Received January 27, 2022, accepted February 25, 2022, date of publication March 3, 2022, date of current version March 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156659

Smartphone-Based Digital Biomarkers for Parkinson's Disease in a Remotely-Administered Setting

MARÍA GOÑI, SIMON B. EICKHOFF, MEHRAN SAHANDI FAR, KAUSTUBH R. PATIL, (Member, IEEE), AND JUERGEN DUKART

Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, 52425 Jülich, Germany
Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

Corresponding author: Juergen Dukart (juergen.dukart@gmail.com)

ABSTRACT Smartphone-based digital biomarker (DB) assessments provide objective measures of daily-life tasks and thus hold the promise to improve diagnosis and monitoring of Parkinson's disease (PD). To date, little is known about which tasks perform best for these purposes and how different confounds including comorbidities, age and sex affect their accuracy. Here we systematically assess the ability of common self-administered smartphone-based tasks to differentiate PD patients and healthy controls (HC) with and without accounting for the above confounds. Using a large cohort of PD patients and healthy volunteers acquired in the mPower study, we extracted about 700 features commonly reported in previous PD studies for gait, balance, voice and tapping tasks. We perform a series of experiments systematically assessing the effects of age, sex and comorbidities on the accuracy of the above tasks for differentiation of PD patients and HC using several machine learning algorithms. When accounting for age, sex and comorbidities, the highest balanced accuracy on hold-out data (73%) was achieved using random forest when combining all tasks followed by tapping using relevance vector machine (67%). Only moderate accuracies were achieved for other tasks (60% for balance, 56% for gait and 53% for voice data). Not accounting for the confounders consistently yielded higher accuracies of up to 77% when combining all tasks. Our results demonstrate the importance of controlling DB data for age and comorbidities.

INDEX TERMS Digital biomarkers, machine learning, Parkinson's disease, smartphones, wearable devices.

I. INTRODUCTION

Diagnosis of Parkinson's disease (PD) still often relies on in-clinic visits and evaluation based on clinical judgement as well as patient and caregiver reported information. This lack of objective measures and the need for in-clinic visits result in the often late and initially inaccurate diagnosis [1]. Recent studies have identified digital assessments as such promising objective biomarkers for PD symptoms including bradykinesia [2], [3], freezing of gait [4], [5], impaired dexterity [6], balance and speech difficulties [7]–[9]. Most of these results were obtained with a moderate number of participants and in a standardized and controlled clinical setting, reducing generalizability and limiting an interpretation with respect to applicability of these measures to an at-home self-administered setting [10]–[12].

As most relevant sensors deployed in these in-clinic studies are also embedded in modern smartphones, this opens the possibility to collect such objective, reliable and quantitative information as digital biomarkers (DB) in an at-home setting and therewith to facilitate diagnosis, health monitoring or treatment management using low-cost, simple and portable technology [13]. Indeed, recent studies applying machine learning algorithms to these high-dimensional data suggested a good diagnostic sensitivity of the respective digital assessments for detection of Parkinson's disease [14]–[17]. However, such at-home assessments create a range of new challenges including selection bias, confounding and sources of noise that need to be understood and dealt with to ensure good reliability of respective outcomes to a level that is sufficient for at home data collection [18]. For example, age, sex and comorbidities are known confounding factors that impact many measures of disease symptoms across neurodegenerative diseases including PD [19]–[23]. Yet, several

The associate editor coordinating the review of this manuscript and approving it for publication was Masood Ur-Rehman.

TABLE 1. Demographics for PD and HC subjects for each experiment. Those cases where age or sex are significantly different between PD and HC are indicated with an asterisk (2 sample t-test for age and Chi-square for sex with 95% confidence).

	Gait		Balance		Voice		Tapping		Multimodal	
	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC
Experiment 1 (all)										
N	610	787	612	803	893	1257	970	1630	597	742
Male/ female	399/ 211*	640/ 147*	401/ 211*	653/ 150*	571/ 322*	1018/ 239*	630/ 340*	1336/ 294*	390/ 207*	607/ 135*
Age (mean±sd)	60.3 ± 8.94*	49.04 ± 10.71*	60.29 ± 8.94*	48.9 ± 10.72*	60.13 ± 8.97*	47.65 ± 10.41*	59.85 ± 9.05*	46.84 ± 10.05*	60.36 ± 8.86*	49.22± 10.78*
UPDRS mean±sd (n)	13.17 ± 7.78 (350)	-	13.14 ± 7.78 (351)	-	13.48 ± 7.93 (566)	-	13.44 ± 7.89 (588)	-	13.15 ± 7.8 (344)	-
UPDRS I mean±sd (n)	5.66 ± 3.64 (361)	-	5.64 ± 3.64 (362)	-	5.64 ± 3.63 (586)	-	5.63 ± 3.63 (608)	-	5.64 ± 3.64 (355)	-
UPDRS II mean±sd (n)	7.59 ± 5.18 (350)	-	7.58 ± 5.18 (351)	-	7.9 ± 5.53 (572)	-	7.86 ± 5.49 (594)	-	7.59 ± 5.21 (344)	-
Experiment 2 (matched)										
N	373	373	376	376	534	534	608	608	361	361
Male/ female	278/ 95	286/ 87	280/ 96	288/ 88	379/ 155	394/ 140	435/ 173	450/ 158	270/ 91	278/ 83
Age (mean±sd)	57.09 ± 9.4	57.09 ± 9.4	57.1 ± 9.4	57.1 ± 9.4	56.54 ± 9.3	56.54 ± 9.3	56.38 ± 9.23	56.38 ± 9.23	57.18 ± 9.36	57.18 ± 9.36
UPDRS mean±sd (n)	14.38 ± 8.48 (206)	-	13.61 ± 8.34 (202)	-	13.76 ± 8.21 (324)	-	13.7 ± 8.08 (349)	-	14.37 ± 8.51 (190)	-
UPDRS I mean±sd (n)	6.16 ± 3.91 (213)	-	5.73 ± 3.85 (207)	-	5.84 ± 3.78 (333)	-	7.8 ± 3.71 (361)	-	5.99 ± 3.82 (198)	-
UPDRS II mean±sd (n)	8.27 ± 5.66 (206)	-	7.95 ± 5.49 (202)	-	7.98 ± 5.6 (328)	-	7.97 ± 5.53 (352)	-	8.45 ± 5.69 (190)	-
Experiment 3 (no comorbidities, matched), experiment 4 (no comorbidities, matched, age controlled), experiment 5 (no comorbidities, matched, sex controlled) and experiment 6 (no comorbidities, matched, controlled)										
N	317	317	320	320	446	446	507	507	306	306
Male/ female	230/ 87	244/ 73	232/ 88	246/ 74	314/ 132	332/ 114	359/ 148	377/ 130	223/ 83	238/ 68
Age (mean±sd)	56.34 ± 9.42	56.34 ± 9.42	56.37 ± 9.41	56.37 ± 9.41	56 ± 9.31	56 ± 9.31	55.71 ± 9.22	55.71 ± 9.22	56.45 ± 9.37	56.45 ± 9.37
UPDRS mean±sd (n)	13.36 ± 7.94 (166)	-	13.53 ± 7.99 (174)	-	13.42 ± 7.63 (275)	-	13.56 ± 7.62 (296)	-	13.5 ± 7.95 (165)	-
UPDRS I mean±sd (n)	5.77 ± 3.71 (172)	-	5.84 ± 3.71 (179)	-	5.56 ± 3.52 (284)	-	5.81 ± 3.54 (304)	-	5.86 ± 3.65 (172)	-
UPDRS II mean±sd (n)	7.65 ± 5.4 (166)	-	7.77 ± 5.44 (174)	-	7.95 ± 5.41 (278)	-	7.85 ± 5.34 (301)	-	7.75 ± 5.38 (165)	-

studies eluded the importance of matching and controlling for these variables [24]–[26], including age, sex [24], [27] or comorbidities which might induce motor (i.e. bradykinesia, tremor or rigidity) and non-motor (i.e. fatigue, restless legs or sleep) symptoms [25]. Other potential data collection biases include small sample sizes [14], [28], inclusion of several recordings per subject [15], [24] or signals of different time lengths [27], which may potentially lead the classifier to detect the idiosyncrasies of each subject rather than specific PD related symptoms, as demonstrated by Neto *et al.* [29]–[31]. In addition, replicability of results is rarely performed in current studies, which may lead to lack of generalizability. Despite the considerable promise for DB in healthcare, these issues limit comparability across studies, hindering interpretation and obstructing translation to the clinic.

Recently, a large dataset of at-home smartphone-based assessments of commonly applied PD tasks including gait, balance, finger tapping and voice evaluations was collected in the mPower study providing a unique resource to examine DB in the study of PD [32], [33]. Indeed, several studies applying machine learning (ML) algorithms have employed this dataset in the study of PD diagnosis, achieving quite different results across studies. Whilst plausible, the impact of the aforementioned confounds on ML-based detection of PD using different at-home digital assessments has not been yet systematically established and has indeed been ignored in many previous studies [15], [24], [27], [34], [35].

Here we systematically explore the influence of accounting for age, sex and comorbidities in the detection of PD in a large at-home dataset. Concretely, we use the mPower dataset to evaluate the ability of common DB task (gait, balance, voice, tapping) for differentiation between PD and HC. In addition, we identify potential DB of Parkinson's disease. With this work, we aim to outline practical suggestions to guide future studies practices and improve comparability across studies.

II. METHODS

A. DATA

Data used in this work were derived from the mPower study [32]. MPower is a mobile application-based study to monitor indicators of PD progression and diagnosis by the

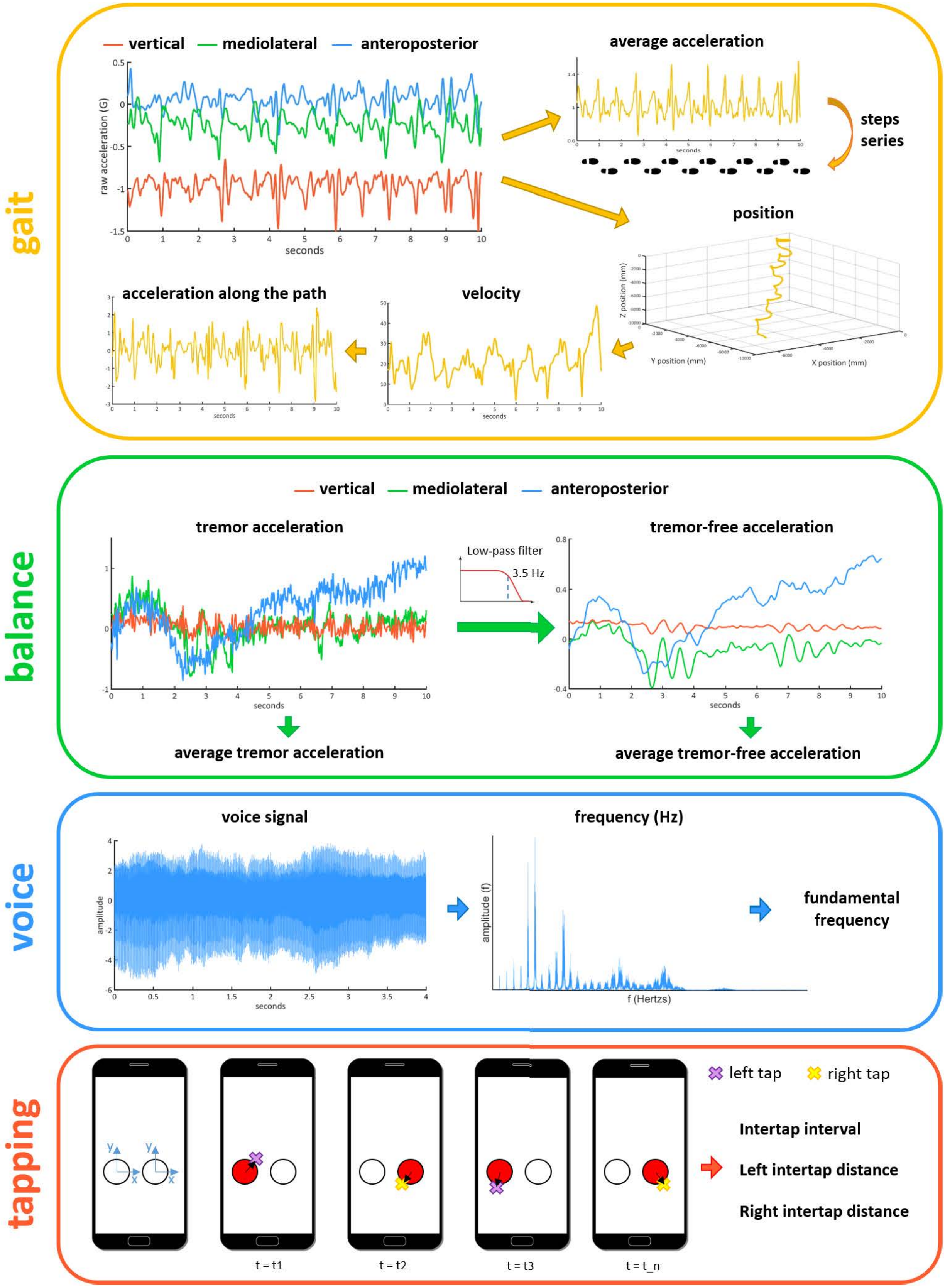

FIGURE 1. Illustration of signal processing and feature extraction based on the raw data for each task.

TABLE 2. List of experiments indicating their corresponding processing steps.

	Exclude comorbidities	Age & sex matching	Control for age	Control for sex	Control for age & sex
E1	-	-	-	-	-
E2	-	x	-	-	-
E3	x	x	-	-	-
E4	x	x	x	-	-
E5	x	x	-	x	-
E6	x	x	x	x	x

E: Experiment

TABLE 3. Balanced accuracy results for CV and holdout datasets and chance level at 95%.

