Motion Sensor-Based Assessment of Parkinson's Disease Motor Symptoms During Leg Agility Tests: Results From Levodopa Challenge

Somayeh Aghanavesi, Filip Bergquist, Dag Nyholm, Marina Senek, and Mevludin Memedi

***Abstract*—Parkinson's disease (PD) is a degenerative, progressive disorder of the central nervous system that mainly affects motor control. The aim of this study was to develop data-driven methods and test their clinimetric properties to detect and quantify PD motor states using motion sensor data from leg agility tests. Nineteen PD patients were recruited in a levodopa single dose challenge study. PD patients performed leg agility tasks while wearing motion sensors on their lower extremities. Clinical evaluation of video recordings was performed by three movement disorder specialists who used four items from the motor section of the unified PD rating scale (UPDRS), the treatment response scale (TRS) and a dyskinesia score. Using the sensor data, spatiotemporal features were calculated and relevant features were selected by feature selection. Machine learning methods like support vector machines (SVM), decision trees, and linear regression, using ten-fold cross validation were trained to predict motor states of the patients. SVM showed the best convergence validity with correlation coefficients of 0.81 to TRS, 0.83 to UPDRS #31 (body bradykinesia and hypokinesia), 0.78 to SUMUPDRS (the sum of the UPDRS items: #26-leg agility, #27-arising from chair, and #29-gait), and 0.67 to dyskinesia. Additionally, the SVM-based scores had similar test–retest reliability in relation to clinical ratings. The SVM-based scores were less responsive to treatment effects than the clinical scores, particularly with regards to dyskinesia. In conclusion, the results from this study indicate that using motion sensors during leg agility tests may lead to valid and reliable objective measures of PD motor symptoms.**

***Index Terms*—Leg agility, Parkinson's disease, support vector machine, stepwise regression, predictive models.**

I. Introduction

PARKINSON'S disease (PD) is a chronic degenerative disorder of the central nervous system. It is characterized by motor symptoms (e.g., bradykinesia, rigidity, and tremor) and non-motor symptoms (sleep problems, impaired cognition, etc.). These symptoms affect the health-related quality of life of individuals diagnosed with PD [1], [2]. The disease is progressive and all currently available therapies are symptomatic without affecting the progression of the disease.

In the early stages the treatment yields a good response with few side effects. However, as the disease progresses therapy complications emerge. The motor states of the patients can fluctuate between "Off", "On" and "On with dyskinesia" motor states. During "Off" periods the medication effect is not optimal, and patients experience parkinsonian symptoms. During "On" periods the medication works better. During "On with dyskinesia" periods patients experience involuntary movements that can be handicapping when excessive. While "Off"-periods are invariably caused by under-treatment, e.g., due to troughs in blood levodopa concentrations, dyskinesia can occur both in response to excessive concentrations of medication in the blood [3], [4] and to decreasing concentrations, when they are suddenly followed by severe "Off". PD treatment is typically guided by patient history based on patient recall, sometimes recorded in paper home diaries [5], and clinical examination using rating scales [6]. The most commonly used rating scale is the Unified PD Rating Scale (UPDRS). However, this scale is associated with a number of limitations including the need for trained experts to use it, the need for the patient to visit the clinic and the existing inter- and intra-observer variability when using it [7]. Because of the infrequent clinical visits these assessments provide only a limited picture of the patients' health status, thus limiting the healthcare providers to offer individualized management of the symptoms and treatment.

To mitigate these problems, coupling sensor technology with data-driven methods such as time series analysis and machine learning may provide means to objectively quantify the states

Manuscript received August 6, 2018; revised September 28, 2018, December 18, 2018, and February 3, 2019; accepted February 4, 2019. Date of publication February 8, 2019; date of current version January 6, 2020. This work was supported in part by the Swedish Knowledge Foundation (Sweden), in part by Swedish Innovation Agency, in part by Acreo, in part by Cenvigo, in part by Sensidose, in part by Uppsala University, in part by Örebro University, and in part by Dalarna University. *(Corresponding author: Somayeh Aghanavesi.)*

S. Aghanavesi is with the Department of Computer Engineering, Dalarna University, Borlänge 781 70, Sweden (e-mail: saa@du.se).

F. Bergquist is with the Department of Pharmacology at Institute of Neuroscience and Physiology, Gothenburg University, Gothenburg 405 30, Sweden (e-mail: filip.bergquist@gu.se).

D. Nyholm and M. Senek are with the Department of Neuroscience, Neurology, Uppsala University, Uppsala 751 85, Sweden (e-mail: dag.nyholm@neuro.uu.se; marina.senek@neuro.uu.se).

M. Memedi is with the Informatics, Örebro University, Örebro 702 81, Sweden (e-mail: mevludin.memedi@oru.se).

Digital Object Identifier 10.1109/JBHI.2019.2898332

TABLE I
CHARACTERISTICS OF PD PATIENTS SHOWN AS MEAN (STANDARD DEVIATION)

Gender	Age (years)	Height (m)	Weight (kg)	Years with PD	Years on levodopa	Affected PD side	Hoehn & Yahr	UPDRS IV
14 Males, 5 Females	71.4 (6.3)	1.75 (0.09)	75.4 (11)	9.7 (6.8)	9.5 (6.5)	9 right, 8 left, 2 not known	3.1(0.8)	6.21(3.13)

