected and ground truth *gait test series*, respectively. If a stride of a detected *gait test series* was also included in a ground truth *gait test series*, it was counted as true positive. All strides in detected *gait test series* that were not included in the ground truth list of strides were counted as false positives.

2) Parameter Optimization: The cost threshold Θ and the turns threshold range $[\eta_{min}, \eta_{max}]$ were tunable parameters and optimized using a grid search. For the turns threshold range, symmetric interval borders around the expected amount of twelve detected turnings were used as parameter options for the grid search (Table II). To get an unbiased estimate of the final performance, the optimization was evaluated using a 4-fold cross validation. The data was split on a participant level. Thus, in each fold, all data of nine participants were used as training set and the data of the three remaining participants as test set. Per fold, the grid search was performed on the training data and the parameter combination with the best F1-score was applied on the test set to calculate the performance metrics for this fold. The final generalization performance was computed as the average over the test sets in all folds.

The duration thresholds κ_{min} and κ_{max} for each fold were determined based on the manually labeled $4 \times 10MWT$s in the

TABLE II
VALUE RANGE OF THE PARAMETERS FOR THE GRID SEARCH IN THE *Gait Test Series* DETECTION

Parameter	Values
Θ	2.3, 2.4, 2.5, 2.6, 2.7
$[\eta_{min}, \eta_{max}]$	[0,24], [2,22], [4,20], [6,18], [8,16], [10,14], [11,13]

respective training data sets as follows: The shortest possible duration of a *gait test series*, κ_{min}, was defined as three times the shortest *Fast* $4 \times 10MWT$. The multiplication by three represented the three consecutive tests in a *gait test series*. We calculated κ_{max} analogously, using three times the duration of the longest *Slow* $4 \times 10MWT$s in the respective training set.

C. Gait Test Series Decomposition

1) Evaluation: To evaluate the *gait test series* decomposition independently of the influence of previous pipeline steps, only the IMU signals of the manually labeled *gait test series* were used as input to the algorithm. Given the pre-segmented *gait test series* data, only the decomposition into the three single $4 \times 10MWT$s was performed, including finding the test borders and assignment of the speed levels.

The *gait test series* decomposition was a three-class classification task and evaluated on the granularity of single strides. In the decomposition algorithm, each stride was assigned to one of the speed level classes, resulting in three separate $4 \times 10MWT$s. For determining the ground truth speed labels, the median speed values of the strides within the manually labelled $4 \times 10MWT$s were calculated and sorted in ascending order. The single strides of the respective $4 \times 10MWT$s were then labelled as *Slow*, *Preferred*, and *Fast*. Based on these ground truth labels and the algorithm predictions, the performance was assessed using stride-by-stride comparisons of the speed labels. The performance parameters (precision, recall, and F1-score) were again calculated participant-wise.

2) Parameter Optimization: The width w of the Gaussian filter window for smoothing the stride speed function v was optimized using a grid search. Values for $w \in \{18, 24, 30, 36, 42\}$ were tested. As described above, the optimization was evaluated using a 4-fold cross validation. The mean F1-score of all speed level classes was used as optimization criterion. In each fold, the value of w providing the highest F1-score in the training set, was applied to the test set to determine the performance. The final performance parameters were calculated as the average of the four test set performance results.

D. Concurrent Validity

Additionally to the evaluation of the two individual algorithm components, a performance assessment of the entire pipeline was done in the form of a concurrent validity analysis [31]. To test the overall validity of the *gait test series* detection and decomposition, we compared gait parameters derived from the detected and decomposed $4 \times 10MWT$s with those from the ground truth $4 \times 10MWT$s.

As the proposed algorithm could potentially produce false negatives or false positives during the *gait test series* detection, a one-to-one comparison between detected and ground truth gait tests was not feasible. Therefore, average gait parameters for detected and ground truth gait tests in *Preferred*, *Slow*, and *Fast* speed were computed and compared per *daily recording*. Gait tests with less than 30 strides were excluded to ensure a sufficient recording of gait variability [32].