	Experiment 1			Experiment 2			Experiment 3			Experiment 4			Experiment 5			Experiment 6		
	CV	H	C	CV	H	C	CV	H	C	CV	H	C	CV	H	C	CV	H	C
Gait	56.6 (54.3-58.9)	57.1	52.6 (51.1-54.2)	50.3 (47.2-53.6)	54.8	50 (47.1-52.6)	56.5 (53.3-59.7)	55.7	49.9 (46.6-53.4)	56.5 (53.3-59.5)	54.8	50 (46.7-53.3)	56.4 (53.1-59.7)	56.2	50 (46.4-53.6)	56.7 (53.3-59.9)	53.8	50.1 (46.9-53.2)
Balance	61.8 (60.4-63.4)	64.7	50.2 (45.7-54.4)	60.4 (58.6-62.4)	58	49.8 (43.2-56)	60 (57.6-62.3)	59.9	49.9 (43.4-56.6)	60.6 (57.2-63.8)	61.3	50.1 (43.4-56.6)	60.1 (56.5-63.3)	61.3	50.1 (43.4-57.6)	60.2 (57.0-63.6)	59.9	50.2 (42.9-57.1)
Voice	62.5 (61.3-63.6)	60.4	50 (46.5-53.5)	53.9 (51.5-56.2)	59.8	50.1 (44.7-55.3)	56.7 (54.4-58.9)	53	50 (43.6-55.7)	56.9 (54.7-59.1)	58.1	50 (44.6-56.1)	60 (57.7-62.1)	60.1	49.8 (43.9-55.4)	59.2 (57.1-61.2)	59.1	50.2 (43.6-55.7)
Tapping	74.8 (74.4-75.2)	72.9	50 (47-52.9)	66.8 (66-67.6)	66.8	49.9 (45.1-55)	67.9 (67-68.9)	67.2	50.1 (44.1-55.9)	68.8 (67.9-69.8)	66.9	50 (44.4-55)	68.7 (67.6-69.7)	68.9	50.2 (45.3-55.3)	68.8 (67.8-69.8)	68.1	50 (45-55.6)
Multimodal	73.9 (72.4-75.5)	76.9	50 (45.1-54.9)	69.4 (67-71.9)	70	50.1 (44.2-56.7)	69.6 (66.9-72.4)	73.5	50 (43.1-56.9)	69.2 (66.2-71.8)	73	50.2 (43.6-56.4)	68 (65-70,8)	69.1	50 (43.6-57.4)	69.9 (67.2-72.8)	70.6	50 (43.1-56.9)

collection of data in subjects with and without PD. Using this app, subjects were presented with a one-time demographic survey about general demographic topics and health history. Completion of the Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Parkinson's Disease Questionnaire short form (PDQ-8) surveys used for PD assessment was requested at baseline as well as monthly throughout the course of the study. Due to the length of the MDS-UPDRS instrument, subjects were presented only a subset of questions focusing largely on the monitor symptoms of PD [32]. Participants had to select "true" or "false" to the following question "Have you been diagnosed by a medical professional with Parkinson Disease?". According to this answer, they were classified as Parkinson's Disease (PD) or Healthy Control (HC). Subjects who did not answer this question were discarded from further analysis. All subjects were presented with different tasks including gait, balance, voice and tapping, which they could complete up to 3 times per day. Subjects who self-identified as having a professional diagnosis of PD were asked to perform these tasks (1) immediately before taking their medication, (2) after taking their medication and (3) at some other time (Table 8). Subjects who self-identified as not having a diagnosis of PD could complete these tasks at any time during the day. In the gait task, subjects were asked to walk 20 steps in a straight line. In the balance task they were required to stand still for 30 seconds. During the voice activity task, subjects were requested to say 'Aaah' into the microphone for 10 seconds. Finally, during the tapping task participants were instructed to alternatively tap two points on the screen within a 20 seconds interval. We additionally excluded those subjects who gave no information about their age, sex or had inconsistencies in their clinical data (e.g. self-reported healthy controls who answered questions about PD diagnosis or PD medication). Since the mPower dataset is strongly slanted toward young HC (Table 15), we restricted our analysis to those subjects within the age range of 35 to 75 years old. This cleaning step resulted in the exclusion of 40-50% of the data depending on the task. To avoid "learning effects" and biases due to several recordings, we only considered the first recording of each

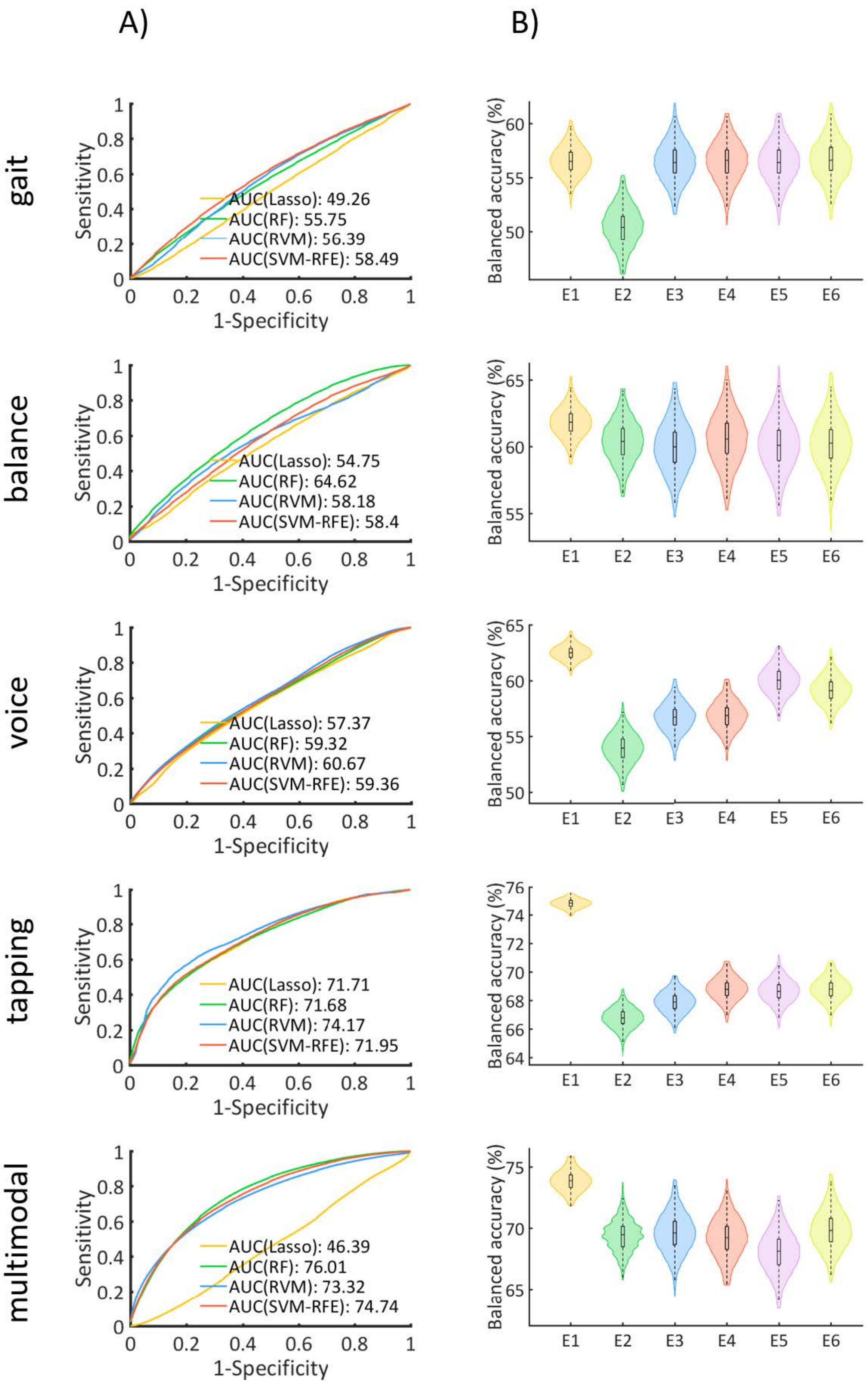

FIGURE 2. A) ROC curves and AUC values for 4 different classifiers for each task, during the main experiment (E3: no comorbidities, matched). B) Balanced accuracy distributions for each task and experiment (E1-E6). E1: all data. E2: age and sex matched. E3: no comorbidities, age and sex matched. E4: no comorbidities, age and sex matched, controlled for age. E5: no comorbidities, age and sex matched, controlled for sex. E6: no comorbidities, age and sex matched, controlled for age and sex.

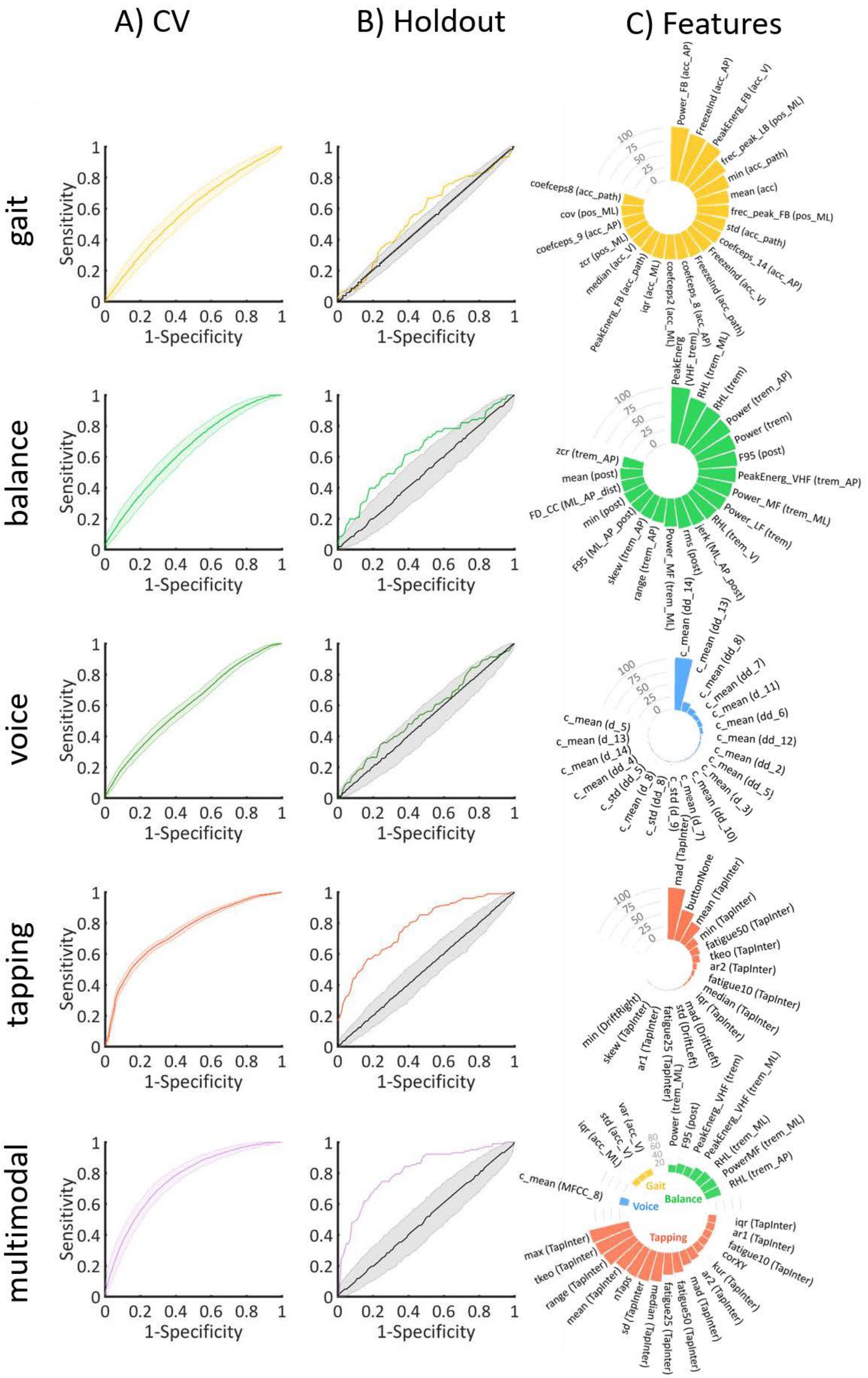

FIGURE 3. A) ROC curves at 95% CI during CV. B) ROC curves at 95% CI during validation of holdout set and at the chance level. C) Scaled average weights of features for each task for the main experiment (E3: no comorbidities, matched). Gait) acc - average acceleration, acc_path – acceleration along path, AP – anteroposterior, FB – freezing band, LB – locomotor band, ML – mediolateral, pos – position, V – vertical, vel – velocity. Balance) trem – tremor, post – postural, dist – distance, LF – low frequency, MF – medium frequency, VHF – very high frequency, RHL – ratio between high and low frequency, F95 –frequency containing 95% of the power spectrum. Voice) c – cepstral coefficient, d – 1st derivative of cepstral coefficient, dd – 2nd derivative of cepstral coefficient. Tapping) TapInter – tap interval. For details on features refer to Appendix A.

TABLE 4. List of gait features.

Feature acronym	Feature description	Signal (acronym)
numSteps	Number of steps during the 10 seconds gait signal.	
MSI	Mean Stride Interval, calculated as the duration of a stride averaged over all strides [58], [64].	
StrideVar	Stride Variability, calculated as the standard deviation divided by the mean stride of the stride interval. Measures consistency and stability [58], [64].	
mean	Mean value of the observations [15], [40].	vertical, anteroposterior, mediolateral and average acceleration (acc_V, acc_AP, acc_ML, acc) vertical, anteroposterior, mediolateral and average position (pos_V, pos_AP, pos_ML, pos) velocity (vel) acceleration along path (acc_path)
min	Minimum value of the observations.	
max	Maximum value of the observations.	
median	Median. Middle value among a dataset [15], [40].	
sd	Standard deviation, calculated as the sum of squares differences between the individual values and the mean. Measures variability [15], [40].	
var	Variance, calculated as the square of the standard deviation. Measures variability.	
range	Range of the observations.	
iqr	Interquartile range, calculated as the difference between 75th and 25th percentiles. Measures dispersion [15], [40].	
rms	Root mean square of the observations.	
cov	Coefficient of variation, calculated as the standard deviation of the signal divided by the mean.	
skew	Skewness. Describes the asymmetry of a signal. A negative value indicates that the distribution is concentrated on the right, while a positive one is concentrated in the left [15], [40], [58].	
kur	Kurtosis. Measures if data is heavy or light-tailed to a normal distribution [40], [58].	
zcr	Zero-crossing rate. Rates sign-changes along a signal [15].	
ApEn	Entropy. Measures uncertainty, ranging from 0-1 where 0 indicates randomness and 1 maximum regularity [15], [58].	
PeakEnerg_LB	Peak of energy in the locomotor band (0.5-3 Hz) [41], [64].	
frec_peak_LB	Frequency at the peak of energy in the locomotor band (0.5-3 Hz) [41].	
Power_LB	Power of the locomotor band (0.5-3 Hz) [41].	
PeakEnerg_FB	Peak of energy in the freezing band (3-8Hz) [42].	
frec_peak_FB	Frequency at the peak of energy in the freezing band (3-8 Hz).	
Power_FB	Power in the freezing band (3-8 Hz).	
FreezeInd	Freeze Index. Calculated as the ratio between the power in the freezing band (3-8 Hz) and the power in the locomotor band (0.5-3 Hz) [42].	
RatioPower	Sum of the power in the freezing (3-8 Hz) and locomotor band (3-8 Hz) [43]	
ar	Coefficient of a 1st order autoregressive model. An autoregressive model forecasts when there is some correlation between current values and their preceding ones [40].	
coefceps_(1-20)	20 Mel Frequency Cepstral Coefficients. Represent the short-term power spectrum [42]	

subject in the analyses. Further details about data cleaning can be found in Appendix A. Demographic details are shown in Table 1.