in PD. This technological framework could offer an objective and reliable assessment of motor states, which may be helpful for better understanding the health condition of the PD patients and individualization of their treatments [8]. Sensor technology, comprised of connected and sensing components embedded into wearable accessories or smartphones, has shown promise for monitoring patients in clinical settings and other home environments [1], [2]. In a recent review study by Johansson *et al*. [9] it was suggested that sensor technology provides useful information of clinical features for monitoring neurological disorders such as PD, stroke and epilepsy. In another study performed by Ramsperger *et al*. [10] it was shown that wearable sensors are promising for quantifying PD motor symptoms in different environmental settings. Previous research has focused on quantifying cardinal PD motor symptoms such as bradykinesia, rigidity, and tremor with the help of sensor technology [9]. Some studies aimed to develop sensor technology that allows quantification of motor states including Off, On and On with dyskinesia [11]–[13]. Promising results for predicting PD symptom severity with regard to UPDRS [14] and quantifying whole-body bradykinesia [15] were reported using multimodal sensors. A review study performed by Rovini *et al*. [8] reported that there are a small number of studies that used sensor data from lower extremities. The most investigated motor tasks for analyzing motor impairments from lower extremities include freezing of gait [16], and timed up and go tests [17]. From a clinical point of view the leg agility test can be seen as the least useful test for measuring PD motor symptoms. However, previous research has shown that objective measures of leg agility were good predictors of the disease severity with regard to UPDRS [13], [18], [19] indicating that they provide more information about the patient health status than what can be captured during clinical observations by movement disorder experts. In severely disabled patients, e.g., at Hoehn and Yahr stage 4 and 5, where there is a higher risk of falling, leg agility test could be a better alternative than other examination tests.

Therefore, the main research question of our study is to test clinimetric properties of data-driven methods to detect and quantify PD motor states using motion sensor data collected during leg agility tests. The aim is to develop and evaluate methods for quantitative assessment of motor states over the time course of a single levodopa dose. To achieve this, in this study motion sensors data were evaluated with regard to clinical ratings on Treatment Response Scale (TRS) [20], parts of the UPDRS III, and dyskinesia score [21]. After processing the sensor data with time series analysis, machine learning methods were employed to map a set of selected spatiotemporal features to the clinical ratings. This was followed by evaluation of their convergent validity, test-retest reliability and responsiveness to treatment.

II. MATERIALS AND METHODS

A. Data Collection

1) Experimental Setup and Participants: Nineteen fluctuating PD patients experiencing motor fluctuations were recruited to a single center, open label, single dose observational study in a hospital in Uppsala, Sweden [22]. Eighteen patients experienced wearing off fluctuations, 13 of them experienced dyskinesia [22]. The study was approved by the regional ethical review board in Uppsala, Sweden. Patient characteristics are shown in Table I.

The administered dose was 150% of their individual levodopa-carbidopa equivalent morning dose in order to induce dyskinesia. Standardized motor tests according to UPDRS-III such as rapid alternating movements of hands, reading a text, finger tapping, leg agility, and walking were performed at different time intervals, including once within 50 minutes before taking the dose, once at the time of dose administration (0 min), and then approximately at 20, 40, 60, 80, 110, 140, 170, 200, 230, 260, 290, 320 and 350 min after dose administration. This was to follow the individual response of the patients to their morning doses from Off motor state to good mobility and/or dyskinetic state and regressing back to Off state. Each patient performed the tests as long as they could, up to 15 trials.

2) Sensor Measurements: For acquisition of sensor data during the motor tests patients were asked to wear motion sensors on their ankles. Each sensor consisted of 3-axial accelerometers and gyroscopes (sampling rate of 102.4 Hz, accelerometer range of +/−16 g and gyroscope range of +/−2000dps). To perform the leg agility test patients were instructed to sit on a straight-back chair and place both feet comfortably on the floor and then to raise and stomp each foot on the floor 10 times as fast as possible. They performed the test first with the right foot and then with the left foot. Fig. 1 shows the test performance and the sensor placement. The sensor data of all time points (x, y, and z axes of accelerometers and gyroscopes) were saved on the SD cards of the sensors and processed offline. Each test occasion was video recorded and timestamps of the sensor data were synchronized with the time points of the videos that were used for the clinical ratings, as explained in the following section.

3) Clinical Assessment of Motor Functions: The video sequences were presented in a randomized order for the movement disorder specialists to blindly rate the performance of the patients with respect to the time from dose administration [20]. The specialists rated four items of the UPDRS-III section including UPDRS #26 (leg agility), UPDRS #27 (arising from chair), UPDRS #29 (gait), UPDRS #31 (body bradykinesia and hypokinesia), each of which were rated on a scale from 0 (normal) to 4 (severely impaired) [23]. The severity of dyskinesia

Fig. 1. Patient wearing a motion sensor on each ankle during the leg agility test. The test was performed first with the right foot and then the left foot.

was also rated on a scale from 0 (no dyskinesia) to 4 (severe dyskinesia) [21]. To rate the overall mobility of patients, the raters used TRS [20], which ranged from −3 (very "Off", severe Parkinsonism) to 0 (On, normal mobility) to 3 (Severe dyskinesia, severe choreatic dyskinesia).

B. Data Processing and Analysis

The procedure was done by first manually segmenting the sensor data, followed by calculating spatiotemporal features from the signals, selecting the set of the most relevant features and application of the machine learning methods.

1) Preprocessing: The sensors continuously recorded data during the test day that was from morning until the last test occasion (350 minutes after dose administration). The total time of recordings for each sensor was about six hours. The data included motion sensor recordings of all the motor tasks. To identify data segments of interest during the leg agility tests the timestamps of each test were extracted through synchronization with the collected videos. Sensor recordings on the Y-axis of acceleration were used to identify the start and the end time points of the leg agility test. The extracted segments consisted of 3D accelerometer and gyroscope data starting from 2 seconds before the first leg agility test movement was identified until 2 seconds after the last movement in the test. This work was performed manually and custom software for plotting and saving the extracted data was built in Matlab.

2) Feature Extraction: This work is based on a data-driven approach where time series analysis techniques were used to extract meaningful motor state information in form of spatiotemporal features from motion sensors data during leg agility tests. These features were then used during machine learning, as described in the following sections. The segmented sensor data were processed to calculate 24 features for each foot separately (Table II). Both acceleration and orientation signals were used as primary inputs to time series analysis techniques. The feature set included statistical features such as mean, standard deviation, and skewness, information-theoretic features such as Approximate Entropy (ApEn), and time-frequency domain features such as Discrete Wavelet Transform (DWT).

