Limited diagnostic accuracy of smartphone-based digital biomarkers for Parkinson's disease in a remotely-administered setting

María Goñi [1,2], Simon Eickhoff [1,2], Mehran Sahandi Far[1,2], Kaustubh Patil [1,2], Juergen Dukart [1,2*]

[1] Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

[2] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

***Corresponding authors:**

Juergen Dukart, PhD

Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany

Email: juergen.dukart@gmail.com

Phone : +49 2461 61 3187

Abstract

Background

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.

Methods

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.

Results

When accounting for age, sex and comorbidities, the highest balanced accuracy on hold-out data (67%) was achieved using relevance vector machine on tapping and when combining all tasks. Only moderate accuracies were achieved for other tasks (60% for balance, 56% for gait and 55% for voice data). Not accounting for the confounders consistently yielded higher accuracies of up to 73% (for tapping) for all tasks.

Discussion

Our results demonstrate the importance of controlling DB data for age and comorbidities. They further point to a moderate power of commonly applied DB tasks to differentiate between PD and HC when conducted in poorly controlled self-administered settings.

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], [8], [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] [11], [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].

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 [14], [15]. Indeed, recent studies applying ML algorithms to this dataset suggest a good diagnostic accuracy of respective digital assessments for PD detection. However, use of different machine learning (ML) algorithms and the focus on one or few tasks limit the comparability across studies with respect to accuracy of different digital assessments for detection of PD [16]–[18]. In addition, such DB assessments may contain different confounds and other 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 [19]. For example, age, sex and comorbidities are known confounding factors that impact many measures of disease symptoms across neurodegenerative diseases including PD [20]–[24]. Several studies eluded the importance of matching and controlling for these variables which might affect motor (i.e. bradykinesia, tremor or rigidity) and non-motor (i.e. fatigue, restless legs or sleep) measures [25]–[28]. Other potential data collection confounds comprise inclusion of several recordings per subject and use of signals of different time length [16], [25], [28], 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]. Whilst plausible, the impact of these 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 [16], [25], [28], [32], [33].

Here we use the mPower dataset to systematically evaluate and compare the ability of common DB tasks (gait, balance, voice, tapping) for detection of PD in an at-home setting. We further systematically test which ML-based algorithms and which task features reported in the literature perform best for differentiation between PD and HC and how age, sex and comorbidities affect the respective accuracies.

METHODS

Data

Data used in this work were derived from the mPower study [14]. MPower is a mobile application-based study to monitor indicators of PD progression and diagnosis by the 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 [14]. 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 S5**). 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, 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 subject in the analyses. Further details about data cleaning can be found in Supplementary Material. Demographic details are shown in **Table 1**.

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)	-

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 [34]. Signals were then standardized to eliminate the gravity component while maintaining the information from outlier data. According to Pittman et al. [25], 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 [35] (**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 [36], while 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 [37].

Voice was recorded at a sample rate of 44.1 Kbps. Pre-processing included a downsampling to 15 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.

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 pre-processing for each task can be found in Supplementary Material.