B. PRE-PROCESSING

The tri-axial accelerometer integrated in the smartphone records acceleration in the 3 axes (vertical, mediolateral and anteroposterior) during the gait and balance tasks. A 4th order 20 Hz cut-off low-pass Butterworth filter was applied to the 3 accelerometer signals. An additional 3rd order 0.3 Hz cut-off high-pass Butterworth filter was applied to minimize the acceleration variability due to respiration [36]. Signals were then standardized to eliminate the gravity component while maintaining the information from outlier data. According to Pittman *et al.* [24], 30% of the devices were not held in the correct position and therefore, we additionally calculated the average acceleration signal. Several signals were extracted from the gait recordings including the step series, position along the 3 axes calculated by double integration, velocity and acceleration along the path [37] (Figure 1).

Two additional signals were considered for the balance task (Figure 1). Tremor frequency in PD is estimated to fall in the 4-7 Hz band [38], whereas postural acceleration measures (tremor-free) fall in the 0-3.5 Hz interval. To extract tremor-free measures of postural acceleration, we applied a 3.5 Hz cut-off low-pass Butterworth filter [39].

Voice was recorded at a sample rate of 44.1 Kbps. Pre-processing included a downsampling to 25 KHz and a noise reduction using a 2nd order Butterworth filter with a low-pass frequency at 400 Hz. The fundamental frequency signal was calculated using a Hamming window of 20 ms with 50% overlap, and verified with the software Praat (Figure 1). Time, frequency and amplitude series were extracted from the voice signals.

TABLE 5. List of balance features.

Acronym	Description	Signal (acronym)
mean	Mean value of the observations.	
min	Minimum value of the observations.	
max	Maximum value of the observations.	
median	Median value of the observations.	
sd	Standard deviation, calculated as the sum of squares differences between the individual values and the mean. Measures variability.	
var	Variance, calculated as the square of the standard deviation. Measures variability.	
range	Range of the observations.	vertical, anteroposterior, mediolateral and average tremor acceleration (trem_V, trem_AP, trem_ML, trem)
iqr	Interquartile range, calculated as the difference between 75th and 25th percentiles. Measures dispersion.	
rms	Root mean square of the observations.	
cov	Coefficient of variation, calculated as the standard deviation of the signal divided by the mean.	vertical, anteroposterior, mediolateral and average postural acceleration (post_V, post _AP, post _ML, post)
skew	Skewness. Describes the asymmetry of a signal. A negative value indicates that the distribution is concentrated on the right, while a positive one is concentrated in the left.	
kur	Kurtosis. Measures if data is heavy or light-tailed to a normal distribution.	
zcr	Zero-crossing rate. Rates sign-changes along a signal.	
ApEn	Entropy. Measures uncertainty, ranging from 0-1 where 0 indicates randomness and 1 maximum regularity.	
Power_MF	Power of the medium frequency band (4-7Hz) [39].	vertical, anteroposterior, mediolateral and average tremor acceleration (trem_V, trem_AP, trem_ML, trem)
PeakEnerg_VHF	Peak of energy in the very high frequency band (>7Hz)	
frec_peak_HF	Frequency at the peak of energy in the high frequency band (>4Hz) [39].	
Power	Power between 3.5-15Hz	
Power_LF	Power in the low frequency band (0.15-3.5Hz)	
RHL	Ratio between the power between 3.5-15Hz and power between 0.15-3.5Hz [39].	
CFREQ	Centroidal frequency for postural measures. Also known as zero-crossing frequency [36], [39], [44].	anteroposterior, mediolateral and average postural acceleration (post _AP, post _ML, post) mediolateral-anteroposterior average postural acceleration (ML_AP_post)
FREQD	Frequency of dispersion of the power spectrum for postural measures [36], [39], [44].	
jerk	Average jerk. Measures vibration as the rate of change in acceleration. Calculated as the derivative of acceleration with respect to time [36], [39].	vertical, anteroposterior, mediolateral and average postural acceleration (post_V, post _AP, post _ML, post) mediolateral-anteroposterior average postural acceleration (ML_AP_post)
TotalPower	Energy between 0.15-3.5Hz for postural measures [36].	
F50	Frequency containing 50% of the total power for postural measures [36], [39].	
F95	Frequency containing 95% of the total power for postural measures [36], [39].	
MDIST	Represents the average distance from the center to each AP and ML points [39], [44].	mediolateral, anteroposterior and average of mediolateral and anteroposterior distance (ML-dist, AP_dist, ML_AP_dist)
RDIST	Root Mean Square distance from the mean center [44].	
TOTEX	Total excursions is the total length of the path. Calculated as the sum of distances between consecutive points [44].	
MVELO	Mean velocity is the average velocity of the center path, calculated as the TOTEX divided by the time [39], [44].	
MFREQ	The mean frequency is the rotational frequency with a radius equal to the mean distance [36], [44].	
AREA_CC	The 95% confidence circle area is the area of a circle enclosing all points in the AP-ML plane with 95% confidence [36], [44].	average of mediolateral and anteroposterior distance (ML_AP_dist)
AREA_CE	The 95% confidence ellipse area is the area of an ellipse enclosing all points in the AP-ML plane with 95% confidence [36], [39], [44].	
AREA_SW	Sway area calculated as the area enclosing the acceleration path [36], [39], [44].	
FD	The fractal dimension indicates the degree to which a curve fills the enclosed metric space [36], [44].	
FD_CC	Fractal dimension based on the 95% confidence circle area [36], [44].	
FD_CE	Fractal dimension based on the 95% confidence ellipse area [36], [44].	

TABLE 6. List of voice features.

Feature acronym	Feature description	Signal (acronym)
amp	Average amplitude [45].	
shim	Absolute shimmer [15], [26], [45].	
shdb	Shimmer in logarithmic domain [45].	
apq3	3 point amplitude perturbation quotient in percentage [45].	
apq5	5 point amplitude perturbation quotient in percentage [45].	
fm	Frequency modulation [45].	
hnr_mean	Mean of the harmonic to noise ratio, which indicates the amount of noise [15], [26], [45].	
hnr_std	Standard deviation of the harmonic to noise ratio [45].	
rpde	Recurrence period density entropy. Characterizes the deviation from signal periodicity [15], [45].	
DFA	Detrended Fluctuation Analysis, which describes turbulent noise [15], [45].	
mean	Mean value [15], [45].	fundamental frequency (f0), amplitude (amp), Teager Kaiser Energy Operator of the fundamental frequency (tkeo_f0), open quotient (oq), glottis quotient open (gqo), glottis quotient closed (gqc)
sd	Standard deviation [15], [45].	
jitt	Absolute jitter [15], [45].	fundamental frequency (f0), period (T)
jitta	Relative or local jitter [45].	
rap	Relative average perturbation [45].	
ppq5	Perturbation quotient using 5 point (cycles) [45].	
range	Range [45].	fundamental frequency (f0)
tkeo_p25	25th percentile of the Teager-Kaiser Energy Operator [45].	
tkeo_p75	75th percentile of the Teager-Kaiser Energy Operator [45].	
ApEn	Pitch Period Entropy. Quantifies the impaired control of stable pitch during a sustained phonation [15], [45].	
p5	5th percentile [45].	Teager Kaiser Energy Operator of the fundamental frequency (tkeo_f0), open quotient (oq), glottis quotient open (gqo), glottis quotient closed (gqc)
p95	95th percentile [45].	
c_mean	Mean of the Mel Frequency Cepstral Coefficients (MFCCs) coefficients, log-energy of the signal and the first and second derivatives of the MFCCs [15], [26], [45].	log energy (log), 0th order cepstral coefficient (0th), 1-12th Mel Frequency Cepstral Coefficients (MFCC_(1-12), 1-14th deltas (d_(1-14)), 1-14th delta-delta (dd_(1-14))
c_std	Standard deviation of the MFCCs coefficients, log-energy of the signal and the first and second derivatives of the MFCCs [26], [45].	

Tapping recordings consist of the {x,y} screen pixel coordinates and timestamp for each tap on the screen. Both the inter-tapping interval (time) and the {x,y} inter-tap distance series were computed (Figure 1). Further details about preprocessing for each task can be found in Appendix A.

C. FEATURE EXTRACTION

A comprehensive search was conducted in PubMed (https://pubmed.ncbi.nlm.nih.gov/) with the following search terms ((Parkinson's disease) AND (walking OR gait OR balance OR voice OR tapping) AND (wearables OR smartphones)) to identify features commonly applied for each task and corresponding signals generated. Based on the results of this search, 423, 183, 124 and 43 features were identified and computed using Matlab R2017a from gait [40], [42], [43], balance [7], [36], [39], [44], voice [25], [26], [45] and tapping data [15], [32], [46], respectively (Table 4-Table 7).

D. MACHINE LEARNING ALGORITHMS

As a different ML algorithm may provide the best performance for a given task, we evaluated four commonly applied algorithms for differentiation between PD and HC:

1) Least Absolute Shrinkage and Selection Operator (LASSO) is a linear method commonly used to deal with high-dimensional data. LASSO applies a regularization process, where it penalizes the coefficients of the regression variables shrinking some of them to zero. During the feature selection process, those variables with non-zero coefficients are selected to be part of the model [47]. LASSO performs well when dealing with linearly separable data and avoiding overfitting.
2) Random Forest (RF) uses an ensemble of decision trees, where each individual tree outputs the classes. The predicted class is decided based on majority vote. Each tree is built based on a bootstrap training set that normally represents two thirds of the total cohort. The left out data is used to get an unbiased estimate of the classification error and get estimates of feature importance. RF runs efficiently in large datasets and deals very well with data with complicated relationships [48].
3) A Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel with Recursive Feature Elimination (SVM-RFE). An SVM is a linear method whose aim is to find the optimal hyperplane that separates

TABLE 7. List of tapping features.

Feature acronym	Feature description	Signal (acronym)
nTaps	Number of taps [46], [65].	
buttonNone	Frequency of tapping outside the button [46], [65].	
corXY	Correlation between X and Y touchscreen coordinates [46], [65].	
mean	Mean value of the observations [15], [46], [65].	Intertap interval (TapInter), Leftdrift (DriftLeft), Right drift (DriftRight)
min	Minimum value of the observations [15], [46], [65].	
max	Maximum value of the observations [15], [46], [65].	
median	Median value of the observations [15], [46], [65].	
mad	Median absolute deviation [46], [65].	
sd	Standard deviation [15], [46], [65].	
range	Range of the observations [15], [46], [65].	
iqr	Interquartile range [46], [65].	
cov	Coefficient of variation [15], [46], [65].	
skew	Skewness [46], [65].	
kur	Kurtosis [46], [65].	
tkeo	Teager-Kaiser Energy Operator. Measures energy variation [15], [46], [65].	Intertap interval (TapInter)
dfa	Detrended Fluctuation Analysis. Measures changes in the signal [15], [46], [65].	
ar1	Coefficient of an autoregressive model at lag 1. Indicates associations between intertap intervals [15], [46], [65].	
ar2	Coefficient of an autoregressive model at lag 2. Indicates associations between intertap intervals [15], [46], [65].	
fatigue10	Increase in the mean intertap interval from the first 10% to the last 10% taps [15], [46], [65].	
fatigue25	Increase in the mean intertap interval from the first 25% to the last 25% taps [15], [46], [65].	
fatigue50	Increase in the mean intertap interval from the first 50% to the last 50 % taps [15], [46], [65].	

between classes. When data is linearly non-separable, it may be transformed to a higher dimensional space using a non-linear transformation function that spreads the data apart such that a linear hyperplane can be found in that space. Here, we used a radial basis kernel function. RFE is a feature selection method that ranks features according to importance, improving both efficiency and accuracy of the classification model. This model is known to remove effectively non-relevant features and achieve high classification performance [49].

4) Relevance Vector Machine (RVM), which follows the same principles of SVM but provides probabilistic classification. The Bayesian formulation prevents from tuning the hyper-parameters of the SVM. Nonetheless, RVMs use an expectation maximization (EM)-like learning that can lead to local minima unlike the standard sequential optimization (SMO)-based algorithms used by SVMs, that guarantee to find a global optima [50].

E. FRAMEWORK

The following six experiments were performed to address the questions on the impact of age, sex and comorbidities that may influence task performance on the classification accuracy for each task and on the combination of all tasks for differentiation between PD and HC (Table 2):

1) Experiment 1 (E1: all) includes all subjects only restricting the age range (35-75 years old).
2) Experiment 2 (E2: matched) includes subjects after an age and sex matching between PD and HC, where we strictly match one HC for each PD subject with the same age and where possible with the same sex.
3) Experiment 3 (E3: no comorbidities, matched) excludes all comorbidities that may affect task performance (see Appendix A) and strictly matches for age and where possible sex on the remaining subjects.
4) Experiments 4-6 (E4-6): Three additional experiments assess if controlling for age and sex impacts the results. These experiments exclude comorbidities, match for age and sex and control for age and/or sex applying multiple regression. For this, age and gender were included as covariates in a multiple regressions using the features for each modality as dependent variables. The estimated beta coefficients for each covariate were used to regress out the estimated effects of age and sex on the respective feature. The resulting residuals for each feature were used for subsequent classification. Experiment 4 (E4): no comorbidities, matched, controlled for age; Experiment 5 (E5): no comorbidities, matched, controlled for sex; Experiment 6 (E6): no comorbidities, matched, controlled for age and sex.