TABLE III
PERFORMANCE MEASURES OF THE *Gait Test Series* DETECTION AS MEAN (SD) IN THE 4-FOLD CROSS-VALIDATION IN [%]. THE CONCEPTS OF STRIDE-BASED AND AREA-BASED EVALUATION ARE EXPLAINED IN DETAIL IN SECTION III-B

Evaluation	Precision	Recall	F1
Stride-based	87.6 (1.2)	90.6 (5.3)	88.9 (3.0)
Area-based	94.2 (2.5)	92.8 (6.5)	93.3 (4.3)

We assessed the concurrent validity of five spatio-temporal gait parameters (stride time, swing time, stance time, stride length, and gait speed) by calculating Pearson's correlation, mean error (ME), mean absolute error (MAE), and the relative absolute error (RAE).

IV. RESULTS

A. Gait Test Series Detection

In total, the data set contained 419 manually labeled *gait test series* including 87088 analyzed strides. The parameter-optimized *gait test series detection* reached an F1-score of 88.9% (±3.0%), averaged over the test data sets in the 4-fold cross-validation (Table III). In three of the four folds, $\Theta = 2.5$ was identified as the optimal cost threshold, and in three folds $[\eta_{min}, \eta_{max}] = [10, 14]$ was determined as optimal turns threshold range. The average area-based F1-score was 93.3% (±4.3%).

Higher precision was observed for lower Θ and a more narrow turns threshold range, whereas the recall was increased for a higher cost threshold and a broader range. Consequently, the highest F1-score was found in the central part of the two-dimensional parameter grid search space.

B. Gait Test Series Decomposition

The *gait test series* decomposition reached an average F1-score of 94.0% (±3.8%) over the four folds, with an average precision of 94.2% (±3.8%) and an average recall of 94.0% (±3.8%). In all four folds, the filter kernel width of $w = 30$ was found as the optimal parameter. The distribution of correct and false predictions for the different speed levels is displayed in the confusion matrix in Fig. 7.

C. Concurrent Validity

Mean and standard deviation (SD) values of the investigated gait parameters from 151 *daily recordings*, as well as agreement measures are given in Table IV. Good agreement between gait parameters derived from detected $4 \times 10MWT$ labels compared to ground truth $4 \times 10MWT$ labels was observed for all gait parameters in all speed levels.

V. DISCUSSION

In this study, we were able to automatically detect $4 \times 10MWT$s in real-world IMU gait recordings of PD patients

TABLE IV

COMPARISON OF THE SPATIO-TEMPORAL PARAMETERS OF THE DETECTED AND THE GROUND TRUTH GAIT TESTS BASED ON DAILY AVERAGES (N = 151 *Daily Recordings*). SHOWN ARE THE MEAN PARAMETERS (SD), PEARSON CORRELATION COEFFICIENT *r*, MEAN ERROR (ME), MEAN ABSOLUTE ERROR (MAE) AND THE RELATIVE ABSOLUTE ERROR (RAE)