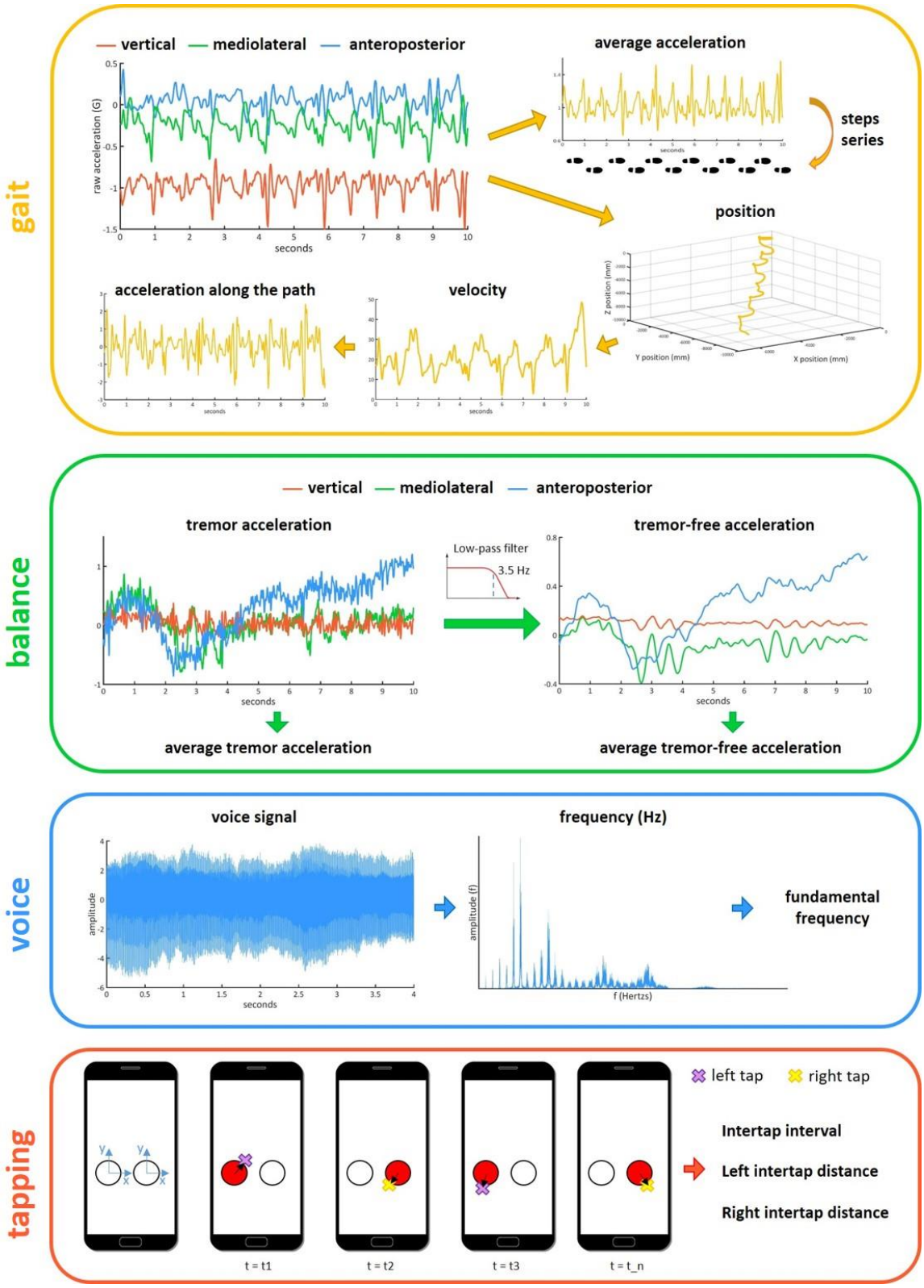

Figure 1. Illustration of signal processing and feature extraction based on the raw data for each task.

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 [38]–[41], balance [7], [34], [37], [42], voice [26], [27], [43] and tapping data [14], [16], [44], respectively (**Table S1-S4**).

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 [45]. 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 [46].

(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 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 [47].

(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 [48].

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 Supplementary Material) 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 to regress out their effects prior to 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 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**.

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

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 the number of trees for RF to 100. A nested cross-validation was implemented to tune the parameters of the SVM-RFE classifier, 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. 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 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.

RESULTS

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 S6** 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)): 69.1%), followed by tapping using RVM (BA: 67.9%), balance using RF (BA: 60%), gait using SVM-RFE (BA: 56.5%) and voice using RVM (BA: 54.8%).