As the performance obtained after removing comorbidities and matching for age and sex (E3) provides a relatively unbiased estimate for differentiation between PD

TABLE 8. Medication status at the time of performing the tasks.

			PD/Total	Before	After	Another	No Med	Empty
E1	Gait	All	653/2711	154	166	259	63	11
		CV	436/2711	99	120	169	39	9
		Holdout	217/903	55	46	90	24	2
	Balance	All	655/2747	156	166	257	64	12
		CV	437/1832	103	99	186	41	8
		Holdout	218/915	53	67	71	23	4
	Voice	All	965/4799	222	229	396	94	24
		CV	644/3200	140	159	265	62	18
		Holdout	321/1599	82	70	131	32	6
	Tapping	All	1054/6221	237	237	446	106	28
		CV	703/4148	156	157	299	72	19
		Holdout	351/2073	81	80	147	34	9
E2	Gait	All	373/746	91	101	135	39	7
		CV	249/498	62	70	86	28	3
		Holdout	124/248	29	31	49	11	4
	Balance	All	376/752	99	96	139	36	6
		CV	251/502	68	60	97	21	5
		Holdout	125/250	31	36	42	15	1
	Voice	All	534/1068	135	131	205	48	15
		CV	356/712	94	89	127	34	12
		Holdout	178/356	41	42	78	14	3
	Tapping	All	608/1216	134	135	262	59	18
		CV	406/812	83	92	182	39	10
		Holdout	202/404	51	43	80	20	8
E3	Gait	All	317/634	82	84	117	28	6
		CV	212/424	54	53	82	19	4
		Holdout	105/210	28	31	35	9	2
	Balance	All	320/640	76	89	116	32	7
		CV	214/428	48	53	87	22	4
		Holdout	106/212	28	36	29	10	3
	Voice	All	446/892	112	103	190	34	7
		CV	298/596	75	71	125	21	6
		Holdout	148/296	37	32	65	13	9
	Tapping	All	507/1014	124	112	211	44	16
		CV	338/676	80	70	147	30	11
		Holdout	169/338	44	42	64	14	5
E4-6	Gait	All	317/634	82	84	117	28	6
		CV	212/424	54	53	82	19	4
		Holdout	105/210	28	31	35	9	2
	Balance	All	320/640	76	89	116	32	7
		CV	214/428	48	53	87	22	4
		Holdout	106/212	28	36	29	10	3
	Voice	All	446/892	112	103	190	34	7
		CV	298/596	75	71	125	21	6
		Holdout	148/296	37	32	65	13	1
	Tapping	All	507/1014	124	112	211	44	16
		CV	338/676	80	70	147	30	11
		Holdout	169/338	44	42	64	14	5

Before - "Immediately before taking their medication", After - "After taking their medication (when they are feeling at their best)", Another - "At some other time", No Med - "I don't take Parkinson's medication", Empty - question unanswered

and HC, these results were used for selection of the best performing ML algorithm for each task and interpretation of the main outcomes throughout this work. Demographic and clinical information for each experiment are provided in Table 1.

Additionally, to compare the performance of our analyses to those in the literature, we performed an analysis including all data without restricting age range (Table 15) and an analysis including all data and both age and sex as features.

F. MODEL PERFORMANCE

Data leakage occurs when information of the holdout test set leaks into the dataset used to build the model, leading to incorrect or overoptimistic predictions. Therefore, in every experiment and task, data was initially split into 2/3 of data to build the predictive model and 1/3 of holdout data to validate this model. To build the model, we performed 1000 repetitions of 10-fold cross-validation (CV) in the 2/3 of the data for each classifier to avoid data leakage and increase robustness. The parameter Lambda of the LASSO model was set to 1 and

TABLE 9. Cross-validation classification performances for each of the tasks (gait, balance, voice, tapping and multimodal features) for four different classifiers.

	% (95% CI)	Gait	Balance	Voice	Tapping	Multimodal
LASSO	Balanced Accuracy	50.14 (48; 52.12)	53.49 (51.4; 55.49)	55.38 (52.69; 57.72)	64.29 (63.02; 65.61)	48.35 (44.85; 51.96)
	Sensitivity	71.07 (65.33; 74.76)	47.39 (44.63; 50.47)	54.55 (48.66; 59.06)	64.89 (63.02; 66.72)	53.37 (37.75; 68.14)
	Specificity	29.22 (25; 33.96)	59.58 (56.31; 62.38)	56.2 (52.85; 60.4)	63.7 (61.54; 65.83)	43.33 (29.66; 58.33)
	PPV	50.1 (48.55; 51.51)	53.97 (51.6; 56.23)	55.46 (52.82; 57.69)	64.14 (62.82; 65.58)	48.45 (45.16; 51.76)
	NPV	50.28 (46.73; 53.84)	53.11 (51.25; 54.99)	55.32 (52.59; 57.81)	64.47 (63.11; 65.92)	48.14 (43.96; 52.37)
	AUC	49.26 (45.95; 52.36)	54.75 (51.84; 57.23)	57.37 (54.83; 59.27)	71.71 (70.2; 73.05)	46.39 (42.05; 50.73)
RF	Balanced Accuracy	54.24 (51.42; 56.84)	59.95 (57.59; 62.27)	56.19 (53.27; 59.4)	65.36 (64.05; 66.64)	69.59 (66.91; 72.43)
	Sensitivity	52.43 (48.59; 55.9)	60.28 (57.01; 63.79)	54.84 (50.84; 59.06)	63.05 (61.24; 64.79)	69.57 (65.69; 73.53)
	Specificity	56.04 (52.36; 59.67)	59.63 (56.31; 62.85)	57.53 (53.69; 61.91)	67.67 (65.68; 69.68)	69.61 (65.69; 73.53)
	PPV	54.4 (51.47; 57.09)	59.9 (57.54; 62.23)	56.37 (53.33; 59.6)	66.11 (64.67; 67.52)	69.61 (66.67; 72.86)
	NPV	54.09 (51.38; 56.67)	60.03 (57.71; 62.51)	56.03 (53.19; 59.12)	64.68 (63.32; 65.95)	69.6 (66.67; 72.8)
	AUC	55.75 (51.73; 59.28)	64.62 (61.72; 67.51)	59.32 (56.71; 62.11)	71.68 (70.23; 73.04)	76.01 (73.79; 78.19)
RVM	Balanced Accuracy	55.42 (53.18; 57.67)	57.07 (54.91; 59)	56.7 (54.36; 58.89)	67.89 (67.01; 68.86)	67.02 (63.48; 70.59)
	Sensitivity	55.54 (52.36; 58.73)	54.05 (51.4; 57.01)	57.74 (54.7; 60.74)	64.34 (63.17; 65.39)	65.43 (60.29; 70.1)
	Specificity	55.31 (52.12; 58.26)	60.09 (57.24; 62.85)	55.66 (52.35; 58.89)	71.43 (70.12; 72.78)	68.6 (63.73; 73.78)
	PPV	55.42 (53.22; 57.62)	57.54 (55.23; 59.69)	56.57 (54.26; 58.63)	69.25 (68.14; 70.42)	67.6 (63.9; 71.69)
	NPV	55.44 (53.14; 57.72)	56.68 (54.59; 58.52)	56.85 (54.47; 59.03)	66.7 (65.8; 67.55)	66.51 (62.74; 70.07)
	AUC	56.39 (53.55; 59.44)	58.18 (55.27; 60.9)	60.67 (58.87; 62.4)	74.17 (73.27; 75.01)	73.32 (70.19; 76.54)
SVM-RFE	Balanced Accuracy	56.5 (53.3; 59.7)	56.19 (52.57; 59.6)	56 (53.2; 58.6)	65.83 (63.76; 67.97)	68.36 (65.47; 71.41)
	Sensitivity	56.56 (53.38; 59.9)	56.17 (52.53; 59.63)	55.56 (53.02; 58.16)	67.61 (65; 70.26)	69.23 (66; 72.63)
	Specificity	56.38 (53.14; 59.6)	56.21 (52.52; 59.95)	56.44 (53.44; 59.21)	64.05 (62.24; 65.94)	67.48 (64.52; 70.79)
	PPV	55.81 (50.94; 60.85)	56.3 (50; 62.62)	59.63 (55.37; 63.59)	60.02 (57.1; 63.02)	65.92 (61.77; 70.59)
	NPV	57.11 (52.36; 61.79)	56.06 (50.47; 62.15)	52.29 (47.65; 56.54)	71.23 (67.46; 74.56)	70.69 (66.67; 74.51)
	AUC	58.49 (55.55; 61.57)	58.4 (54.35; 62.1)	59.36 (56.3; 61.85)	71.95 (70.01; 73.75)	74.74 (72.29; 77.23)

AUC – Area Under the Curve, NPV – Negative Predictive Values, PPV – Positive Predictive Values, RF – Random Forest, RVM – Relevance Vector Machine, SVM-RFE – Support Vector Machine-Recursive Feature Elimination

the number of trees for RF to 100. A nested cross-validation was implemented to tune the parameters of the SVM-RFE classifier. The procedure consists of an inner CV to select the best parameters of the model following a grid search for the regularization constant (C) ranging from 2^{-7} to 2^{7} and for gamma (γ) ranging from 2^{-4} to 2^{4} for the SVM. Then, the outer loop is used to assess the model selected in the inner CV. Extensive parameter optimization was applied only on SVM-RFE classifier, given that the other algorithms have already embedded optimization and that 1000 repetitions of 10-fold cross-validation and multiple experiments would have taken based on the estimated from a single run each at least several months on the high-throughput cluster available to us. For each model, we report the following measures of predictive performance: balanced accuracy (BA), sensitivity, specificity, positive (PPV) and negative predictive value (NPV), mean receiver operating characteristic (ROC) curves with 95% confidence intervals and area under the curve (AUC). Comparisons between models are based on BA.

Once the best predictive model with the highest cross-validation BA was identified using the CV dataset, it was validated using the holdout dataset, reporting the

TABLE 10. Classification performance for the gait task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	56.56 (54.34; 58.85)	57.11	52.63 (51.12; 54.21)	50.33 (47.19; 53.62)	54.84	50.03 (47.13; 52.58)	56.5 (53.3; 59.7)	55.71	49.92 (46.63; 53.36)
Sensitivity (%)	52.63 (49.52; 55.89)	39.41	31.41 (26.11; 36.45)	50.34 (47.15; 53.75)	50	45.21 (38.71; 50.81)	56.56 (53.38; 59.9)	56.19	50.33 (43.81; 57.14)
Specificity (%)	60.49 (59.12; 62.02)	74.81	73.86 (71.97; 76.13)	50.33 (47.15; 53.54)	59.68	54.85 (54.35; 55.56)	56.38 (53.14; 59.6)	55.24	49.52 (49.45; 49.58)
PPV (%)	38.22 (34.64; 42.02)	54.79	50 (50; 50)	49.68 (45.38; 54.22)	55.36	50 (50; 50)	55.81 (50.94; 60.85)	55.66	50 (50; 50)
NPV (%)	73.32 (70.1; 76.19)	61.44	56.35 (52.98; 59.56)	50.98 (46.19; 55.82)	54.41	50.05 (44.12; 55.15)	57.11 (52.36; 61.79)	55.77	49.85 (43.27; 56.73)
AUC (%)	59.45 (57.41; 61.38)	59.88	50.03 (44.42; 55.17)	50.98 (47.84; 54.13)	56.5	50.07 (42.9; 56.99)	58.49 (55.55; 61.57)	55.85	49.95 (41.48; 57.74)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	56.48 (53.3; 59.46)	54.76	49.97 (46.7; 53.31)	56.4 (53.07; 59.68)	56.19	50 (46.43; 53.57)	56.65 (53.3; 59.91)	53.81	50.05 (46.85; 53.19)
Sensitivity (%)	56.56 (53.24; 59.66)	54.29	49.46 (42.86; 56.19)	56.46 (53.08; 59.95)	56.19	49.99 (42.86; 57.14)	56.71 (53.32; 60.05)	51.43	47.71 (40.95; 54.29)
Specificity (%)	56.39 (53.16; 59.43)	55.24	50.48 (50.42; 50.55)	56.35 (52.93; 59.62)	56.19	50 (50; 50)	56.58 (53.24; 59.91)	56.19	52.39 (52.1; 52.75)
PPV (%)	55.82 (51.42; 60.38)	54.81	50 (50; 50)	56 (50.94; 60.38)	56.19	50 (50; 50)	56.12 (51.42; 60.38)	54	50 (50; 50)
NPV (%)	57.12 (52.36; 61.79)	54.72	49.94 (43.4; 56.6)	56.82 (52.36; 61.32)	56.19	49.99 (42.86; 57.14)	57.15 (52.83; 61.79)	53.64	50.08 (43.64; 56.36)
AUC (%)	58.51 (55.51; 61.3)	56.25	49.99 (41.91; 58.25)	58.2 (54.83; 61.24)	56.1	50.02 (42.65; 58.05)	58.33 (55.25; 61.52)	55.99	50.07 (42.16; 57.99)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

aforementioned performance metrics. In addition, to test whether the BA of the predictive model is higher than chance level (0.5 for binary classification), we ran 1000 permutations randomly permuting the predicted classes, reporting BA at 95% confidence intervals.

III. RESULTS

A. CLASSIFIER SELECTION AND RESULTS FOR THE CV DATASET

Four different classifiers (random forest: RF, Least Absolute Shrinkage and Selection Operator: LASSO, support vector machine: SVM, relevance vector machine: RVM-RFE) were applied to each of the four tasks and their combination during the main experiment (E3: no comorbidities, matched for age and sex). Table 9 provides detailed information on the classification performance for each ML algorithm and each task. The ROC curves and corresponding AUC values for the four classifiers for each of the tasks during the cross-validation (CV) step are displayed in Figure 2A. RF, RVM and SVM-RFE performed similarly across all tasks, whereas LASSO was the classifier performing the poorest. Best performance was achieved on the combination of all tasks using RF (balanced accuracy (BA)): 69.6%), followed by tapping using RVM (BA: 67.9%), balance using RF (BA: 60%), voice using RVM (BA: 56.7%) and gait using SVM-RFE (BA: 56.5%).