		Detected Labels	Manual Labels	r	ME	MAE	RAE
Slow	**Stride time [s]**	1.307 (0.182)	1.315 (0.186)	0.987	0.008 (0.030)	0.018 (0.026)	0.013 (0.019)
	Swing time [s]	0.438 (0.051)	0.441 (0.052)	0.955	0.003 (0.016)	0.007 (0.014)	0.016 (0.037)
	Stance time [s]	0.869 (0.140)	0.874 (0.143)	0.987	0.005 (0.023)	0.014 (0.019)	0.016 (0.021)
	Stride length [m]	0.934 (0.193)	0.937 (0.196)	0.945	0.003 (0.064)	0.021 (0.061)	0.028 (0.117)
	Gait speed [m/s]	0.738 (0.205)	0.735 (0.210)	0.971	-0.003 (0.050)	0.021 (0.045)	0.033 (0.114)
Preferred	**Stride time [s]**	1.127 (0.119)	1.121 (0.113)	0.957	-0.006 (0.034)	0.020 (0.028)	0.018 (0.025)
	Swing time [s]	0.401 (0.038)	0.400 (0.036)	0.955	-0.001 (0.011)	0.006 (0.010)	0.015 (0.026)
	Stance time [s]	0.726 (0.087)	0.721 (0.084)	0.940	-0.005 (0.030)	0.017 (0.025)	0.023 (0.033)
	Stride length [m]	1.053 (0.190)	1.063 (0.198)	0.937	0.007 (0.068)	0.029 (0.062)	0.032 (0.112)
	Gait speed [m/s]	0.957 (0.232)	0.970 (0.242)	0.953	0.009 (0.073)	0.033 (0.066)	0.039 (0.127)
Fast	**Stride time [s]**	1.012 (0.113)	1.000 (0.116)	0.971	-0.014 (0.027)	0.020 (0.024)	0.020 (0.025)
	Swing time [s]	0.376 (0.034)	0.374 (0.034)	0.980	-0.002 (0.007)	0.004 (0.006)	0.012 (0.017)
	Stance time [s]	0.636 (0.085)	0.626 (0.091)	0.968	-0.012 (0.022)	0.017 (0.019)	0.028 (0.032)
	Stride length [m]	1.171 (0.242)	1.187 (0.260)	0.986	0.020 (0.042)	0.031 (0.034)	0.025 (0.026)
	Gait speed [m/s]	1.194 (0.315)	1.220 (0.335)	0.985	0.031 (0.057)	0.044 (0.048)	0.035 (0.034)

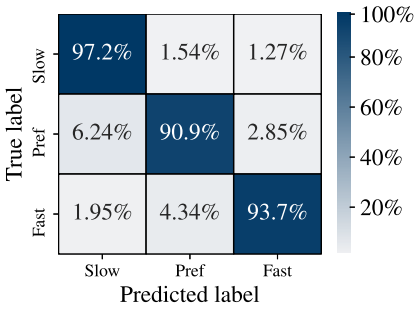

Fig. 7. Confusion matrix for the *gait test series* decomposition.

with a novel algorithmic pipeline. The detection algorithm makes use of the periodic pattern of recurring turning movements followed by straight walks. This pattern is captured by the superior-inferior gyroscope axis signal (gyr_{si}) and can be found during continuous recordings using template matching based on sDTW.

A major challenge for detecting standardized unsupervised gait tests are deviations from the predefined study protocol resulting in non-compliant gait tests. For instance, the walkway length cannot be expected to be exactly 10 m, resting periods between individual $4 \times 10MWT$s were omitted in many cases, and the order of speed levels in consecutive $4 \times 10MWT$s varied between and within participants.

These boundary conditions required a higher algorithmic complexity and advancements compared to our previously presented algorithm for the detection of $4 \times 10MWT$s from short, supervised recordings in the laboratory [22]. Hence, a substantially extended pipeline consisting of two separated parts for the detection and decomposition of *gait test series* was developed and evaluated on real-world data.

A. Gait Sequence Detection

A previously reported challenge were very short resting periods between single gait tests that lead to several tests being included in one gait sequence [22]. In our unsupervised study, the existence of resting periods between the single $4 \times 10MWT$s in one *gait test series* could not be ensured. In fact, we observed that some participants performed no or only very short resting periods. To achieve a generalizable solution for all participants, we decided to always merge successive gait sequences separated by short breaks. This concatenation ensured a robust processing, independent of the resting period duration between individual $4 \times 10MWT$s in a *gait test series*. Consequently, a *gait test series* was expected to be entirely captured in one gait sequence.

B. Gait Test Series Detection

The sDTW approach for the template matching allowed the detection of *gait test series* with $4 \times 10MWT$s of different speed levels using a uniform template based on the replication of a single turning peak (Fig. 3). Due to the ability to stretch or squeeze the template to achieve an optimal fit with the probe signal, also potential deviations from the path length of 10 m could be handled successfully.