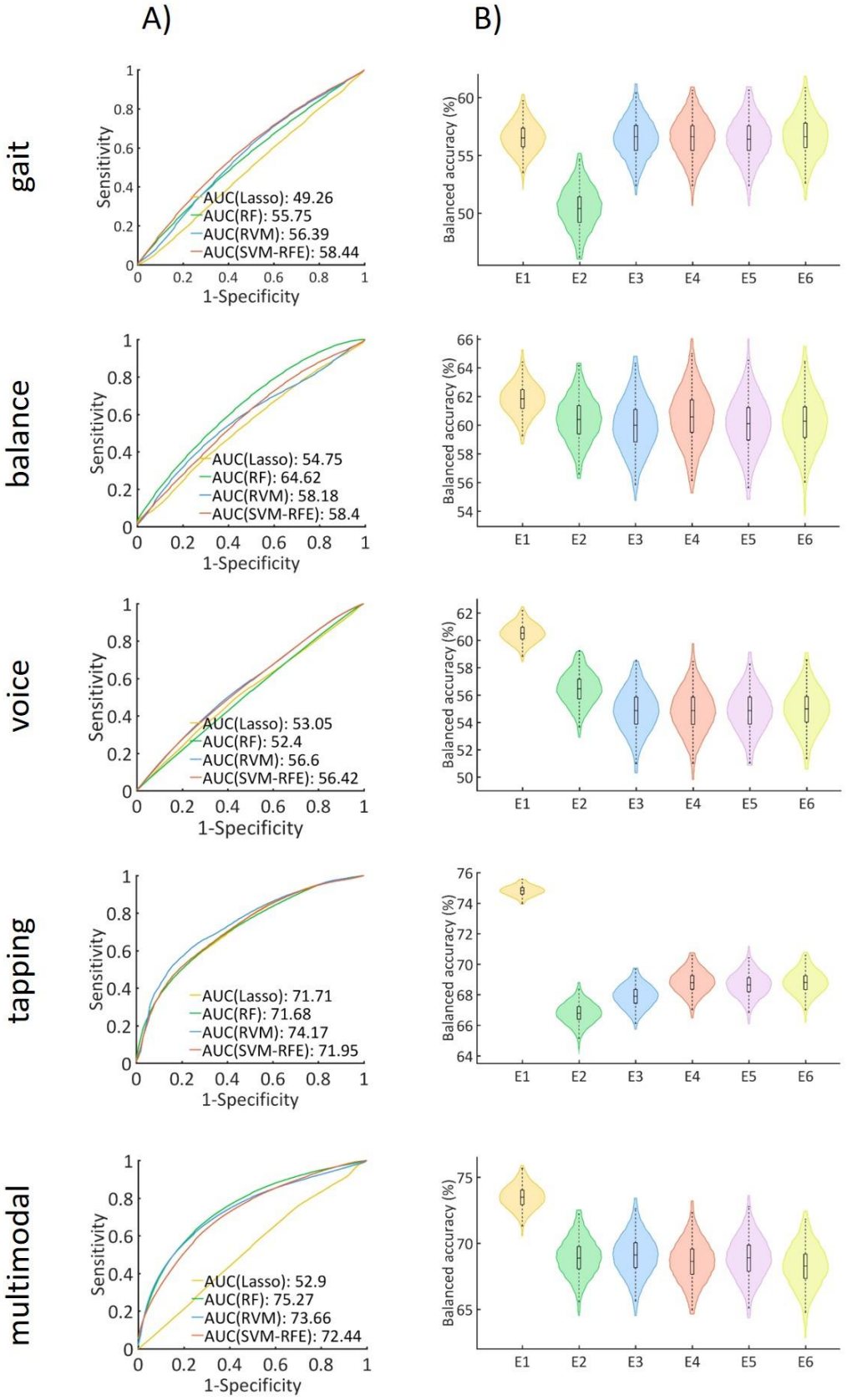
A)
B)
gait
balance
voice
tapping
multimodal
Sensitivity
1-Specificity
AUC(Lasso): 49.26
AUC(RF): 55.75
AUC(RVM): 56.39
AUC(SVM-RFE): 58.44
AUC(Lasso): 54.75
AUC(RF): 64.62
AUC(RVM): 58.18
AUC(SVM-RFE): 58.4
AUC(Lasso): 53.05
AUC(RF): 52.4
AUC(RVM): 56.6
AUC(SVM-RFE): 56.42
AUC(Lasso): 71.71
AUC(RF): 71.68
AUC(RVM): 74.17
AUC(SVM-RFE): 71.95
AUC(Lasso): 52.9
AUC(RF): 75.27
AUC(RVM): 73.66
AUC(SVM-RFE): 72.44
Balanced accuracy (%)
E1
E2
E3
E4
E5
E6

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.

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 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 S7-S11**. 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: 60.5%; tapping: 74.8; multimodal combining all four tasks: 73.5%). 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%) and tapping (E2: 66.8%; E3: 67.9%) but lower BA for balance (E2: 60.4%; E3: 60.0%) and voice (E2: 56.4%; E3: 54.8%). After additionally regressing out the effects of age and/or sex (E4-E6) the change in the BA was negligible (< 1%) (**Table 3**, **Table S7-S11**).