B. COMPARISON OF EXPERIMENTS IN THE CROSS-VALIDATION SETTING

ML algorithms performing best for each task in the main experiment (E3: no comorbidities, matched for age and sex) were applied to corresponding task data of the other five

TABLE 11. Classification performance for the balance task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	61.82 (60.41; 63.4)	64.73	50.15 (45.68; 54.38)	60.42 (58.57; 62.35)	58	49.81 (43.2; 56)	59.95 (57.59; 62.27)	59.91	49.93 (43.4; 56.6)
Sensitivity (%)	44.96 (42.89; 47.43)	50.25	33.72 (28.43; 38.73)	57.72 (54.78; 60.36)	58	49.81 (43.2; 56)	60.28 (57.01; 63.79)	65.57	55.59 (49.06; 62.26)
Specificity (%)	78.68 (76.96; 80.41)	79.21	66.59 (62.55; 70.41)	63.11 (60.36; 66.14)	58	49.81 (44; 56)	59.63 (56.31; 62.85)	54.25	44.27 (37.74; 50.94)
PPV (%)	61.63 (59.37; 64.05)	64.87	43.54 (36.94; 49.69)	61.02 (58.98; 63.14)	58	49.81 (43.31; 56.1)	59.9 (57.54; 62.23)	58.9	49.94 (44.07; 55.93)
NPV (%)	65.26 (64.21; 66.46)	67.57	56.8 (53.5; 59.94)	59.89 (58.1; 61.72)	58	49.81 (43.31; 56.1)	60.03 (57.71; 62.51)	61.17	49.92 (42.71; 57.61)
AUC (%)	67.02 (65.38; 68.73)	70.45	50.15 (44.85; 51.19)	65.15 (62.85; 67.43)	61.02	49.86 (42.16; 57.21)	64.61 (61.72; 67.51)	63.09	49.99 (42.18; 57.44)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	60.58 (57.24; 63.79)	61.32	50.11 (43.4; 56.6)	60.12 (56.54; 63.32)	61.32	50.08 (43.4; 57.55)	60.24 (57.01; 63.55)	59.91	50.2 (42.93; 57.08)
Sensitivity (%)	60.67 (55.84; 65.42)	66.04	54.82 (48.11; 61.32)	59.72 (54.67; 64.02)	61.32	50.08 (43.4; 57.55)	60.06 (55.14; 64.95)	58.49	48.78 (41.51; 55.66)
Specificity (%)	60.49 (56.08; 64.95)	56.6	45.39 (38.68; 51.89)	60.5 (56.08; 64.95)	61.32	50.08 (43.4; 57.55)	60.42 (56.08; 64.95)	61.32	51.61 (44.34; 58.49)
PPV (%)	60.57 (57.11; 63.92)	60.34	50.1 (43.97; 56.03)	60.2 (56.63; 63.5)	61.32	50.08 (43.4; 57.55)	60.29 (57.05; 63.57)	60.19	50.2 (42.72; 57.28)
NPV (%)	60.61 (57.11; 63.98)	62.5	50.12 (42.71; 57.29)	60.04 (56.49; 63.33)	61.32	50.08 (43.4; 57.55)	60.22 (56.81; 63.7)	59.63	50.19 (43.12; 56.88)
AUC (%)	65.45 (62.32; 68.41)	63.59	50.08 (41.98; 56.94)	64.96 (61.84; 67.77)	62.76	50.06 (42.24; 58.54)	65.24 (62.32; 68.02)	62.65	50.28 (42.04; 58.05)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

experiments (E1: all subjects, E2: matched for age and sex, E4-6: same as E3 but additionally regressing out the effects of age and/or sex). Classification performance for each task and experiment during the CV and over holdout sets is summarized in Table 3 and Table 10-Table 14. BA distributions for each experiment and task during the CV are displayed in Figure 2B.

In the CV, E1 (all data) resulted in the highest but modest BA for all tasks (gait: 56.6%; balance: 61.8%; voice: 62.5%; tapping: 74.8; multimodal combining all four tasks: 73.9%). Removal of comorbidities in E3 had a marginal effect on BA as compared to E2 (matched for age and sex) with increased BA for gait (E2: 50.3%; E3: 56.5%), voice (E2: 53.9%; E3: 56.7%) and tapping (E2: 66.8%; E3: 67.9%) but lower BA for balance (E2: 60.4%; E3: 60.0%). After additionally regressing out the effects of age and/or sex (E4-E6) the change in the BA was negligible for all tasks (< 1%) except for voice when regressing out sex (E3: 56.7%; E5: 60%) and both age and sex (E3: 56.7%; E6: 59.2%) (Table 3, Tables 10–14).

Analyses including all data without trimming for age range led to the highest accuracy of 74.4% using tapping data, followed by 72.7% for the multimodal case and 58%, 52.9% and 51% for balance, voice and gait data respectively. In all cases specificity was close to 100% whereas sensitivity was exceedingly low (Table 16-Table 20). When including both age and sex as additional features, accuracy increased to 80.8% for tapping data, 75.3% for the multimodal case and 73.1%, 69% and 57% for voice, balance and gait data respectively with high specificities and low sensitivities.

TABLE 12. Classification performance for the voice task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	62.49 (61.33; 63.62)	60.41	49.99 (46.46; 53.5)	53.94 (51.48; 56.18)	59.83	50.05 (44.66; 55.34)	56.7 (54.36; 58.89)	53.04	49.98 (43.58; 55.74)
Sensitivity (%)	46.54 (44.63; 48.49)	44.44	32.25 (28.11; 36.36)	50.43 (47.05; 53.65)	56.74	46.96 (41.57; 52.25)	57.74 (54.7; 60.74)	52.03	48.97 (42.57; 54.73)
Specificity (%)	78.43 (77.15; 79.71)	76.37	67.73 (64.8; 70.64)	57.45 (53.93; 60.67)	62.92	53.14 (47.75; 58.43)	55.66 (52.35; 58.89)	54.05	50.99 (44.6; 56.76)
PPV (%)	60.55 (58.79; 62.19)	57.14	41.47 (36.15; 46.75)	54.24 (51.62; 56.67)	60.48	50.05 (44.31; 55.69)	56.57 (54.26; 58.63)	53.1	49.98 (43.45; 55.86)
NPV (%)	67.35 (66.5; 68.19)	65.98	58.51 (55.98; 61.03)	53.68 (51.36; 55.76)	59.26	50.04 (44.97; 55.03)	56.85 (54.47; 50.03)	52.98	49.98 (43.71; 55.63)
AUC (%)	68.99 (68.14; 69.79)	66.95	49.93 (45.49; 54.03)	55.5 (53.36; 57.6)	62.48	50.14 (44.18; 56.09)	60.67 (58.87; 62.4)	54.99	50.01 (43.02; 56.65)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	56.85 54.7; 59.06)	58.11	49.98 (44.6; 56.08)	59.99 (57.72; 62.08)	60.14	49.82 (43.92; 55.41)	59.15 (57.05; 61.24)	59.12	50.15 (43.58; 55.74)
Sensitivity (%)	56.64 (53.69; 59.56)	53.38	45.25 (39.87; 51.35)	59.55 (56.38; 62.42)	53.38	43.06 (37.16; 48.65)	58.83 (56.04; 61.41)	58.78	49.82 (43.24; 55.41)
Specificity (%)	57.06 (54.03; 60.07)	62.84	54.71 (49.32; 60.81)	60.43 (57.38; 63.42)	66.89	56.58 (50.68; 62.16)	59.46 (56.71; 62.42)	59.46	50.49 (43.92; 56.08)
PPV (%)	56.88 (54.59; 59.25)	58.96	49.98 (44.03; 56.72)	60.08 (57.76; 62.2)	61.72	49.79 (42.97; 56.25)	59.21 (57; 61.36)	59.18	50.15 (43.54; 55.78)
NPV (%)	56.82 (54.64; 59.06)	57.41	49.98 (45.06; 55.56)	59.91 (57.69; 62)	58.93	49.84 (44.64; 54.76)	59.09 (57.02; 61.23)	59.06	50.15 (43.62; 55.71)
AUC (%)	58.44 (56.37; 60.38)	59.13	50.03 (43.7; 56.47)	63.3 (61.25; 65.05)	61.64	49.85 (43.28; 56.58)	61.72 (59.65; 63.53)	63.26	50.25 (43.32; 57.15)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

C. RESULTS FOR THE HOLDOUT DATASET

Best performing classifiers trained on the 2/3 of the initial dataset used for cross-validation were applied to the 1/3 holdout dataset. Results for the holdout dataset were highly similar to the CV results (Table 3, Tables 10–14). All results are summarized in Figure 3 and Table 3. The multimodal combination of all tasks resulted in the best performance for differentiation of PD and HC in the holdout cohort (BA: 73.5%) followed by the tapping features (67.2%). Voice features achieved the lowest BA of 53% followed by gait (55.7%) and balance (59.9%) features (Table 3). For the base experiment E3, the difference in BA between CV and holdout sets was less than 1% for all tasks except for a 3.7% reduction in BA for voice data and a 3.9% increase for the multimodal feature combination. Exclusion of comorbidities resulted in only minor changes for gait, balance and tapping (<2%) with a 6.8% drop only observed using voice data and a 3.5% increase for the multimodal case. BA performance for all tasks increased by 1.4% (gait) to 7.4% (voice) for all tasks when using the dataset only restricting the age range (E1) as compared to E3. No systematic effects of additionally controlling for age and/or sex prior to classification (E4-E6) were observed with BA changes being small and inconsistent across tasks and experiments.

Analyses including all data without trimming for age range reached the highest accuracy in the holdout set of 73.3% using multimodal features, followed by 71.1% for the tapping task and 55.8%, 52.6% and 51.6% for balance, voice and gait data respectively (Table 16-Table 20). When including both age and sex as additional features, accuracy in the holdout data

TABLE 13. Classification performance for the tapping task.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	74.81 (74.41; 75.23)	72.9	49.99 (46.98; 52.9)	66.78 (65.95; 67.55)	66.83	49.86 (45.05; 54.95)	67.89 (67.01; 68.86)	67.16	50.09 (44.08; 55.92)
Sensitivity (%)	61.09 (60.36; 61.75)	57.59	28.86 (25.08; 32.51)	63.77 (62.56; 64.9)	59.9	42.93 (38.12; 48.02)	64.34 (63.17; 65.39)	68.05	50.98 (44.97; 56.81)
Specificity (%)	88.52 (88.13; 88.91)	88.21	71.13 (68.88; 73.3)	69.8 (68.6; 70.94)	73.76	56.8 (51.98; 61.88)	71.43 (70.12; 72.78)	66.27	49.21 (43.2; 55.03)
PPV (%)	76.01 (75.38; 76.71)	74.4	37.29 (32.4; 42)	67.86 (66.88; 68.74)	69.54	49.84 (44.25; 55.75)	69.25 (68.14; 70.42)	66.86	50.09 (44.19; 55.81)
NPV (%)	79.26 (78.95; 79.56)	77.76	62.7 (60.71; 64.61)	65.83 (65.01; 66.55)	64.78	49.88 (45.65; 54.35)	66.7 (65.8; 67.55)	67.47	50.1 (43.98; 56.02)
AUC (%)	83.48 (83.14; 83.77)	84.42	50.01 (46.37; 53.95)	73.36 (72.56; 73.99)	74.97	49.88 (44.38; 55.73)	74.17 (73.27; 75.01)	77.51	50.13 (43.63; 56.88)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	68.8 (67.9; 69.75)	66.86	49.98 (44.38; 55.03)	68.66 (67.6; 69.68)	68.93	50.23 (45.27; 55.33)	68.8 (67.75; 69.75)	68.05	49.97 (44.97; 55.62)
Sensitivity (%)	65.45 (64.35; 66.86)	66.27	49.39 (43.79; 54.44)	65.86 (64.65; 67.01)	68.64	49.94 (44.97; 55.03)	65.6 (64.35; 66.86)	67.46	49.38 (44.38; 55.03)
Specificity (%)	72.16 (70.71; 73.52)	67.46	50.57 (44.97; 55.62)	71.46 (69.97; 72.93)	69.23	50.53 (45.56; 55.62)	72 (70.56; 73.37)	68.64	50.56 (45.56; 56.21)
PPV (%)	70.16 (69.03; 71.31)	67.07	49.98 (44.31; 55.09)	69.78 (68.53; 71.02)	69.05	50.23 (45.24; 55.36)	70.09 (68.85; 71.23)	68.26	49.97 (44.91; 55.69)
NPV (%)	67.62 (66.8; 68.63)	66.67	49.98 (44.44; 54.97)	67.67 (66.62; 68.6)	68.82	50.23 (45.29; 55.29)	67.67 (66.71; 68.58)	67.84	49.97 (45.03; 55.56)
AUC (%)	74.88 (73.97; 75.7)	78.05	50 (44.41; 56.22)	74.41 (73.5; 75.22)	77.95	50.15 (44.24; 56.09)	74.82 (73.94; 75.58)	78.38	49.99 (44.11; 56.22)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

raised to 78.9% for tapping data, 75.9% for the multimodal case and 74.6%, 66% and 58.3% for voice, balance and gait data respectively with very high specificities and very low sensitivities.

D. PREDICTIVE FEATURES

Best performance during CV for the main experiment E3 was achieved using the multimodal set of features. Figure 3 shows the scaled average absolute feature weights for RVM and SVM-RFE and the scaled average importance scores for RF, calculated with the out-of-bag (OOB) permuted predictor delta error across 1000 repetitions during the CV. Features with the highest importance scores belong to the tapping task followed by the balance task. Tapping features with the highest importance scores comprised the range of intertap interval (100), maximum value of the intertap interval (99.8) and Teager-Kaiser energy operator of the intertap interval (83.2). Balance features with highest importance scores were the power ratio between high (3.5-15 Hz) and low (0.15-3.5 Hz) frequency for anteroposterior acceleration (31.5) and energy in the medium frequency band for medio-lateral acceleration (25.3). Gait and voice tasks had the least contributions in terms of importance scores.