After the sDTW, a postprocessing pipeline was applied focusing on determining *gait test series* borders as precisely as possible. Due to the unsupervised scenario, resting periods were not only often missing between subsequent $4 \times 10MWT$s, but also directly before and after a *gait test series*. Therefore, gait sequences detected in real-world data could include a *gait test series*, but in addition also other gait signals. Possible reasons are, that participants were walking to prepare an obstacle-free straight walkway and the smartphone for operating the annotation app. This is in contrast to the standardized clinical recordings in [22], where no other gait movements besides the $4 \times 10MWT$ were contained in an observed gait sequence. Hence, the gait sequence borders could directly be used as the test borders. Therefore, in this study, all *gait test series* candidates were undertaken turning and length checking, leading to precisely set borders (Fig. 4).

One reason for the errors observed in the study can be found in the static template with a preset amount of twelve turning peaks. In case of non-compliant tests that were performed with a reduced number of passes and turnings, matches could be missed during the template matching. Non-compliant tests with an increased number of turnings could be detected, but the borders only covered a part of the true *gait test series*.

The predicted *gait test series* borders were not directly compared with the ones from the ground truth. However, we consider the comparison of lists of strides from detected and ground truth *gait test series* more valuable, as it gives an estimate for the expected accuracy for an automated stride-based analysis of gait parameters (Table IV). Therefore, it was important to evaluate if the predicted *gait test series* borders contain all strides that belong to the corresponding tests.

We also computed the area-based performance, which helped to understand if the algorithm was generally able to detect the regions of interest independently of the accurate determination of the test borders (Table III). A high area-based accuracy allows to avoid timestamp annotations performed by the patients. Our algorithm can predict the coarse regions of the gait tests within the IMU data and thus assist a researcher during the manual test annotation when a fully automated processing is not desired. An overlap of one sample is sufficient, as the larger signal region in the range of minutes can then be displayed to the human annotator, who may verify or correct the test borders manually. Given the area-based F1-score of 93.3%, it can be concluded that the presented algorithm provides a reliable assistance for an annotation tool, where the performance could also be further tuned towards an improved recall, by adapting the thresholds accordingly.

C. Gait Test Series Decomposition

In the second part of the proposed pipeline, the beginning and end of individual $4 \times 10MWT$s were determined and corresponding speed levels were assigned (Fig. 6). The algorithm detected macroscopic changes of the gait speed within a *gait test series*, to determine the transitions between the $4 \times 10MWT$s (Fig. 5). Within every *gait test series*, the three $4 \times 10MWT$s were sorted according to the respective median gait speed. Thus, pre-determined speed thresholds were avoided, considering that the self-selected gait speeds can considerably differ between patients in different disease stages [19].

The *gait test series* decomposition had a performance with an F1-score of 94.0% ($\pm$3.8%). The low standard deviation in the cross-validation, as well as the consistent value for the optimized hyper parameter w indicate a good generalizability of the results. Errors could occur, when the differences between the three speed levels were not very pronounced. This could either cause the borders to be shifted by some strides or even the switch of an entire speed level between tests.

D. Concurrent Validity

In addition to the performance evaluation of the *gait test series* detection and decomposition against the ground truth annotations, we investigated the spatio-temporal gait parameters derived from the gait tests. The calculated parameters are in agreement with values previously measured for patients with PD [33]. The results of the concurrent validity study further show a very good agreement between the gait parameters derived from gait tests detected by the proposed algorithm and the parameters derived from the ground truth gait tests (Table IV). The error ranges are below differences that are clinically relevant, for example for the distinction of patients with different PD disease stages [19]. The good agreement of the gait parameters is a direct consequence of the similarity of the lists of strides contained in the detected and the ground truth gait tests. Hence, these findings underline the good performance of the proposed algorithmic pipeline.

E. Limitations

The separate evaluation of the *gait test series* decomposition was only feasible using data from the ground truth *gait test series* where always three subsequent $4 \times 10MWT$s were performed. The performance for potentially detected *gait test series* with too many or too few $4 \times 10MWT$s was not explicitly investigated. However, *gait test series* with an unusual number of passes, like $2 \times$ 10MWTs or $6 \times$ 10MWTs, were present in the data set and did not decrease the accuracy, as the gait speed changes were still present. Nevertheless, the results of the concurrent validity study are promising and indicate that the entire pipeline consisting of *gait test series* detection and *gait test series* decomposition is robust for real-world gait data.