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	60.5 (59.7-61.4)	62.1	49.9 (46.3-53.2)	56.4 (55-58)	54.8	50.1 (44.7-55.3)	54.8 (52.8-56.5)	55.4	50.1 (44.6-55.4)	54.8 (52.9-56.8)	57.1	49.9 (44.3-55.7)	54.9 (52.9-56.7)	57.4	50.2 (44.6-55.4)	54.9 (52.9-56.9)	57.4	50.1 (45.3-55.4)

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	73.5 (72-75)	76.8	49.9 (45-54.1)	68.9 (66.4-71.4)	70.8	49.9 (43.3-55.8)	69.1 (66.4-71.8)	66.7	50.1 (43.1-56.9)	68.6 (65.8-71.3)	69.6	50.1 (43.1-56.9)	68.9 (66.2-71.6)	69.6	49.9 (43.1-56.9)	68.3 (65.2-71)	66.7	49.9 (43.1-55.9)

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**, **Table S7-S11**). All results are summarized in **Figure 3** and **Table 3**. Tapping features resulted in the best performance for differentiation of PD and HC in the holdout cohort (BA: 67.2%) followed by the multimodal combination of all tasks with a very similar BA (66.7%). Voice features achieved the lowest BA of 55.4% followed by gait (55.7%) and balance (59.9%) features. For the base experiment E3, the difference in BA between CV and holdout sets was less than 1% for all tasks with a 2.4% reduction in BA only observed for the multimodal feature combination. Exclusion of comorbidities resulted in only minor changes for all tasks (<2%) with a drop of 4% in BA only observed for the multimodal case (**Table 3**, **Table S11**). BA performance for all tasks increased by 1.4% (gait) to 10% (combined features) 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.