TABLE II
EXTRACTED FEATURES FROM RAW ACCELEROMETER AND GYROSCOPE DATA DURING LEG AGILITY TASKS

Feature #	Description
1 2	The mean magnitude of acceleration and orientation.
3 4	The standard deviation magnitude of acceleration and orientation.
5 6	The skewness of magnitude of acceleration and orientation.
7 8	Maximum magnitude of acceleration and orientation.
9 10	Mean of the third level high-frequency components after applying DWT on magnitude of acceleration and orientation.
11 12	Standard deviation of the third level high-frequency components after applying DWT on magnitude of acceleration and orientation.
13 14	ApEn of magnitude of acceleration and orientation.
15	Number of peaks in magnitude acceleration.
16	Standard deviation of peaks of magnitude of acceleration.
17	Slope of the regression line calculated for peaks magnitude of acceleration over time.
18 19	Mean and standard deviation of the time segments when foot was on the floor.
20 21	Energy of magnitude of acceleration and orientation.
22 23	Mean and maximum of displacement.
24	The total area under the curve of magnitude of acceleration.

The six original signals from motions sensors included *Xacc, Yacc, Zacc, Xgyr, Ygyr* and *Zgyr*, which were used to calculate magnitude of acceleration and orientation (*Macc* and *Mgyr*) signals, using the following equations:

$$Macc = \sqrt{X_{acc^2} + Y_{acc^2} + Z_{acc^2}} \quad (1)$$

$$Mgyr = \sqrt{X_{gyr^2} + Y_{gyr^2} + Z_{gyr^2}} \quad (2)$$

Where X_{acc}, Y_{acc} and Z_{acc} represent the acceleration and X_{gyr}, Y_{gyr} and Z_{gyr} represent the orientation in the three axes, respectively.

The first eight features that were calculated were the mean, standard deviation, maximum, and skewness for the two magnitude of acceleration and orientation.

Features #9, #10, #11, and #12 were based on a three level DWT with Daubechies wavelet function family [24], [25] applied on *Macc* and *Mgyr* signals. First level consisted of low frequency components of 0-25.6 Hz and high frequency components of 25.6-51.2 Hz. After decomposing the low frequency components of the first level, the second level consisted of low frequency components of 0-12.8 Hz and high frequency components of 12.8-25.6 Hz. Finally, the third level signals were further decomposed to the low (0-6.4 Hz) and high (6.4-12.8 Hz) frequencies. Next, mean and standard deviation of *Macc* and *Mgyr* were calculated for third level high-frequency components to generate the four features.

ApEn is a nonlinear and statistical method [26] and was applied on the *Macc* and *Mgyr* signals to quantify the amount of irregularities [27] in motion during leg agility tests. The window size was set to 2 and the similarity measure was set to 0.2 (20% of the standard deviation of the signal) [26]. This analysis resulted in features #13 and #14.

To quantify the number of foot taps during a test trial, peaks of *Macc* were counted and this resulted in feature #15. Feature #16 was defined as the standard deviation of *Macc* peaks to measure the variability in acceleration. Going further, the slope of the regression line was calculated for peaks in magnitude of acceleration to measure the overall trend of acceleration during the test trial, resulting in feature #17.

Time segments when the foot was on the floor were identified and their mean and standard deviation were calculated to represent the amount of delay the patients exhibited when trying to raise the foot (features #18, #19).

The next two features (#20 and #21) were related to the energy of the *Macc* and *Mgyr* signals and were calculated using the following equations:

$$\mathrm{E_{acc}} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} |\mathrm{M_{acc}}|^2 \tag{3}$$

$$\mathrm{E}_{gyr} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} |\mathrm{M_{gyr}}|^2 \tag{4}$$

Where N was the number of data points in the signal and $\mathrm{E_{acc}}$ and $\mathrm{E_{gyr}}$ were the energy of *Macc* and *Mgyr* signals, respectively. In order to obtain a comparable energy score for all patients and since the signal energy increased with the length of the signal, $\mathrm{E_{acc}}$ and $\mathrm{E_{gyr}}$ were divided to their respective signal lengths.

To obtain the amount of displacement of foot during the tests the following equation was applied on the magnitude of acceleration signal:

$$\mathrm{S} = v_0\,(t_{i+1} - t_i) + \frac{M_{acc_i}(\mathrm{t_{i+1}} - \mathrm{t_i})^2}{2} \tag{5}$$

Where *S* represents the amount of displacement in meters, *t* was the timestamp in seconds and *Macc* was the magnitude of acceleration at timestamp *i* in m/s^2. The velocity at t_i was assumed to be zero and displacement was measured by calculating the area under the line of *Macc* shaping a trapezoid between every two sequential time points of t_{i+1}, t_i. To calculate features #22 and #23 mean and maximum of displacement were calculated.

The final feature (feature #24) was based on the total area under the *Macc* signal curve during the test trial.

Finally, for each patient and test occasion, the mean values of the features for each foot were calculated and used in the subsequent analysis. The feature extraction analysis was performed using a custom software written for Matlab.

3) Feature Selection: To find a subset with the most relevant features in relation to the clinical ratings, forward-backward stepwise regression and principal component analysis (PCA) were applied. During stepwise regression, the 24 features were used as independent variables and mean ratings of the three raters on each of the clinical scales were used as dependent variables. In total, four stepwise regression models were built and each one of them resulted in a different list of features to be considered as predictors in the machine learning process, as explained below. Stepwise regression models start with no features and then iteratively include or remove features based on significance of the predictors providing the best fit. In the PCA, the dimensions of the 24 features were reduced into a smaller set of variables called principal components (PC). The most relevant PCs were retained in subsequent analysis based on the amount of variation explained by each PC with eigenvalues equal to or higher than 1 [28].