VI. Conclusion and Outlook

We developed and evaluated a novel method for the automatic detection of unsupervised standardized $4 \times 10MWT$s in real-world IMU recordings of PD patients. First, gait test series of subsequently performed $4 \times 10MWT$s of different gait speeds, were detected and extracted. Second, these series were decomposed into single gait tests and the corresponding gait speed levels were assigned. Finally, we performed a concurrent validity study, proving that our algorithm is able to provide digital mobility outcomes that are similar to those calculated from manually labeled tests.

The proposed method works well for series of three subsequent $4 \times 10MWT$s, as they contain a characteristic pattern of turnings, that is different from typical real-world movements. The reliable detection of single $4 \times 10MWT$s or shorter tests such as the TUG is potentially more challenging and might require advanced techniques to detect them while preventing a high false positive rate. Hence, in future research, the detection of other gait tests should be further investigated considering more complex machine learning methods.

In this study, we used a sensor setup with two sensors attached to the shoes. However, a single sensor at the lower back is often used in real-world monitoring. The presented algorithm focusing on turnings is potentially also applicable for a single sensor setting including a gyroscope, which should be investigated in future research. Initial experiments beyond the scope of this study already showed promising results.

From a clinical point of view, further research is required to interpret and understand the outcomes of unsupervised gait tests with respect to standardized tests in the hospital on one hand and to fully unsupervised real-world gait on the other hand. Our algorithm provides a reliable tool to support clinicians and patients with the recording and processing of the required data, which is potentially also applicable to other diseases with motor impairments. To the best of our knowledge, this is the first study presenting an algorithm for the detection

of unsupervised standardized gait tests within real-world data. Therefore, we make and essential contribution to the field of real-world gait analysis, enabling further clinical research for a better understanding and treatment of movement disorders.

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.

REFERENCES

[1] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol., Neurosur. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] M. Skorvanek *et al.*, "Differences in MDS-UPDRS scores based on Hoehn and Yahr stage and disease duration," *Movement Disorders Clin. Pract.*, vol. 4, no. 4, pp. 536–544, Jul. 2017.

[3] D. A. Heldman *et al.*, "The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures," *Movement Disorders*, vol. 26, no. 10, pp. 1859–1863, Aug. 2011.

[4] B. Post, M. P. Merkus, R. M. de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?" *Movement Disorders*, vol. 20, no. 12, pp. 1577–1584, 2005.

[5] J. Klucken *et al.*, "Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e56956.

[6] S. D. Din, C. Kirk, A. J. Yarnall, L. Rochester, and J. M. Hausdorff, "Body-worn sensors for remote monitoring of Parkinson's disease motor symptoms: Vision, state of the art, and challenges ahead," *J. Parkinson's Disease*, vol. 11, no. s1, pp. 1–13, 2021.

[7] D. Trojaniello *et al.*, "Estimation of step-by-step spatio-temporal parameters of normal and impaired gait using shank-mounted magneto-inertial sensors: Application to elderly, hemiparetic, parkinsonian and choreic gait," *J. Neuroeng. Rehabil.*, vol. 11, no. 1, pp. 1–12, 2014.

[8] 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.

[9] L. Rochester *et al.*, "A roadmap to inform development, validation and approval of digital mobility outcomes: The mobilise-D approach," *Digit. Biomarkers*, vol. 4, no. 1, pp. 13–27, Jan. 2021.

[10] E. Warmerdam *et al.*, "Long-term unsupervised mobility assessment in movement disorders," *Lancet Neurol.*, vol. 19, no. 5, pp. 462–470, May 2020.

[11] T. Iluz *et al.*, "Automated detection of missteps during community ambulation in patients with Parkinson's disease: A new approach for quantifying fall risk in the community setting," *J. NeuroEng. Rehabil.*, vol. 11, no. 1, pp. 1–9, Dec. 2014.

[12] 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," *PLoS ONE*, vol. 9, no. 5, May 2014, Art. no. e96675.

[13] 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. A, Biomed. Sci. Med. Sci.*, vol. 74, no. 4, pp. 500–506, Apr. 2019.