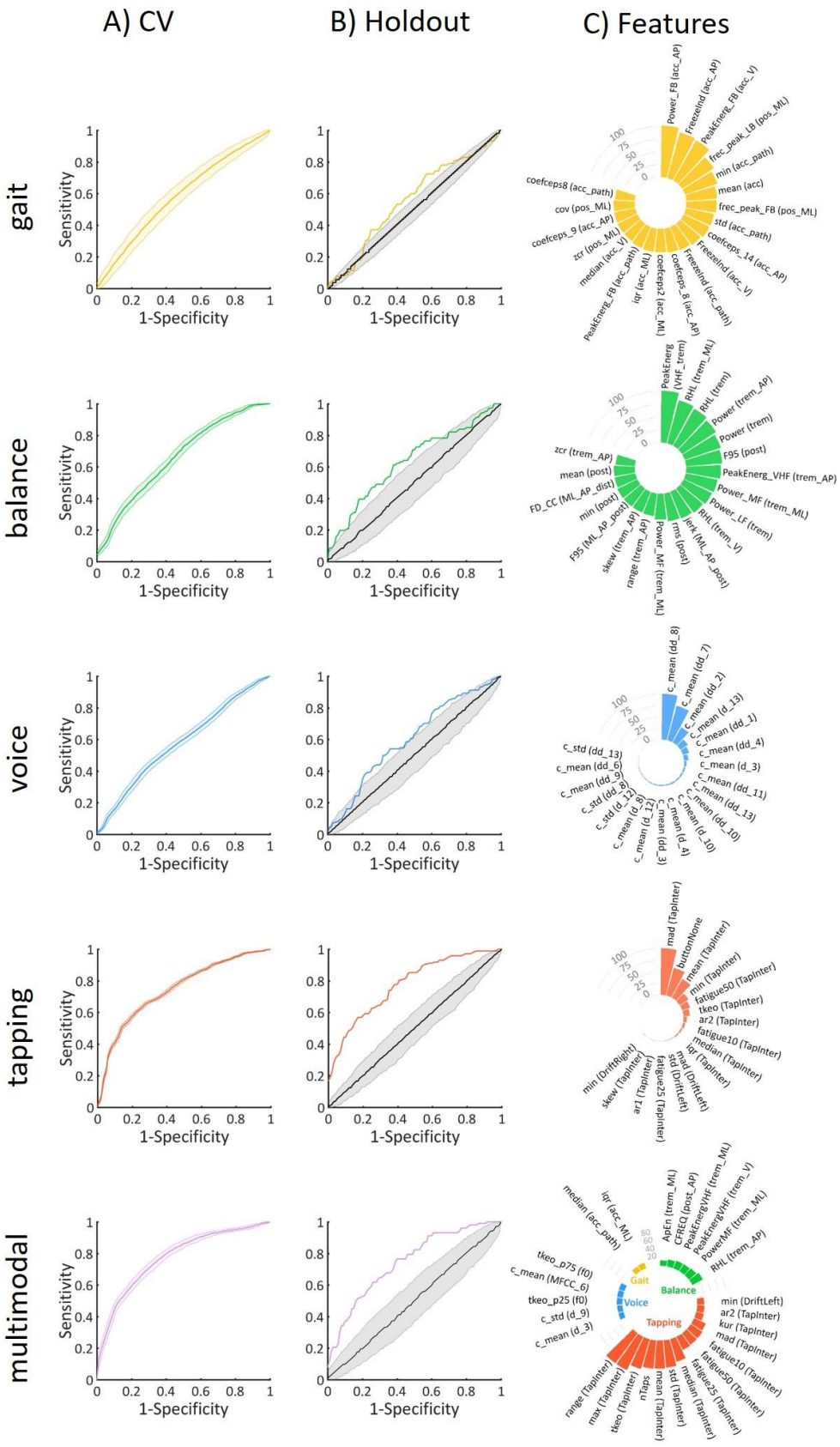
A) CV
B) Holdout
C) Features
gait
balance
voice
tapping
multimodal
Sensitivity
1-Specificity
Power_FB (acc_AP)
FreezeInd (acc_AP)
PeakEnerg_FB (acc_V)
frec_peak_LB (pos_ML)
min (acc_path)
mean (acc)
frec_peak_FB (pos_ML)
std (acc_path)
coefceps_14 (acc_AP)
FreezeInd (acc_V)
FreezeInd (acc_path)
coefceps_8 (acc_AP)
coefceps2 (acc_ML)
iqr (acc_ML)
PeakEnerg_FB (acc_path)
median (acc_V)
zcr (pos_ML)
coefceps_9 (acc_AP)
cov (pos_ML)
coefceps8 (acc_path)
PeakEnerg (VHF_trem)
RHL (trem_ML)
RHL (trem)
Power (trem_AP)
Power (trem)
F95 (post)
PeakEnerg_VHF (trem_AP)
Power_MF (trem_ML)
Power_LF (trem)
RHL (trem_V)
jerk (ML_AP_post)
rms (post)
Power_MF (trem_ML)
range (trem_AP)
skew (trem_AP)
F95 (ML_AP_post)
min (post)
FD_CC (ML_AP_dist)
mean (post)
zcr (trem_AP)
c_mean (dd_8)
c_mean (dd_7)
c_mean (dd_2)
c_mean (d_13)
c_mean (dd_1)
c_mean (dd_4)
c_mean (d_3)
c_mean (dd_11)
c_mean (dd_13)
c_mean (dd_10)
c_mean (d_10)
c_mean (d_4)
c_mean (dd_3)
c_mean (d_12)
c_mean (d_8)
c_std (d_12)
c_std (dd_8)
c_mean (dd_9)
c_mean (dd_6)
c_std (dd_13)
mad (TapInter)
buttonNone
mean (TapInter)
min (TapInter)
fatigue50 (TapInter)
tkeo (TapInter)
ar2 (TapInter)
fatigue10 (TapInter)
median (TapInter)
iqr (TapInter)
mad (DriftLeft)
std (DriftLeft)
fatigue25 (TapInter)
ar1 (TapInter)
skew (TapInter)
min (DriftRight)
median (acc_path)
iqr (acc_ML)
ApEn (trem_ML)
CREQ (post_AP)
PeakEnergVHF (trem_ML)
PowerMF (trem_V)
RHL (trem_AP)
Gait
Balance
Voice
Tapping
tkeo_p75 (f0)
c_mean (MFCC_6)
tkeo_p25 (f0)
c_std (d_9)
c_mean (d_3)
min (DriftLeft)
ar2 (TapInter)
kur (TapInter)
mad (TapInter)
fatigue10 (TapInter)
fatigue50 (TapInter)
fatigue25 (TapInter)
median (TapInter)
std (TapInter)
mean (TapInter)
nTaps
tkeo (TapInter)
max (TapInter)
range (TapInter)

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 supplementary material.