TABLE III
RELEVANT SPATIOTEMPORAL FEATURES WERE EXTRACTED DURING FEATURE SELECTION PROCEDURES

Feature #	Response in stepwise regression			
	TRS	UPDRS #31	SUMUPDRS	Dyskinesia
1[a]	X			
2[b]	X	X	X	X
3[a]		X	X	
4[b]		X	X	
5[a]		X	X	
6[b]	X	X	X	X
7[a]			X	X
8[b]	X	X	X	X
11[a]	X	X	X	
12[b]	X			
13[a]				X
14[b]	X			X
18[c]				X
20[a]	X		X	
21[b]		X	X	
22[a]	X	X	X	X
23[a]	X	X	X	
24[a]		X	X	

[a]accelerometer-Related Feature, [b]gyroscope-Related Feature, [c]time-Related Feature.

4) Machine Learning: Supervised machine learning methods including support vector machines (SVM), decision trees (DT) and linear regression (LR) with 10-fold cross validation were employed to map the selected features to the mean clinical ratings on the four scales of TRS, UPDRS #31 (bradykinesia), SUMUPDRS (defined as the sum of UPDRS #26 (leg agility), UPDRS #27 (arising from chair), and UPDRS #29 (gait)), and dyskinesia. For each of the scales there were individual models fitted and evaluated for their predictive performance. The SVMs were trained using radial basis kernel function and 169 support vectors. The DTs were trained using information gain criterion for selecting the most relevant features to the responses. The validation with 10-fold was done to evaluate the prediction performance of the methods. With this approach, the data was randomly divided into ten sets where during each step nine

TABLE IV
ABSOLUTE CORRELATION COEFFICIENTS (RMSE) BETWEEN MEAN RATING OF THREE RATERS AND MACHINE LEARNING-BASED SCORES USING STEPWISE REGRESSION AND PCA

	Stepwise			PCA		
	SVM	**LR**	**DT**	**SVM**	**LR**	**DT**
TRS	0.81 (0.77)	0.74 (0.86)	0.64 (1.02)	0.60 (1.05)	0.55 (1.07)	0.57 (1.01)
UPDRS #31	0.83 (0.53)	0.77 (0.59)	0.73 (0.65)	0.61 (0.74)	0.54 (0.76)	0.56 (0.77)
SUMUPDRS	0.78(1.65)	0.76(1.59)	0.74(1.74)	0.66(1.90)	0.61(1.93)	0.63(1.99)
Dyskinesia	0.67 (0.50)	0.56 (0.58)	0.35 (0.7)	0.41 (0.65)	0.38 (0.64)	0.29 (0.71)

All the coefficients had a P-value < 0.001. TRS is overall mobility, UPDRS #31 is body bradykinesia, SUMUPDRS is the sum of the UPDRS #26 (leg agility), UPDRS #27 (arising from chair) and UPDRS #29 (gait) and Dyskinesia is the severity of dyskinesia (involuntary movements).

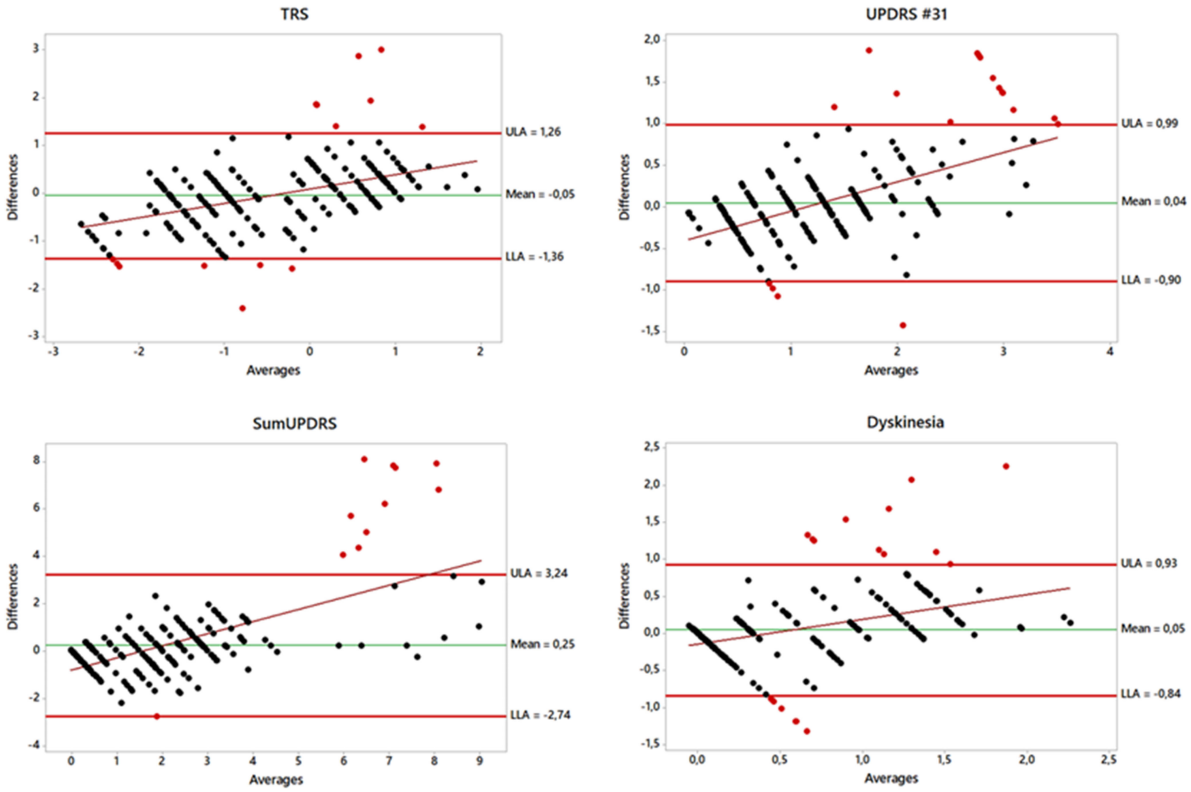

Fig. 2. Differences in scores measured by the SVMs and mean ratings of the three raters on four scales. Middle, green line represents mean bias; upper, red line shows the upper limit of agreement (ULA) calculated as mean +1.96 standard deviation; lower, red line shows the lower limit of agreement (LLA) calculated as mean – 1.96 standard deviation. The slope represents the regressed line between SVMs and mean ratings.