[14] A. J. Espay *et al.*, "Technology in Parkinson's disease: Challenges and opportunities," *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, Apr. 2016.

[15] S. D. Din, A. Godfrey, C. Mazzà, S. Lord, and L. Rochester, "Free-living monitoring of Parkinson's disease: Lessons from the field," *Movement Disorders*, vol. 31, no. 9, pp. 1293–1313, Sep. 2016.

[16] 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 Disease*, vol. 10, no. 4, pp. 1763–1773, 2020, doi: 10.3233/JPD-202129.

[17] A. Salarian, F. B. Horak, C. Zampieri, P. Carlson-Kuhta, J. G. Nutt, and K. Aminian, "iTUG, a sensitive and reliable measure of mobility," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 3, pp. 303–310, Jun. 2010.

[18] C. Zampieri, A. Salarian, P. Carlson-Kuhta, J. G. Nutt, and F. B. Horak, "Assessing mobility at home in people with early Parkinson's disease using an instrumented timed up and go test," *Parkinsonism Rel. Disorders*, vol. 17, no. 4, pp. 277–280, May 2011.

[19] J. C. M. Schlachetzki *et al.*, "Wearable sensors objectively measure gait parameters in Parkinson's disease," *PLoS ONE*, vol. 12, no. 10, 2017, Art. no. e0183989.

[20] M. Heijmans *et al.*, "Monitoring Parkinson's disease symptoms during daily life: A feasibility study," *NPJ Parkinson's Disease*, vol. 5, no. 1, pp. 1–6, Dec. 2019.

[21] M. Naftali and L. Findlater, "Accessibility in context: Understanding the truly mobile experience of smartphone users with motor impairments," in *Proc. 16th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, 2014, pp. 209–216.

[22] S. Fischer *et al.*, "Automatic clinical gait test detection from inertial sensor data," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jul. 2020, pp. 789–792.

[23] M. Müller, *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, ch. 4, pp. 69–84.

[24] 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.

[25] I. Ghersi, M. H. Ferrando, C. G. Fliger, C. F. C. Arenas, D. J. E. Molina, and M. T. Miralles, "Gait-cycle segmentation method based on lower-trunk acceleration signals and dynamic time warping," *Med. Eng. Phys.*, vol. 82, pp. 70–77, Aug. 2020.

[26] N. Roth, C. F. Martindale, B. M. Eskofier, H. Gaßner, Z. Kohl, and J. Klucken, "Synchronized sensor insoles for clinical gait analysis in home-monitoring applications," *Current Directions Biomed. Eng.*, vol. 4, no. 1, pp. 433–437, Sep. 2018.

[27] F. Ferraris and M. Parvis, "Procedure for effortless in-field calibration of three-axis rate gyros and accelerometers," *Sensors Mater.*, vol. 7, no. 5, pp. 311–330, 1995.

[28] 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.

[29] R. Tavenard *et al.*, "Tslearn, a machine learning toolkit for time series data," *J. Mach. Learn. Res.*, vol. 21, no. 118, pp. 1–6, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-091.html

[30] P. Virtanen *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Feb. 2020.

[31] F. Kluge, H. Gaßner, J. Hannink, C. Pasluosta, J. Klucken, and B. M. Eskofier, "Towards mobile gait analysis: Concurrent validity and test-retest reliability of an inertial measurement system for the assessment of spatio-temporal gait parameters," *Sensors*, vol. 17, no. 7, p. 1522, Jun. 2017.

[32] B. Galna, S. Lord, and L. Rochester, "Is gait variability reliable in older adults and Parkinson's disease? Towards an optimal testing protocol," *Gait Posture*, vol. 37, no. 4, pp. 580–585, Apr. 2013.

[33] O. Sofuwa, A. Nieuwboer, K. Desloovere, A.-M. Willems, F. Chavret, and I. Jonkers, "Quantitative gait analysis in Parkinson's disease: Comparison with a healthy control group," *Arch. Phys. Med. Rehabil.*, vol. 86, no. 5, pp. 1007–1013, May 2005.