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 AP acceleration (31.5) and energy in the medium frequency band for mediolateral acceleration (25.3). Gait and voice tasks had the least contributions in terms of importance scores.

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 the utility of smartphones-based assessment to differentiate between PD and controls may be limited in such a self-administered and loosely-controlled setting. Moreover, we show that, depending on the constellation, not accounting for confounds in PD digital biomarker task data may lead to under- but also over-optimistic results.

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 [26], [27], others simply applied a single classifier [28], [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 [49], [50]. For instance, RF, RVM and Gaussian SVM are non-linear algorithms, offering more flexibility regarding the type of data. On the contrary, LASSO is a linear classifier and thus, its performance depends on whether the data is linearly separable. While 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, tapping features reached a BA of 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 [27], [28]. Yet, these studies followed certain “optimistic” approaches as discussed below.

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. 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 [25], [27], [28], [32]. 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. [51] 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 [26] performed analysis with all 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 [52]. 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.

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 [53], [54]. 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 [55]. 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 [56]. 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 [57], [58].

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

Importantly, the interpretation of our findings is limited by several aspects. Potential limitations include 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 [51].

REFERENCES

[1] C. H. Adler *et al.*, 'Low clinical diagnostic accuracy of early vs advanced Parkinson disease: clinicopathologic study', *Neurology*, vol. 83, no. 5, pp. 406–412, Jul. 2014, doi: 10.1212/WNL.0000000000000641.

[2] J.-W. Kim *et al.*, 'Analysis of lower limb bradykinesia in Parkinson's disease patients', *Geriatrics & Gerontology International*, vol. 12, no. 2, pp. 257–264, 2012, doi: 10.1111/j.1447-0594.2011.00761.x.

[3] J.-F. Daneault *et al.*, 'Estimating Bradykinesia in Parkinson's Disease with a Minimum Number of Wearable Sensors', in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Jul. 2017, pp. 264–265, doi: 10.1109/CHASE.2017.94.

[4] H. Zach *et al.*, 'Identifying freezing of gait in Parkinson's disease during freezing provoking tasks using waist-mounted accelerometry', *Parkinsonism & Related Disorders*, vol. 21, no. 11, pp. 1362–1366, Nov. 2015, doi: 10.1016/j.parkreldis.2015.09.051.

[5] A. Suppa *et al.*, 'l-DOPA and Freezing of Gait in Parkinson's Disease: Objective Assessment through a Wearable Wireless System', *Front. Neurol.*, vol. 8, 2017, doi: 10.3389/fneur.2017.00406.

[6] N. 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', *Front. Aging Neurosci.*, vol. 7, 2015, doi: 10.3389/fnagi.2015.00151.

[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*, 2015, doi: 10.1371/journal.pone.0123705.

[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',

arXiv:1612.07608 [cs], Dec. 2016, Accessed: Jul. 15, 2020. [Online]. Available: http://arxiv.org/abs/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, doi: 10.1016/j.csl.2013.12.001.

[10] A. J. Espay *et al.*, 'Technology in Parkinson's disease: Challenges and opportunities', *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, 2016, doi: https://doi.org/10.1002/mds.26642.

[11] E. Rovini, C. Maremmani, and F. Cavallo, 'How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review', *Front. Neurosci.*, vol. 11, 2017, doi: 10.3389/fnins.2017.00555.

[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 (English Edition)*, vol. 34, no. 1, pp. 38–54, Jan. 2019, doi: 10.1016/j.nrleng.2018.12.002.

[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, 2013, doi: https://doi.org/10.1002/mds.25628.

[14] B. M. Bot *et al.*, 'The mPower study, Parkinson disease mobile data collected using ResearchKit', *Sci Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.11.

[15] A. Zhan *et al.*, 'Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity', *JAMA Neurol*, vol. 75, no. 7, pp. 876–880, Jul. 2018, doi: 10.1001/jamaneurol.2018.0809.

[16] S. Arora *et al.*, 'Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study', *Parkinsonism & Related Disorders*, vol. 21, no. 6, pp. 650–653, Jun. 2015, doi: 10.1016/j.parkreldis.2015.02.026.