sets were used for training the models and one set was used for testing them. Then the correlation and the error between total predicted set and the actual set were calculated. For training and testing the machine learning methods a custom software in R was written and the default parameters of the methods in R libraries were used.

5) Statistical Analysis: Convergence validity of the machine learning methods was assessed through Pearson correlation coefficients and Root Mean Squared Error (RMSE). Agreements between the scores obtained by the machine learning methods and scores obtained by the three raters were analyzed using Bland-Altman plots [29]. The analysis included plotting the error against the mean of the machine learning- and clinician-based scores.

To assess test-retest reliability of the machine learning-based scores and inter-rater agreements one-way consistency intra-class correlation coefficients (ICCs) and their 95% confidence intervals (CI) were calculated. For the test-retest reliability analysis the data during the two baseline measurements, i.e., before receiving the dose and when the dose was administered, were used.

TABLE V
ICCS (95% CONFIDENCE INTERVAL) OF THE MEAN CLINICAL RATINGS AND SVM-BASED SCORES DURING THE TWO BASELINE MEASUREMENTS

	Clinical scores	**SVM**
TRS	0.91 (0.78-0.96)	0.81 (0.57-0.92)
UPDRS #31	0.85 (0.65-0.94)	0.89 (0.73-0.95)
SUMUPDRS	0.91 (0.78-0.97)	0.91 (0.78-0.96)

Responsiveness of the machine learning-based scores to treatment effects was assessed by calculating effect sizes. Effect sizes

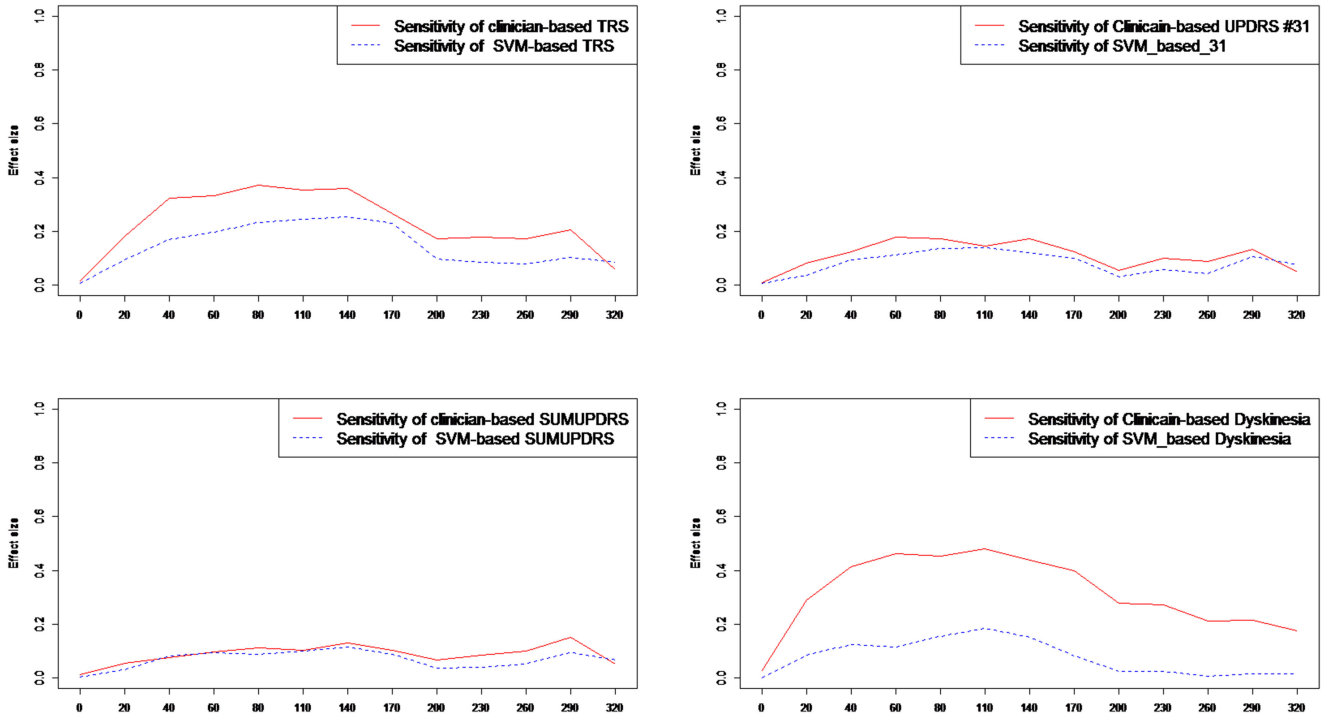

Fig. 3. Responsiveness to treatment analysis of the mean clinical ratings (solid lines) and SVM-based scores (dashed lines) for each of the four scales across the test occasions. The horizontal axis represents the minutes after taking the levodopa dose. The vertical axis represents the effect sizes representing the changes in scores between baseline and later tests e.g., changes between baseline test and first test, baseline test and second test and so on. Number of tests per time slot: 0 (n = 19), 20 (n = 19), 40 (n = 19), 60 (n = 19), 80 (n = 19), 110 (n = 19), 140 (n = 19), 170 (n = 19), 200 (n = 19), 230 (n = 18), 260 (n = 15), 290 (n = 13), 320 (n = 11). Since for test #15 at 350 min there was only observation performed by a patient it was decided to not include that observation in the calculation of effect sizes.

were used to detect changes from baseline (no medication) to the follow up time points when patients were on medication where a high effect size indicated that the methods were responsive to treatment [30].