IV. DISCUSSION

Here, we systematically evaluated the ability of four commonly applied DB tasks to differentiate between PD and HC in a self-administered remote setting. Our findings indicate that, depending on the constellation, not accounting for confounds in PD digital biomarker task data may lead to under- but also over-optimistic results.

TABLE 14. Classification performance for the multimodal features.

	Experiment 1			Experiment 2			Experiment 3		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	73.85 (72.41; 75.47)	76.88	50.04 (45.13; 54.88)	69.39 (67.01; 71.89)	70	50.08 (44.17; 56.67)	69.59 (66.91; 72.43)	73.53	50.04 (43.14; 56.86)
Sensitivity (%)	67.09 (64.57; 69.6)	68.34	38.61 (33.17; 43.97)	66.67 (63.07; 70.12)	65	45.08 (39.17; 51.67)	69.57 (65.69; 73.53)	77.45	53.96 (47.06; 60.78)
Specificity (%)	80.61 (78.59; 82.53)	85.43	61.47 (57.09; 65.79)	72.12 (68.47; 75.52)	75	55.08 (49.17; 61.67)	69.61 (65.69; 73.53)	69.61	46.12 (39.22; 52.94)
PPV (%)	73.57 (71.52; 75.63)	79.07	44.67 (38.37; 50.87)	70.53 (67.71; 73.18)	72.22	50.09 (43.52; 57.41)	69.61 (66.67; 72.86)	71.82	50.04 (43.64; 56.36)
NPV (%)	75.29 (73.92; 76.79)	77.01	55.41 (51.46; 59.31)	68.4 (65.97; 70.97)	68.18	50.07 (44.7; 56.06)	69.6 (66.67; 72.8)	75.53	50.04 (42.55; 57.45)
AUC (%)	82.25 (81.39; 83.15)	85.63	49.96 (44.69; 55.2)	76.01 (74.21; 77.94)	78.81	50.02 (42.89; 57.43)	76.01 (73.79; 78.19)	80.49	50.08 (42.04; 58.23)

	Experiment 4			Experiment 5			Experiment 6		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	69.24 (66.18; 71.81)	73.04	50.18 (43.63; 56.37)	68.03 (64.95; 70.83)	69.12	50.02 (43.63; 57.35)	69.86 (67.16; 72.79)	70.59	50.01 (43.14; 56.86)
Sensitivity (%)	67.88 (63.73; 71.57)	70.59	47.72 (41.18; 53.92)	65.77 (61.77; 69.61)	63.73	44.62 (38.24; 51.96)	65.98 (62.26; 69.61)	61.76	41.18 (34.31; 48.04)
Specificity (%)	70.6 (66.18; 74.51)	75.49	52.62 (46.08; 58.82)	70.3 (65.69; 74.51)	74.51	55.41 (49.02; 62.75)	73.74 (69.61; 77.94)	79.41	58.83 (51.96; 65.69)
PPV (%)	69.8 (66.5; 72.82)	74.23	50.18 (43.3; 56.7)	68.91 (65.37; 72.21)	71.43	50.02 (42.86; 58.24)	71.56 (68.27; 75.28)	75	50.01 (41.67; 58.33)
NPV (%)	68.75 (65.85; 71.71)	71.96	50.16 (43.93; 56.08)	67.26 (64.39; 70.19)	67.26	50.02 (44.25; 56.64)	68.43 (65.84; 71.27)	67.5	50 (44.17; 55.83)
AUC (%)	74.78 (72.53; 76.97)	80.27	50.04 (42.59; 58.06)	73.49 (71.17; 75.83)	76.68	50.04 (41.42; 58.59)	75.83 (73.72; 78.15)	77.58	50 (42.26; 57.92)

AUC – Area Under the Curve, CV – Cross-Validation, NPV – Negative Predictive Values, PPV – Positive Predictive Values

TABLE 15. Demographics for PD and HC subjects including all data.

	Gait		Balance		Voice		Tapping		Multimodal	
	PD	HC	PD	HC	PD	HC	PD	HC	PD	HC
N	653	2058	655	2092	965	3834	1054	5167	640	1940
Male/ female	436 ± 217	1678 ± 38	438 ± 217	438 ± 385	629 ± 336	3108 ± 726	697 ± 357	4190 ± 977	427 ± 213	1582 ± 358
Age (mean±sd)	60.45 ± 10.72	34.77 ± 14.29	60.44 ± 10.71	34.76 ± 14.23	60.33 ± 11.04	32.77 ± 13.12	59.77 ± 11.42	32.18 ± 12. 53	60.51 ± 10.69	34.84 ± 14.41

A. IDENTIFICATION OF PARKINSON'S DISEASE

Out of the four evaluated machine learning algorithms, similar performance was achieved for all classifiers except LASSO which showed the poorest performance. Whereas some previous studies using the mPower dataset selected different algorithms according to tasks [25], [26], others simply applied a single classifier [27], [29]. No single classifier performed best for all four tasks in our study. This is in line with previous research showing that the selection of the classifier depends mainly on the type and complexity of the data [51], [52]. For instance, RF, RVM and Gaussian SVM are non-linear algorithms, offering more flexibility regarding the

TABLE 16. Classification performance for the gait task.

	Additional experiment: All data			Additional experiment: All data + age + gender		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	50.95 (50.29; 51.63)	51.55	50.01 (49.13; 50.95)	57.2 (55.81; 58.68)	58.25	50.01 (48.54; 51.57)
Sensitivity (%)	2.74 (1.61; 4.13)	3.69	1.35 (0; 2.77)	16.19 (13.3; 19.27)	17.51	5 (2.77; 7.37)
Specificity (%)	99.16 (98.8; 99.6)	99.42	98.68 (98.25; 99.13)	98.21 (97.63; 98.76)	98.98	95.02 (94.32; 95.77)
PPV (%)	50.8 (33.33; 66.67)	66.67	24.34 (0; 50)	74.26 (68.28; 80.66)	84.44	24.12 (13.33; 35.56)
NPV (%)	76.24 (75.99; 76.49)	76.54	75.97 (75.65; 76.32)	78.67 (78.11; 79.26)	79.14	75.97 (75.41; 76.57)
AUC (%)	66.08 (64.67; 67.52)	67.81	49.93 (45.49; 54.46)	86.49 (85.66; 87.32)	88.43	50.01 (45.45; 54.51)

TABLE 17. Classification performance for the balance task.

	Additional experiment: All data			Additional experiment: All data + age + gender		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	58.04(57.02; 59.1)	55.75	49.99 (48.22; 51.53)	69.03 (67.41; 70.53)	65.97	50.07 (47.6; 52.41)
Sensitivity (%)	19.12 (17.16; 21.28)	14.22	5.46 (2.75; 7.8)	41.9 (38.9; 45.08)	36.24	12.02 (8.26; 15.6)
Specificity (%)	96.96 (96.34; 97.56)	97.27	94.53 (93.69; 95.27)	96.16 (95.56; 96.77)	95.7	88.12 (86.94; 89.24)
PPV (%)	66.35 (61.03; 71.31)	62	23.78 (12; 34)	77.36 (74.48; 80.26)	72.48	24.05 (16.51; 31.19)
NPV (%)	79.28 (78.88; 79.71)	78.38	76.17 (75.49; 76.76)	84.09 (83.35; 84.8)	82.75	76.21 (75.19; 77.17)
AUC (%)	73.42 (72.51; 74.34)	72.05	49.98 (45.45; 54.5)	89.46 (88.93; 90.02)	89.57	49.99 (45.79; 54.5)

type of data. On the contrary, LASSO is a linear classifier and thus, its performance depends on whether the data is linearly separable. Whereas the generalizability of this observation is limited by the use of only one linear classifier, it may point to a better usability of non-linear approaches for classification of digital assessments.

For discrimination of PD and HC, combination of all tasks reached a BA of 74%, followed by tapping that achieved 67%, outperforming other tasks which were close to chance level. These results are in line with previous literature using the mPower dataset, where tapping reached the highest accuracies and gait and voice were closer to chance level [29]. Several studies reported higher accuracies for this type of data [24], [27]. Yet, these studies followed certain "optimistic" approaches as discussed below.

B. POTENTIAL CONFOUNDERS

Exclusion of comorbidities resulted in increased accuracies by a few percent, suggesting that other diseases may add more variability to the signal. Prediction performances considerably decreased for all tasks after matching for age and sex indicating the importance of controlling for such confounds in DB data. When including all data without trimming age range, accuracies greatly increase. Nonetheless, specificity

TABLE 18. Classification performance for the voice task.

	Additional experiment: All data			Additional experiment: All data + age + gender		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	52.91 (52.13; 53.68)	52.57	50.03 (49.26; 51.01)	73.12 (71.9; 74.35)	74.6	50 (48.09; 52.19)
Sensitivity (%)	7.59 (6.06; 9.16)	6.23	2.17 (0.94; 3.74)	50.68 (48.29; 53.11)	53.89	14.58 (11.53; 18.07)
Specificity (%)	98.24 (97.93; 98.55)	98.9	97.89 (97.57; 98.28)	95.55 (95.11; 96.01)	95.31	85.43 (84.66; 86.31)
PPV (%)	52.05 (45.24; 58.7)	58.82	20.49 (8.82; 35.29)	74.18 (72.28; 76.2)	74.25	20.08 (15.88; 24.89)
NPV (%)	80.84 (80.58; 81.1)	80.77	79.93 (79.68; 80.26)	88.49 (88; 89)	89.17	79.93 (79.21; 80.75)
AUC (%)	73.23 (72.43; 74.01)	74.91	50.09 (46.63; 53.52)	91.39 (91.05; 91.72)	91.16	49.99 (46.49; 53.58)

TABLE 19. Classification performance for the tapping task.

	Experiment 7			Experiment 8		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	74.42 (73.79; 74.99)	71.08	50.03 (48.27; 51.87)	80.81 (79.99; 81.58)	78.91	50 (48.22; 51.82)
Sensitivity (%)	52.65 (51.49; 53.77)	45.87	10.9 (7.98; 13.96)	65.68 (64.01; 67.28)	61.25	13.22 (10.26; 16.24)
Specificity (%)	96.18 (95.91; 96.43)	96.28	89.16 (88.56; 89.78)	95.94 (95.68; 96.24)	96.57	86.78 (86.18; 87.4)
PPV (%)	73.75 (72.41; 75.15)	71.56	17.01 (12.44; 21.78)	76.76 (75.53; 78.17)	78.47	16.94 (13.14; 20.8)
NPV (%)	90.87 (90.66; 91.07)	89.72	83.08 (82.52; 83.66)	93.2 (92.9; 93.49)	92.44	83.07 (82.49; 83.66)
AUC (%)	89.39 (89.08; 89.71)	88.64	50.04 (46.6; 53.56)	94.51 (94.29; 94.72)	94.78	50.12 (46.59; 53.37)

values are exceedingly high whereas sensitivity values are vastly low. This indicates a greater prediction ability for the HC group, which is considerably larger than the PD group for subjects under 35 years old. Including age and sex as part of the features resulted in further accuracy increases, yet with very low sensitivities. Since the dataset is strongly slanted toward young HC, the model is most likely distinguishing HC based on age and gender in this case. Such effects may also explain the high accuracies in some of the previous studies using mPower dataset, where no proper matching for these confounds was performed, age and/or sex were used as features despite a large imbalance across groups or non-balanced accuracies were reported [24], [26], [27], [34]. In example, in the overall mPower dataset HC outnumber PD by a factor of five and age and sex alone provide a high discrimination accuracy between PD and HC with PD being on average 28 years older and more often female (34% of PD vs 19% of HC). Our findings are also in line with previous studies demonstrating a similarly strong decrease in accuracies when accounting for respective confounds. Neto *et al.* [53] studied the effect of confounders on gait data. They reached very high accuracy when not accounting for confounders, compared with a very modest accuracy when using unconfounded measures. Schwab and Karlent [25] performed analysis with all

TABLE 20. Classification performance for the multimodal features.

	Experiment 7			Experiment 8		
	CV (95% CI)	Holdout	Chance (95% CI)	CV (95% CI)	Holdout	Chance (95% CI)
Balanced accuracy (%)	72.67 (71.82; 73.48)	73.26	50.07 (47.2; 52.97)	75.34 (74.44; 76.19)	75.88	50.06 (47.32; 53.06)
Sensitivity (%)	49.44 (47.78; 50.94)	49.77	14.89 (10.8; 19.25)	54.66 (53.05; 56.32)	55.63	16.79 (12.68; 21.13)
Specificity (%)	95.91 (95.52; 96.33)	96.75	85.25 (83.75; 86.69)	96.03 (95.67; 96.41)	96.13	83.32 (81.89; 84.83)
PPV (%)	79.95 (78.26; 81.67)	83.46	24.97 (17.83; 32.28)	81.96 (80.51; 83.51)	82.58	24.93 (18.82; 31.69)
NPV (%)	85.18 (84.77; 85.57)	85.38	75.23 (73.98; 76.5)	86.52 (86.08; 86.95)	86.79	75.23 (74; 76.57)
AUC (%)	88.99 (88.46; 89.48)	89.22	50.15 (45.55; 54.75)	92.39 (91.97; 92.81)	92.15	50.03 (45.74; 54.63)

the tasks from the mPower dataset with and without including age and sex, the latter resulting in a similarly low accuracy as in our study.

For all classification experiments, we used only one recording per subject to prevent the classifier from detecting the idiosyncrasies of each subject rather than specific PD related symptoms [29]–[31]. Single measures are likely to contain more noise due to higher variation in task administration as well as in individual performance in a poorly-controlled setting [54]. Using multiple time points may therefore further increase the discrimination between PD and HC as demonstrated in several previous studies [29]–[31]. Yet, our results in this respect highlight the need of further understanding and better control of the individual parameters which impact the task performance during a single administration.

C. PREDICTORS OF PARKINSON'S DISEASE

Features with largest weights in the multimodal discrimination between PD and HC were derived from the tapping task. These features mostly related to the inter-tapping interval (time), presumably reflecting bradykinesia-like symptoms. These results are in line with previous studies, where tapping features related to speed and accuracy had the strongest correlation with clinical scores [55], [56]. Balance task features related to tremor measures had larger weights than postural ones. In addition, features from the frequency domain had greater weights than spatiotemporal features. Spatiotemporal features have been extensively studied and applied, due to their ease of computation and interpretability [57]. However, these features offer information limited primarily to leg movement, whilst frequency features add information regarding asymmetry and variability. Furthermore, balance features with higher weights belonged to the mediolateral and anteroposterior signals, related to stability. Even though gait had limited contribution to the classification accuracy, acceleration features had the highest weights from this task. This observation is in line with previous findings where acceleration proved to better capture PD-related gait changes [58]. In line with some previous studies, features with the highest weights from the voice task were all based on Mel Frequency Cepstral Coefficients which can detect subtle changes in speech articulation that are common in PD [59], [60].