[17] D. Joshi, A. Khajuria, and P. Joshi, 'An automatic non-invasive method for Parkinson's disease classification', *Computer Methods and Programs in Biomedicine*, vol. 145, pp. 135–145, Jul. 2017, doi: 10.1016/j.cmpb.2017.04.007.

[18] H. H. Manap, N. Md Tahir, and A. I. M. Yassin, 'Statistical analysis of parkinson disease gait classification using Artificial Neural Network', in *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec. 2011, pp. 060–065, doi: 10.1109/ISSPIT.2011.6151536.

[19] 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, p. 6139716, 2017, doi: 10.1155/2017/6139716.

[20] K. Szewczyk-Krolikowski *et al.*, '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 Relat. Disord.*, vol. 20, no. 1, pp. 99–105, Jan. 2014, doi: 10.1016/j.parkreldis.2013.09.025.

[21] 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, doi: 10.1007/s00415-016-8384-9.

[22] S. Nazem *et al.*, 'Montreal cognitive assessment performance in patients with Parkinson's disease with "normal" global cognition according to mini-mental state examination score', *J Am Geriatr Soc*, vol. 57, no. 2, pp. 304–308, Feb. 2009, doi: 10.1111/j.1532-5415.2008.02096.x.

[23] M. M. Wickremaratchi *et al.*, 'The motor phenotype of Parkinson's disease in relation to age at onset', *Movement Disorders*, vol. 26, no. 3, pp. 457–463, 2011, doi: 10.1002/mds.23469.

[24] L. M. Shulman, R. L. Taback, J. Bean, and W. J. Weiner, 'Comorbidity of the nonmotor symptoms of Parkinson's disease', *Movement Disorders*, vol. 16, no. 3, pp. 507–510, 2001, doi: 10.1002/mds.1099.

[25] B. Pittman, R. H. Ghomi, and D. Si, 'Parkinson's Disease Classification of mPower Walking Activity Participants', in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 4253–4256, doi: 10.1109/EMBC.2018.8513409.

[26] P. Schwab and W. Karlen, 'PhoneMD: Learning to Diagnose Parkinson's Disease from Smartphone Data', *arXiv:1810.01485 [cs, q-bio]*, Nov. 2018, Accessed: Jun. 25, 2020. [Online]. Available: http://arxiv.org/abs/1810.01485.

[27] 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, 2019, doi: 10.1109/TBME.2018.2873252.

[28] S. Mehrang, M. Jauhiainen, J. Pietil, 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', *Conf Proc IEEE Eng Med Biol Soc*, vol. 2018, pp. 2913–2916, 2018, doi: 10.1109/EMBC.2018.8512921.

[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', *arXiv:1706.09574 [stat]*, Jun. 2017, Accessed: Jul. 14, 2020. [Online]. Available: http://arxiv.org/abs/1706.09574.

[30] E. Chaibub Neto *et al.*, 'Detecting the impact of subject characteristics on machine learning-based diagnostic applications', *npj Digital Medicine*, vol. 2, no. 1, Art. no. 1, Oct. 2019, doi: 10.1038/s41746-019-0178-x.

[31] E. C. Neto *et al.*, 'Learning Disease vs Participant Signatures: a permutation test approach to detect identity confounding in machine learning diagnostic applications', *arXiv:1712.03120 [stat]*, Jul. 2018, Accessed: Jul. 17, 2020. [Online]. Available: http://arxiv.org/abs/1712.03120.

[32] M. Giuliano, A. García-López, S. Pérez, F. D. Pérez, O. Spositto, and J. Bossero, 'Selection of voice parameters for Parkinson´s disease prediction from collected mobile data', in *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, Apr. 2019, pp. 1–3, doi: 10.1109/STSIVA.2019.8730219.

[33] J. Prince and M. de Vos, 'A Deep Learning Framework for the Remote Detection of Parkinson'S Disease Using Smart-Phone Sensor Data', *Conf Proc IEEE Eng Med Biol Soc*, vol. 2018, pp. 3144–3147, 2018, doi: 10.1109/EMBC.2018.8512972.

[34] R. Martinez-Mendez, M. Sekine, and T. Tamura, 'Postural sway parameters using a triaxial accelerometer: Comparing elderly and young healthy adults', *Computer Methods in Biomechanics and Biomedical Engineering*, 2012, doi: 10.1080/10255842.2011.565753.