To achieve this, analysis of variance (ANOVA) models were fitted on the data between different time points starting from test #1 and test #2, test #1 and test #3, and so on. The statistical analyses were performed in R and Minitab 18.1.

To assess the differences in motor test results between right and left legs in patients who were asymmetric, t-tests were performed on the first principal component (PC1) of the 24 features of each individual leg.

III. Results

A. Evaluation of Relevance of Features

After applying stepwise regression models for each of the four scales, different sets of features were identified as the most relevant ones to be used in machine learning. The results from this analysis are summarized in Table III. Most of the features were selected as relevant predictors for all four scales. Six features (#): 9, 10, 15, 16, 17, and 19 were not selected at all by the regression models. There were four features which were selected in the four regression models including:

- Feature #2 (Mean magnitude of orientation)
- Feature #6 (Skewness of magnitude of orientation)
- Feature #8 (Maximum magnitude of orientation)
- Feature #22 (Mean of displacement)

where features #2, #6, and #8 were based on gyroscope signals and #22 was based on accelerometer signals. When investigating the distribution of the features in the individual models it was noticed that both gyroscope- and accelerometer-related features were equally important as predictors of the 4 clinical scales. From the PCA, 5 PCS were retained accounting for 84% of the variance in the data.

B. Inter-Rater Agreements

There were moderate to good agreements between the three clinical raters with ICCs of 0.80 for TRS, 0.58 for UPDRS #31, 0.69 for SUMUPDRS and 0.67 for Dyskinesia [31]. Based on these results it was decided to take the mean clinical score of the three raters per patient and time point and use them in the machine learning methodology.

C. Convergent Validity

After employing machine learning methods (SVM, LR, and DT) on the selected features by stepwise regression and PCA, the absolute correlation coefficients between scores produced by machine learning methods and mean clinical ratings ranged from 0.29 to 0.83 (Table IV). The best combination (feature selection plus machine learning method) was stepwise regression and SVM. The validity results from applying PCA as a feature selection method were lower than applying stepwise regression. These results were consistent for the three machine learning

methods. Therefore, it was decided to use the scores produced from stepwise regression and SVM in subsequent analysis.

After analysing the measurement bias between the scores produced by SVM and the mean ratings of the three raters on the four scales, it was found that the SVMs overestimated cases at the lower end of the scales and underestimated cases at the higher end of the scales (Fig. 2). The errors (mean biases ±95% CI) were -0.05 ± 1.31 for TRS, 0.04 ± 0.95 for UPDRS #31, 0.25 ± 2.99 for SUMUPDRS and 0.05 ± 0.88 for Dyskinesia. These results indicate evidence of a small bias between the scores produced by SVMs and clinicians since the mean biases were close to zero.

D. Test-Retest Reliability

There were high ICCs for clinical rating scales and the SVM-based scores during the first two baseline measurements (Table V), indicating good test-retest reliability. For this analysis, data from 18 patients were used since one of the patients did not have a baseline measurement. The test-retest reliability for dyskinesia was not assessed since the patients exhibited no dyskinesia during the first two baseline measurements.

E. Responsiveness to Treatment

As shown in Fig. 3, the SVM-based scores on the four scales were reasonably responsive in relation to the clinical scores. The biggest gap between the effect sizes could be seen when assessing responsiveness of dyskinesia.

F. Analysis of Results in Asymmetric Patients

In the sample set, 89% of patients had asymmetrical motor symptoms. There were 9 patients who were affected mostly on their right side and 8 patients on the left side (Table I). In the two groups of patients, the PC1s between right and left legs were not different (P-value = 0.40 for right side affected PD patients; P-value = 0.48 for left side affected PD patients). These results indicate that the PD asymmetry did not have any effect on the performance of the leg agility tests, as measured by the PC1.

IV. Discussion and Conclusions

Adequate assessment of the severity of motor states and therapy-related complications of PD patients is needed in order to individualize and optimize the treatments. In this paper, we have presented the development and evaluation of methods for scoring PD motor states in an objective manner by employing machine learning methods and sensor technology. The results from this study indicate that sensor-based objective measures can be used for assessing PD motor states, as scored by the clinical TRS scale. The methodology could be further integrated into sensor-based dosing systems for individualizing dose schedules, which in turn could improve care in general as well as PD outcomes [32]. Employing the proposed methodology in clinical tools may lead to a more efficient approach for PD management and follow-up of treatment effects of patients on an individual basis.

After extracting 24 spatiotemporal features from motion sensor data, the final feature sets selected by the stepwise regression models included features from both accelerometers and gyroscopes, indicating that both sensors were equally important for capturing clinically-relevant movement information. When investigating the number of features that were represented in all four models it was found that there were only 4 features that were selected, out of which 3 were based on gyroscope and 1 was based on the accelerometer. The selected features were then used as inputs to machine learning methods to quantify the states on the clinical scales. Comparing the performance of the machine learning methods when using inputs from stepwise regression and PCA, it was found that stepwise regression was superior in terms of convergent validity. Similar results were found by our previous work [33] where step-wise regression outperformed PCA in terms of SVM's validity and responsiveness to treatment. Our results were in line with the results reported in studies by other research groups [34], [35] where stepwise regression provided better results than PCA in terms of validity and test-retest reliability.