D. LIMITATIONS AND FURTHER RESEARCH

Whereas sensors-integrated in smartphones open new opportunities for at-home continuous, reliable, non-invasive and low-cost monitoring of PD, our finding highlight the need for further development, optimization and standardization of specific measures for such applications.

The interpretation of our findings is limited by several aspects, including the lack of standardization, poor control of environmental and medication effects during performance of the tasks and intentionally or unintentionally incorrect information provided by the participants. In addition, removal of comorbidities and matching for age and sex led to exclusion of about 50% of data, which may affect the training of classifiers [53].

Further use of smartphones in the detection of Parkinson's disease symptoms include detection of hypomimia from face expressions, socializing and lifestyle behavior and typing patterns among others [61], [62].

APPENDIX A
SUPPLEMENTARY METHODS

A. DATA CLEANING

MPower dataset offers demographic, PDQ8 and MDS-UPDRS surveys and task-based data. The demographics table contains data for 6805 subjects. In order to establish a diagnosis, participants had to select "true" or "false" to the following question "Have you been diagnosed by a medical professional with Parkinson Disease?". According to this answer, they are classified as Parkinson's Disease (PD) or Healthy Control (HC). Some subjects left this question unanswered and thus they were discarded from further analysis. Those subjects classified as PD which did not completed the PDQ8 and MDS-UPDRS questionnaire were also excluded. Subjects with no information on age, sex or any task data were also removed, resulting in 6614 subjects. Those empty, null or corrupted files for each task were deleted, resulting in 2807 subjects with gait and balance data, 4925 with voice data and 6366 with tapping data. Since a large number of subjects are HC under 35 years old, our analysis focused on a subset of subjects within the age range of 35 to 75 years old, leading to 1435 subjects with gait and balance data, 2186 subjects with voice data and 2644 subjects with tapping data. Finally, all subjects with inconsistencies for each of the tasks were discarded (i.e., subjects that reported not to have been diagnosed with Parkinson's disease but filled in PD medication questions, year of diagnosis of PD, surgery or deep brain stimulation). This last elimination resulted in 1416 subjects with gait and balance data, 2153 subjects with voice data and 2600 subjects with tapping data.

B. SIGNALS LENGTH

Gait task consists of walking 20 steps in a straight line. In order to analyse the same signal length for each subject, we examined how many subjects had gait data for different time durations. We observed that after 10 seconds, participation was dropping heavily. Therefore, we selected a time length of 10 seconds and discarded those participants with shorter signals. Following the same reasoning, we chose voice signals of 7 seconds, trimming the first second and last two seconds, and tapping signals of 20 seconds. Similarly, balance task consists of standing still for 30 seconds although just 20 seconds were selected. Nonetheless, whereas gait, voice and tapping are independent tasks, and therefore they are started by the user, balance task starts straight after the gait task. This is, as soon as the gait task ends, the app plays out loud "turn around and stand still for 30 seconds". As a result, most of the balance recordings include initial slots of noise, which most likely coincide with the time that subjects listen to the instructions, react, turn around and start the task. Therefore, we trimmed the first 5 seconds of the signal, resulting in balance signals of 15 seconds for all subjects. Final number of subjects consisted of 1397 subjects with gait data, 1415 subjects with balance data, 2150 subjects with voice data and 2600 subjects with tapping data.

C. PRE-PROCESSING AND SIGNAL EXTRACTION

Gait and balance data consists on vertical (V), anteroposterior (AP) and mediolateral (ML) acceleration signals. For these 3 gait acceleration signals, we applied a Butterworth low pass filter with cut-off frequency at 20 Hz followed by a 3° order high pass filter at 0.3 Hz. According to Pittman *et al.* [24], around 30% of devices were not held in the correct position. Therefore, the greatest gravitational displacement is not always along the vertical axis. Then, we standardized these three signals and calculated an additional average acceleration signal. Based on the standardized acceleration signal, we extracted the step series. We calculated position signals along the three axes by double integrating the acceleration signals and the average position. Then, we extracted velocity and acceleration along the path by derivation [37].

Balance acceleration signals were filtered with a low pass Butterworth filter at 20 Hz. Since tremor in PD usually falls in the 4-7Hz frequency band [38], [39], the interval 0-3.5 Hz is considered for tremor-free or postural acceleration measures. Hence, we applied a Butterworth filter at 3.5 Hz to extract postural acceleration measures. We also calculated the average of the tremor acceleration in the 3 axes and the average of the postural acceleration in the 3 axes.

Voice signals were recorded at a sample frequency of 44.1 KHz. We downsampled the signal to 25KHz, applied a second order Butterworth filter with cut-off frequency at 400 Hz followed by a pre-emphasis FIR filter for noise reduction and correct for distortions. We extracted the fundamental frequency (f0) series, which was verified with the Praat software.

Tapping recordings consists of the {x,y} screen pixel coordinates and timestamp for each tap on the screen. Signals derived out of these recordings were the inter-tapping interval (time) and the {x,y} inter-tap distance series.

D. FEATURE EXTRACTION

1) GAIT
We extracted 11 signals from the original accelerometer recordings during gait tasks. These are V, AP and ML acceleration, the step series, the average of the acceleration in the three axes, the V, AP and ML position, the average position in the three axes, the velocity and the acceleration along the path. Table 4 collects a list of features extracted for these signals along with their acronyms.

2) BALANCE
Balance signals consist in the V, AP and ML tremor acceleration (4-7 Hz), the average of these 3 signals, the V, AP and ML postural acceleration (0-3.5 Hz) and the average of these 3 signals. We extracted displacement-related postural features from ML, AP and average of both distance signals, following the formulation in Martinez-Mendez *et al.* [36] (Table 5).

3) VOICE
Most of voice features were extracted following the formulation in Tsanas *et al.* [45]. Tsanas *et al.* state that the period (T) signal provides different information than f0. Therefore, we additionally extracted the T series. Further signals include glottis quotient and 14 Mel Frequency Cepstral Coefficients (MFCCs),

including the 0^{th} coefficient and the log-energy of the signal, along with their associated delta and delta-delta coefficients as applied in the Voicebox Matlab ToolBox [63] (Table 6).

4) TAPPING
We considered a set of features computed from the inter-tapping interval (time) and the {x,y} inter-tap distance signals, according to Bot *et al.* [46] (Table 7).

E. COMORBIDITIES

Comorbidities selected for removal in the experiments E3-E6 include ‘‘Alzheimer Disease or Alzheimer dementia’’, ‘‘Dementia’’, ‘‘Schizophrenia or Bipolar Disorder’’, ‘‘Alcoholism’’, ‘‘Multiple Sclerosis’’, ‘‘Leukemia or Lymphoma’’, ‘‘Acute Myocardial Infarction/Heart Attack’’, ‘‘Stroke/Transient Ischemic Attack’’, ‘‘Breast Cancer’’, ‘‘Colorectal Cancer’’, ‘‘Prostate Cancer’’, ‘‘Lung Cancer’’, ‘‘Endometrial/Uterine Cancer’’, ‘‘Any other kind of cancer OR tumor’’, ‘‘Heart Failure/Congestive Heart Failure’’, ‘‘Ischemic Heart Disease’’. These comorbidities were removed since they may lead to brain damage or to undertake chemotherapy or other therapy, which might induce brain changes.

F. MEDICATION STATUS

Table 8 shows the number of subjects that performed the task just before taking their medication, after taking their medication, at another random time, number of those who were not taking any medication and number of those who did not give any information about their medication status.

G. SELECTION OF THE BEST CLASSIFIER DURING THE MAIN EXPERIMENT (NO COMORBIDITIES; MATCHED)

Table 9 shows the classification performance for the four classifiers under consideration for each task.

APPENDIX B
SUPPLEMENTARY RESULTS

Table 10-Table 14 summarize the results for each task (gait, balance, voice, tapping) and the combination of all the tasks, for the experiment 1 (all data), experiment 2 (matched data), experiment 3 (no comorbidities and matched data), experiment 4 (no comorbidities, matched, controlled for age), experiment 5 (no comorbidities, matched, controlled for sex) and experiment 6 (no comorbidities, matched, controlled for age and sex).

A. ADDITIONAL EXPERIMENTS

Our results may differ to those in the current literature using the mPower dataset since we follow different approaches. To explain these discrepancies and compare with the literature, we included two additional experiments including all data without trimming for age range and all data including both age and sex as features in the analyses (Table 15). Classification performances for both additional experiments for each tasks are summarized in Table 16-Table 20.

ACKNOWLEDGMENT

Data used in this study were contributed by users of the Parkinson m-Power mobile application as part of the m-Power study developed by Sage Bionetworks and described in Synapse [doi:10.7303/syn4993293].

AUTHOR CONTRIBUTION

María Goñi performed the overall analyses and wrote manuscript. Juergen Dukart designed the overall study and contributed to writing the manuscript. Kaustubh R. Patil, Simon B. Eickhoff, and Mehran Sahandi Far provided input on the analyses. All authors contributed to interpretation of results, reviewed, and commented on the manuscript.

AUTHOR DECLARATION

The access to the Mpower data was granted after registration in the Synapse system, signing an oath, submitting an Intended Data Use Statement and accepting data-specific Conditions. All appropriate institutional forms have been archived.

REFERENCES

[1] C. H. Adler, T. G. Beach, J. G. Hentz, H. A. Shill, J. N. Caviness, E. Driver-Dunckley, M. N. Sabbagh, L. I. Sue, S. A. Jacobson, C. M. Belden, and B. N. Dugger, ‘‘Low clinical diagnostic accuracy of early vs advanced Parkinson disease: Clinicopathologic study,’’ *Neurology*, vol. 83, no. 5, pp. 406–412, Jul. 2014.

[2] J.-W. Kim, Y. Kwon, Y.-M. Kim, H.-Y. Chung, G.-M. Eom, J.-H. Jun, J.-W. Lee, S.-B. Koh, B. K. Park, and D.-K. Kwon, ‘‘Analysis of lower limb bradykinesia in Parkinson’s disease patients,’’ *Geriatrics Gerontology Int.*, vol. 12, no. 2, pp. 257–264, Apr. 2012.

[3] J.-F. Daneault, S. I. Lee, F. N. Golabchi, S. Patel, L. C. Shih, S. Paganoni, and P. Bonato, ‘‘Estimating bradykinesia in Parkinson’s disease with a minimum number of wearable sensors,’’ in *Proc. IEEE/ACM Int. Conf. Connected Health: Appl., Syst. Eng. Technol. (CHASE)*, Jul. 2017, pp. 264–265.

[4] H. Zach, A. M. Janssen, A. H. Snijders, A. Delval, M. U. Ferraye, E. Auff, V. Weerdesteyn, B. R. Bloem, and J. Nonnekes, ‘‘Identifying freezing of gait in parkinson’s disease during freezing provoking tasks using waist-mounted accelerometry,’’ *Parkinsonism Rel. Disorders*, vol. 21, no. 11, pp. 1362–1366, Nov. 2015.

[5] A. Suppa, A. Kita, G. Leodori, A. Zampogna, E. Nicolini, P. Lorenzi, R. Rao, and F. Irrera, ‘‘L-DOPA and freezing of gait in Parkinson’s disease: Objective assessment through a wearable wireless system,’’ *Frontiers Neurol.*, vol. 8, p. 406, Aug. 2017.

[6] N.-H. Ko, C. M. Laine, B. E. Fisher, and F. J. Valero-Cuevas, ‘‘Force variability during dexterous manipulation in individuals with mild to moderate Parkinson’s disease,’’ *Frontiers Aging Neurosci.*, vol. 7, p. 151, Aug. 2015.

[7] R. P. Hubble, G. A. Naughton, P. A. Silburn, and M. H. Cole, ‘‘Wearable sensor use for assessing standing balance and walking stability in people with Parkinson’s disease: A systematic review,’’ *PLoS ONE*, vol. 10, no. 4, Apr. 2015, Art. no. e0123705.

[8] H. Dubey, J. C. Goldberg, M. Abtahi, L. Mahler, and K. Mankodiya, ‘‘EchoWear: Smartwatch technology for voice and speech treatments of patients with Parkinson’s disease,’’ 2016, *arXiv:1612.07608*.

[9] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, ‘‘Fully automated assessment of the severity of Parkinson’s disease from speech,’’ *Comput. Speech Lang.*, vol. 29, no. 1, pp. 172–185, Jan. 2015.

[10] A. J. Espay, P. Bonato, F. B. Nahab, W. Maetzler, J. M. Dean, J. Klucken, B. M. Eskofier, A. Merola, F. Horak, A. E. Lang, and R. Reilmann, ‘‘Technology in parkinson’s disease: Challenges and opportunities,’’ *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, Apr. 2016.

[11] E. Rovini, C. Maremmani, and F. Cavallo, ‘‘How wearable sensors can support Parkinson’s disease diagnosis and treatment: A systematic review,’’ *Frontiers Neurosci.*, vol. 11, p. 555, Oct. 2017.

[12] M. Linares-del Rey, L. Vela-Desojo, and R. Cano-de la Cuerda, "Mobile phone applications in Parkinson's disease: A systematic review," *Neurología*, vol. 34, no. 1, pp. 38–54, Jan. 2019.

[13] W. Maetzler, J. Domingos, K. Srulijes, J. J. Ferreira, and B. R. Bloem, "Quantitative wearable sensors for objective assessment of Parkinson's disease," *Movement Disorders*, vol. 28, no. 12, pp. 1628–1637, Oct. 2013.

[14] S. Arora, V. Venkataraman, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, "High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3641–3644.

[15] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Rel. Disorders*, vol. 21, no. 6, pp. 650–653, Jun. 2015.

[16] A. Benba, A. Jilbab, and A. Hammouch, "Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines," *Int. J. Electr. Eng. Inform.*, vol. 7, no. 2, pp. 297–307, Jun. 2015.