The proposed methods showed good validity and test-retest reliability in PD motor symptom quantification. The best combination of feature selection and machine learning methods was stepwise regression and SVM with correlation coefficients ranging from 0.67 to 0.83 in relation to the clinician-based ratings on four scales (TRS, bradykinesia scale, SumUPDRS and dyskinesia). Assessing the responsiveness of the scores derived from SVM showed the ability of this method to capture the treatment effects in relation to TRS and dyskinesia rating scales (Fig. 3). The results in the present study were similar to those reported in our previous research where hand pronation-supination data was used for the same purpose [25]. The correlation coefficient (RMSE) for automatic scoring of TRS were 0.81 (0.77), slightly lower than in the previous study 0.82 (0.73). Similarly, in the study performed by Parisi *et al.* [19] a high correlation (0.74) was found between automatic and clinician-based scores on UPDRS, using data from leg agility tests in PD patients. Das *et al.* [36] quantified PD motor states using an optical motion capture system during various motor tasks, including leg agility tests, in PD patients treated with DBS and reported that their method showed that SVM classifiers had accuracy of 95.8% in discriminating mild vs. severe states and 70% in discriminating On vs. Off motor states. Furthermore, quantitative measurement of movements in PD patients during leg agility tests was evaluated using a Kinect-based system in relation to a Vicon 3D motion analysis system and good accuracy of Kinect was reported in measuring spatiotemporal characteristics of movements [37].

This study has the following limitations. First, the number of patients is low and results require further confirmation in a larger trial. Second, as seen in Fig. 2 the range of the states was underestimated by SVM. However, the relative agreement was still very good with 95% CIs of the errors of ±1.31 units for TRS, ±0.95 units for UPDRS #31, ±2.99 units for SUMUPDRS and ±0.88 units for Dyskinesia. After investigating the individual outliers, it was found that they belonged to 2 patients, who were levodopa non-responsive subjects since they had steady mean TRS ratings across the time course of single

dose. This could also impact the scoring on the same scale by SVM. In this study, a large proportion of measurements assigned by TRS ratings were between -2 and $+1$, which made it difficult for the machine learning methods to predict cases outside of this range. Compared to our previous work on quantifying PD motor states using motion sensor data during walking tests [12], our results show that analysis of leg agility data is less suitable for capturing dyskinesias since the correlations and RMSE in this study were lower; 0.67, 0.5 vs. 0.79, 0.47. This was also reflected in the lower responsiveness to treatment as compared to clinical TRS, as shown in Fig. 3. Due to the abovementioned limitations the methods tended to underestimate the scores and to concentrate their predictions around the mean of the population. Nevertheless, detection of severe dyskinesias is complicated in patients with mild to moderately impaired motor states, as shown in the study performed by Tsipouras *et al*. [11]. Accuracy of sensor-based systems to detect dyskinesias has also been shown to be different depending on the presence of dyskinesias in different parts of the body. For instance, in the study performed by Perez-Lopez *et al*. there was a low responsiveness and specificity of the methods when using a belt attached on the torso and assessing mild dyskinesias in distal parts of extremities.

The methods presented in this study could be useful in monitoring changes in PD and possibly making individualized treatments feasible [32], [38], [39]. This would mean including functionalities to the system for on-line analysis of the data where the data of interest would automatically be identified and processed accordingly. Future research will focus on evaluating the feasibility of the methods to be used for home monitoring. Another interesting area would be to evaluate the clinimetric properties of the methods when using sensor data from multiple tests and determine the feasibility for aiding individualized adjustments of doses in PD patients [38]. The machine learning methods could be further optimized by training and testing them with datasets from larger set of patients and with a more homogeneous distribution of cases across different severity levels of the UPDRS and TRS scales. In addition, training the raters on how to rate the patients particularly according to the TRS scale and adding more raters for evaluating the states would reduce the inter-rater variability, which in turn could improve the accuracy of the machine learning methods.

In conclusion, this study demonstrates good clinimetric properties of a data-driven methodology that quantifies the motor states in PD using motion sensors data during leg agility tests. The proposed methodology could form the basis for developing systems for follow up of the effects of treatment and individualizing treatments in PD.

References

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

[2] J. A. Stamford, P. N. Schmidt, and K. E. Friedl, "What Engineering technology could do for quality of life in Parkinson's disease: A review of current needs and opportunities," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1862–1872, Nov. 2015.

[3] J. G. Nutt, S. T. Gancher, and W. R. Woodward, "Does an inhibitory action of levodopa contribute to motor fluctuations?" *Neurology*, vol. 38, no. 10, pp. 1553–1557, Oct. 1988.

[4] M. Merello and A. J. Lees, "Beginning-of-dose motor deterioration following the acute administration of levodopa and apomorphine in parkinsons-disease," *Neurol. Neurosurgery Psychiatry*, vol. 55, no. 11, pp. 1024–1026, Nov. 1992.

[5] R. A. Hauser, F. Deckers, and P. Lehert, "Parkinson's disease home diary: Further validation and implications for clinical trials," *Movement Disorders*, vol. 19, no. 12, pp. 1409–1413, Dec. 2004.

[6] 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," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, Nov. 2008.

[7] 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, Dec. 2005.

[8] E. Rovini, C. Maremmani, and F. Cavallo, "How wearable sensors can support Parkinson's disease diagnosis and treatment: A systematic review," *Front. Neurosci.*, vol. 11, Oct. 2017, Art. no. 555.

[9] D. Johansson, K. Malmgren, and M. Alt Murphy, "Wearable sensors for clinical applications in epilepsy, Parkinson's disease, and stroke: A mixed-methods systematic review," *J. Neurol.*, vol. 265, pp. 1740–1752, Feb. 2018.

[10] R. Ramsperger *et al.*, "Continuous leg dyskinesia assessment in Parkinson's disease -clinical validity and ecological effect," *Parkinsonism Related Disorders*, vol. 26, pp. 41–46, May 2016.

[11] M. G. Tsipouras, A. T. Tzallas, G. Rigas, S. Tsouli, D. I. Fotiadis, and S. Konitsiotis, "An automated methodology for levodopa-induced dyskinesia: Assessment based on gyroscope and accelerometer signals," *Artif. Intell. Med.*, vol. 55, no. 2, pp. 127–135, Jun. 2012.

[12] I. Thomas, F. Bergquist, R. Constantinescu, D. Nyholm, M. Senek, and M. Memedi, "Using measurements from wearable sensors for automatic scoring of Parkinson's disease motor states: Results from 7 patients," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Seogwipo, South Korea, 2017, pp. 131–134.

[13] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation," *J. Neuroeng. Rehabil.*, vol. 15, no. 1, pp. 97–110, Nov. 2018.

[14] S. Arora *et al.*, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Related Disorders*, vol. 21, no. 6, pp. 650–653, Jun. 2015.

[15] S. Memar, M. Delrobaei, M. Pieterman, K. McIsaac, and M. Jog, "Quantification of whole-body bradykinesia in Parkinson's disease participants using multiple inertial sensors," *Neurological Sci.*, vol. 387, pp. 157–165, Apr. 2018.

[16] A. Delval, L. Defebvre, and C. Tard, "Freezing during tapping tasks in patients with advanced Parkinson's disease and freezing of gait," *PLoS One*, vol. 12, no. 9, 2017, Art. no. e0181973.

[17] L. Palmerini, S. Mellone, G. Avanzolini, F. Valzania, and L. Chiari, "Quantification of motor impairment in parkinson's disease using an instrumented timed Up and Go Test," *IEEE Trans. Neural. Syst. Rehabil. Eng.*, vol. 21, no. 4, pp. 664–673, Jul. 2013.

[18] M. Giuberti *et al.*, "On the characterization of leg agility in patients with Parkinson's disease," in *Proc. IEEE Int. Conf. Body Sens. Netw.*, 2013, pp. 1–6.

[19] F. Parisi *et al.*, "Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, sit-to-stand, and gait tasks in parkinson's disease," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1777–1793, Nov. 2015.

[20] D. Nyholm *et al.*, "Duodenal levodopa infusion monotherapy vs oral polypharmacy in advanced Parkinson disease," *Neurology*, vol. 64, no. 2, pp. 216–223, 2005.

[21] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, no. 1, pp. 41–47, Jan. 2007.

[22] M. Senek *et al.*, "Levodopa/carbidopa microtablets in Parkinson's disease: A study of pharmacokinetics and blinded motor assessment," *Clin. Pharmacol.*, vol. 73, no. 5, pp. 563–571, May 2017.

[23] S. Fahn and R. L. Elton, "Unified Parkinson's disease rating scale," in *Recent Developments in Parkinson's Disease*. Florham Park, NJ, USA: Macmillan Health Care Inf., 1987, vol. 2, pp. 153–163.

[24] J. Westin *et al.*, "A new computer method for assessing drawing impairment in Parkinson's disease," *Neurosci. Methods*, vol. 190, no. 1, pp. 143–148, Jun. 2010.
[25] I. Thomas *et al.*, "A treatment-response index from wearable sensors for quantifying parkinson's disease motor states," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1341–1349, Sep. 2018.
[26] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Natl. Acad. Sci. USA*, vol. 88, no. 6, pp. 2297–301, Mar. 1991.
[27] S. Aghanavesi, M. Memedi, M. Dougherty, D. Nyholm, and J. Westin, "Verification of a method for measuring Parkinson's disease related temporal irregularity in spiral drawings," *Sensors*, vol. 17, no. 10, Oct. 2017, Art. no. E2341.
[28] R. K. White, "Practical applications of quantitative structure-activity-relationships (Qsar) in environmental chemistry and toxicology," *Risk Anal.*, vol. 12, no. 1, pp. 156–157, Mar. 1992.
[29] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, no. 8476, pp. 307–310, Feb. 1986.
[30] C. G. Goetz *et al.*, "Which dyskinesia scale best detects treatment response?" *Movement Disorders*, vol. 28, no. 3, pp. 341–346, Mar. 2013.
[31] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016.
[32] I. Thomas *et al.*, "Sensor-based algorithmic dosing suggestions for oral administration of levodopa/carbidopa microtablets for Parkinson's disease: a first experience," *J. Neurol*, vol. 266, no. 3, pp. 651–658, Mar. 2019.
[33] F. Javed, I. Thomas, and M. Memedi, "A comparison of feature selection methods when using motion sensors data: A case study in Parkinson's disease," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA, 2018, pp. 5426–5429.
[34] H. Ssegane, E. W. Tollner, Y. M. Mohamoud, T. C. Rasmussen, and J. F. Dowd, "Advances in variable selection methods I: Causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships," *Hydrology*, vol. 438, pp. 16–25, May 2012.
[35] Z. Y. Liu, J. F. Huang, J. J. Shi, R. X. Tao, W. Zhou, and L. L. Zhang, "Characterizing and estimating rice brown spot disease severity using stepwise regression, principal component regression and partial least-square regression," *Zhejiang Univ. Sci. B*, vol. 8, no. 10, pp. 738–744, Oct. 2007.
[36] S. Das *et al.*, "Quantitative measurement of motor symptoms in Parkinson's disease: A study with full-body motion capture data," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 6789–6792.
[37] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the microsoft kinect sensor for measuring movement in people with Parkinson's disease," *Gait Posture*, vol. 39, no. 4, pp. 1062–1068, Apr. 2014.
[38] D. Johansson *et al.*, "Individualization of levodopa treatment using a microtablet dispenser and ambulatory accelerometry," *CNS Neurosci. Therapeutics*, vol. 24, pp. 439–447, 2018.
[39] N. Khobragade, D. Graupe, and D. Tuninetti, "Towards fully automated closed-loop deep brain stimulation in parkinson's disease patients: A LAMSTAR-based tremor predictor," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, vol. 2015, pp. 2616–2619.