[17] N. Kostikis, D. Hristu-Varsakelis, M. Arnaoutoglou, and C. Kotsavasiloglou, "A smartphone-based tool for assessing parkinsonian hand tremor," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1835–1842, Nov. 2015.

[18] M. Suzuki, H. Mitoma, and M. Yoneyama, "Quantitative analysis of motor status in Parkinson's disease using wearable devices: From methodological considerations to problems in clinical applications," *Parkinsons Dis*, vol. 2017, May 2017, Art. no. 6139716.

[19] K. Szewczyk-Krolikowski, P. Tomlinson, K. Nithi, R. Wade-Martins, K. Talbot, Y. Ben-Shlomo, and M. T. M. Hu, "The influence of age and gender on motor and non-motor features of early Parkinson's disease: Initial findings from the Oxford Parkinson disease center (OPDC) discovery cohort," *Parkinsonism Rel. Disorders*, vol. 20, no. 1, pp. 99–105, Jan. 2014.

[20] M. Picillo, A. Nicoletti, V. Fetoni, B. Garavaglia, P. Barone, and M. T. Pellecchia, "The relevance of gender in Parkinson's disease: A review," *J. Neurol.*, vol. 264, no. 8, pp. 1583–1607, Aug. 2017.

[21] S. Nazem, A. D. Siderowf, J. E. Duda, T. T. Have, A. Colcher, S. S. Horn, P. J. Moberg, J. R. Wilkinson, H. I. Hurtig, M. B. Stern, and D. Weintraub, "Montreal cognitive assessment performance in patients with Parkinson's disease with 'normal' global cognition according to mini-mental state examination score," *J. Amer. Geriatrics Soc.*, vol. 57, no. 2, pp. 304–308, Feb. 2009.

[22] M. M. Wickremaratchi, M. D. W. Knipe, B. S. D. Sastry, E. Morgan, A. Jones, R. Salmon, R. Weiser, M. Moran, D. Davies, L. Ebenezer, S. Raha, N. P. Robertson, C. C. Butler, Y. Ben-Shlomo, and H. R. Morris, "The motor phenotype of Parkinson's disease in relation to age at onset," *Movement Disorders*, vol. 26, no. 3, pp. 457–463, Feb. 2011.

[23] L. M. Shulman, R. L. Taback, J. Bean, and W. J. Weiner, "Comorbidity of the nonmotor symptoms of Parkinson's disease," *Movement Disorders, Official J. Movement Disorder Soc.*, vol. 16, no. 3, pp. 507–510, 2001.

[24] B. Pittman, R. H. Ghomi, and D. Si, "Parkinson's disease classification of mPower walking activity participants," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4253–4256.

[25] P. Schwab and W. Karlen, "PhoneMD: Learning to diagnose Parkinson's disease from smartphone data," 2018, *arXiv:1810.01485*.

[26] J. Prince, F. Andreotti, and M. De Vos, "Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1402–1411, May 2019.

[27] S. Mehrang, M. Jauhiainen, J. Pietila, J. Puustinen, J. Ruokolainen, and H. Nieminen, "Identification of Parkinson's disease utilizing a single self-recorded 20-step walking test acquired by Smartphone's inertial measurement unit," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2913–2916.

[28] M. Memedi, A. Sadikov, V. Groznik, J. Žabkar, M. Možina, F. Bergquist, A. Johansson, D. Haubenberger, and D. Nyholm, "Automatic spiral analysis for objective assessment of motor symptoms in Parkinson's disease," *Sensors*, vol. 15, no. 9, pp. 23727–23744, Sep. 2015.

[29] E. C. Neto, T. M Perumal, A. Pratap, B. M Bot, L. Mangravite, and L. Omberg, "On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: The mPower case study," 2017, *arXiv:1706.09574*.

[30] E. C. Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, P. Snyder, B. M. Bot, A. D. Trister, S. H. Friend, L. Mangravite, and L. Omberg, "Detecting the impact of subject characteristics on machine learning-based diagnostic applications," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–6, Oct. 2019.

[31] E. C. Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, B. M. Bot, A. D Trister, S. H Friend, L. Mangravite, and L. Omberg, "Learning disease vs participant signatures: A permutation test approach to detect identity confounding in machine learning diagnostic applications," 2017, *arXiv:1712.03120*.

[32] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, and A. D. Trister, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Sci. Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016.

[33] A. Zhan, S. Mohan, C. Tarolli, R. B. Schneider, J. L. Adams, S. Sharma, M. J. Elson, K. L. Spear, A. M. Glidden, M. A. Little, and A. Terzis, "Using smartphones and machine learning to quantify Parkinson disease severity: The mobile Parkinson disease score," *JAMA Neurol.*, vol. 75, no. 7, pp. 876–880, Jul. 2018.

[34] M. Giuliano, A. García-López, S. Pérez, F. D. Pérez, O. Sposito, and J. Bossero, "Selection of voice parameters for Parkinson's disease prediction from collected mobile data," in *Proc. 22nd Symp. Image Signal Process. Artif. Vis. (STSIVA)*, Apr. 2019, pp. 1–3.

[35] J. Prince and M. de Vos, "A deep learning framework for the remote detection of Parkinson'S disease using smart-phone sensor data," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 3144–3147.

[36] R. Martinez-Mendez, M. Sekine, and T. Tamura, "Postural sway parameters using a triaxial accelerometer: Comparing elderly and young healthy adults," *Comput. Methods Biomech. Biomed. Eng.*, vol. 15, no. 9, pp. 899–910, Sep. 2012.

[37] K. Seifert and O. Camacho, "Implementing positioning algorithms using accelerometers," *Freescale Semicond.*, vol. 1, p. 13, Feb. 2007.

[38] K. E. Lyons, R. Pahwa, and R. Pahwa, *Handbook of Essential Tremor and Other Tremor Disorders*. Boca Raton, FL, USA: CRC Press, 2005.

[39] L. Palmerini, L. Rocchi, S. Mellone, F. Valzania, and L. Chiari, "Feature selection for accelerometer-based posture analysis in Parkinson's disease," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 481–490, May 2011.

[40] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. R. Dorsey, S. Saria, and A. Terzis, "High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection," 2016, *arXiv:1601.00960*.

[41] A. Weiss, S. Sharifi, M. Plotnik, J. P. P. van Vugt, N. Giladi, and J. M. Hausdorff, "Toward automated, at-home assessment of mobility among patients with Parkinson disease, using a body-worn accelerometer," *Neurorehabilitation Neural Repair*, vol. 25, no. 9, pp. 810–818, Nov. 2011.

[42] R. San-Segundo, R. Torres-Sánchez, J. Hodgins, and F. De la Torre, "Increasing robustness in the detection of freezing of gait in Parkinson's disease," *Electronics*, vol. 8, no. 2, p. 119, Jan. 2019.

[43] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Troster, "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 436–446, Mar. 2010.

[44] T. E. Prieto, J. B. Myklebust, R. G. Hoffmann, E. G. Lovett, and B. M. Myklebust, "Measures of postural steadiness: Differences between healthy young and elderly adults," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 9, pp. 956–966, Sep. 1996.

[45] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *J. Roy. Soc. Interface*, vol. 8, no. 59, pp. 842–855, Jun. 2011.

[46] B. M. Bot. *mPower: Public Researcher Portal*. Accessed: Jun. 25, 2020. [Online]. Available: https://www.synapse.org/#!Synapse:syn4993293/files/

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B*, vol. 58, no. 1, pp. 267–288, 1996.

[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[49] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, Jan. 2002.

[50] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[51] S. Bind, A. K. Tiwari, and A. K. Sahani, "A survey of machine learning based approaches for Parkinson disease prediction," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1648–1655, 2015.
[52] P. B. Brazdil, C. Soares, and J. P. da Costa, "Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results," *Mach. Learn.*, vol. 50, no. 3, pp. 251–277, Mar. 2003.
[53] E. C. Neto, A. Pratap, T. M Perumal, M. Tummalacherla, B. M Bot, L. Mangravite, and L. Omberg, "Using permutations to assess confounding in machine learning applications for digital health," 2018, *arXiv:1811.11920*.
[54] M. S. Far, S. B. Eickhoff, M. Goñi, and J. Dukart, "Exploring test retest reliability and longitudinal stability of digital biomarkers for Parkinson's disease in the m-power dataset: Cohort study," *J. Med. Internet Res.*, vol. 23, no. 9, p. e26608, Sep. 2021.
[55] M. Memedi, T. Khan, P. Grenholm, D. Nyholm, and J. Westin, "Automatic and objective assessment of alternating tapping performance in Parkinson's disease," *Sensors*, vol. 13, no. 12, pp. 16965–16984, Dec. 2013.
[56] C. Y. Lee, S. J. Kang, S.-K. Hong, H.-I. Ma, U. Lee, and Y. J. Kim, "A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease," *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0158852.
[57] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge, and D. C. Ackland, "Classification of parkinson's disease gait using spatial-temporal gait features," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1794–1802, Nov. 2015.
[58] E. Sejdic, K. A. Lowry, J. Bellanca, M. S. Redfern, and J. S. Brach, "A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 3, pp. 603–612, May 2014.
[59] T. Khan. Running-Speech MFCC are Better Markers of Parkinsonian Speech Deficits Than Vowel Phonation and Diadochokinetic. DiVA. Accessed: Dec. 2020. [Online]. Available: http://mdh.diva-portal.org/smash/record.jsf?pid=diva2%3A705196&dswid=6494/
[60] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.
[61] Y. Seliverstov, D. Diagovchenko, M. Kravchenko, M. Babin, E. Fedotova, and M. Belyaev, "Hypomimia detection with a smartphone camera as a possible self-screening tool for Parkinson disease," *J. Neurol.*, vol. 90, no. 15, p. 3.047, 2018.
[62] R. B. Schneider, L. Omberg, E. A. Macklin, M. Daeschler, L. Bataille, S. Anthwal, T. L. Myers, E. Baloga, S. Duquette, P. Snyder, and K. Amodeo, "Design of a virtual longitudinal observational study in Parkinson's disease (AT-HOME PD)," *Ann. Clin. Transl. Neurol.*, vol. 8, no. 2, pp. 308–320, Feb. 2020.
[63] M. Brookes. *VOICEBOX: Speech Processing Toolbox for MATLAB*. Accessed: Dec. 2020. [Online]. Available: http://www.ee.ic.ac.U.K./hp/staff/dmb/voicebox/voicebox.html/
[64] A. Mirelman, T. Heman, K. Yasinovsky, A. Thaler, T. Gurevich, K. Marder, S. Bressman, A. Bar-Shira, A. Orr-Urtreger, N. Giladi, and J. M. Hausdorff, "Fall risk and gait in Parkinson's disease: The role of the LRRK2 G2019S mutation," *Movement Disorders*, vol. 28, no. 12, pp. 1683–1690, Oct. 2013.
[65] B. M. Bot. *Sage-Bionetworks: MPower-Sdata*. Accessed: Jul. 2020. [Online]. Available: https://github.com/Sage-Bionetworks/mPower-sdata/

MARÍA GOÑI received the B.S. degree in technical industrial engineering from the University of Cantabria, Spain, in 2008, the M.S. degree in biomedical engineering from the University of Pais Vasco, Spain, in 2014, the Ph.D. degree in neurosciences from the University of Aberdeen, U.K., in 2019, and the B.S. degree in electronics engineering from the University of Alcalá, Spain, in 2021.

Since 2019, she has been a Postdoctoral Researcher with the Institute of Neuroscience and Medicine, Research Centre Jülich, Germany. Her research interests include the identification of prognostic biomarkers in neuropsychiatric diseases by integrating different neuroimaging modalities and sensor-based data and the application of machine learning and other analytical techniques.

SIMON B. EICKHOFF is currently a Full Professor and the Chair of the Institute for Systems Neuroscience, Heinrich Heine University, Düsseldorf, and the Director of the Institute of Neuroscience and Medicine (INM-7, Brain and Behavior), Forschungszentrum Jülich. He is a Visiting Professor at the Institute of Automation, Chinese Academy of Sciences. He is working at the interface between neuroanatomy, data-science, and brain medicine. He aims to obtain a more detailed characterization of the organization of the human brain and its inter-individual variability in order to better understand its changes in advanced age and neurological and psychiatric disorders. This goal is pursued by the development and application of novel analysis tools and approaches for large-scale, multi-modal analysis of brain structure, function and connectivity and by machine-learning for single subject prediction of cognitive and socio-affective traits, and ultimately precision medicine.

MEHRAN SAHANDI FAR received the master's degree in computer science from Eastern Mediterranean University. He is currently pursuing the Ph.D. degree with the Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, and a Researcher with the Research Center Jülich's Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour). His current research interests include digital biomarker development and remote monitoring using smart devices.

KAUSTUBH R. PATIL (Member, IEEE) received the B.E. degree in electronics engineering from Shivaji University, India, in 2003, the M.Sc. degree in artificial intelligence and intelligent systems from the University of Porto, Portugal, in 2007, and the Ph.D. degree from the Max Planck Institute of Computer Science, Germany, in 2013. From 2013 to 2016, he was a Postdoctoral Fellow at UCL and MIT. He joined FZJ, in 2017, where he is currently leading the Applied Machine Learning Group. He works on the application of machine learning techniques to better understand biological systems. He is an Associate Editor of IEEE Access.

JUERGEN DUKART received the Diploma degree in psychology from Ruhr University Bochum, in 2008, and the Ph.D. degree in neuroscience from the Max Planck Institute for Human Cognitive and Brain Sciences, in 2011. He was a Postdoctoral Researcher at the University of Lausanne, Switzerland. Then, he moved to the Pharmaceutical Company F. Hoffmann-La Roche, where he worked for five years in different functions, such as the Head of Clinical Imaging and a Biomarker Experimental Medicine Leader being responsible for imaging and overall biomarker strategy in various phase I to phase III clinical trials. Since 2019, he has been leading the Group Biomarker Development, Institute of Neuroscience and Medicine (INM-7), Research Centre Jülich, Germany. He is the author of more than 60 articles. His main research interests include the development of technology-based biomarkers for early detection, follow-up, stratification, and monitoring of treatment effects in neurological and psychiatric diseases. He is an Associate Editor of the journal *Frontiers in Human Neuroscience*.