[35] K. Seifert and O. Camacho, 'Implementing Positioning Algorithms Using Accelerometers', p. 13.

[36] K. E. Lyons, R. Pahwa, and R. Pahwa, *Handbook of Essential Tremor and Other Tremor Disorders*. CRC Press, 2005.

[37] L. Palmerini, L. Rocchi, S. Mellone, F. Valzania, and L. Chiari, 'Feature selection for accelerometer-based posture analysis in Parkinsons disease', *IEEE Transactions on Information Technology in Biomedicine*, 2011, doi: 10.1109/TITB.2011.2107916.

[38] A. Zhan *et al.*, 'High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection', *arXiv:1601.00960 [cs]*, Jan. 2016, Accessed: Jun. 25, 2020. [Online]. Available: http://arxiv.org/abs/1601.00960.

[39] 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 and Neural Repair*, 2011, doi: 10.1177/1545968311424869.

[40] 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*, 2019, doi: 10.3390/electronics8020119.

[41] M. Bächlin *et al.*, 'Wearable assistant for Parkinsons disease patients with the freezing of gait symptom', *IEEE Transactions on Information Technology in Biomedicine*, 2010, doi: 10.1109/TITB.2009.2036165.

[42] 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 Transactions on Biomedical Engineering*, vol. 43, no. 9, pp. 956–966, Sep. 1996, doi: 10.1109/10.532130.

[43] 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', *Journal of The Royal Society Interface*, vol. 8, no. 59, pp. 842–855, Jun. 2011, doi: 10.1098/rsif.2010.0456.

[44] B. M. Bot, 'mPower: Public Researcher Portal'. https://www.synapse.org/#!Synapse:syn4993293/files/ (accessed Jun. 25, 2020).

[45] R. Tibshirani, 'Regression Shrinkage and Selection Via the Lasso', *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[46] L. Breiman, 'Random forests', *Machine learning*, vol. 45, pp. 5–32, 2001.

[47] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, 'Gene Selection for Cancer Classification using Support Vector Machines', *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan. 2002, doi: 10.1023/A:1012487302797.

[48] M. E. Tipping, 'Sparse Bayesian Learning and the Relevance Vector Machine', *Journal of Machine Learning Research*, vol. 1, no. Jun, pp. 211–244, 2001.

[49] S. Bind, A. K. Tiwari, and A. K. Sahani, *A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction*. .

[50] P. B. Brazdil, C. Soares, and J. P. da Costa, 'Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results', *Machine Learning*, vol. 50, no. 3, pp. 251–277, Mar. 2003, doi: 10.1023/A:1021713901879.

[51] E. C. Neto *et al.*, 'Using permutations to assess confounding in machine learning applications for digital health', *arXiv:1811.11920 [stat]*, Nov. 2018, Accessed: Jul. 17, 2020. [Online]. Available: http://arxiv.org/abs/1811.11920.

[52] M. Sahandi 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 | medRxiv'. https://www.medrxiv.org/content/10.1101/2020.12.16.20247122v1 (accessed Dec. 21, 2020).

[53] M. Memedi, T. Khan, P. Grenholm, D. Nyholm, and J. Westin, 'Automatic and objective assessment of alternating tapping performance in parkinson's disease', *Sensors (Switzerland)*, 2013, doi: 10.3390/s131216965.

[54] 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, doi: 10.1371/journal.pone.0158852.

[55] 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 Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1794–1802, Nov. 2015, doi: 10.1109/JBHI.2015.2450232.

[56] 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 Transactions on Neural Systems and Rehabilitation Engineering*, 2014, doi: 10.1109/TNSRE.2013.2265887.

[57] T. Khan, 'Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic', 2014.

[58] 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*

Transactions on Biomedical Engineering, vol. 59, no. 5, pp. 1264–1271, May 2012, doi: 10.1109/TBME.2012.2183367.

[59] A. Mirelman *et al.*, 'Fall risk and gait in Parkinson's disease: The role of the LRRK2 G2019S mutation', *Movement Disorders*, vol. 28, no. 12, pp. 1683–1690, 2013, doi: 10.1002/mds.25587.

[60] Brookes, Mike, *VOICEBOX: Speech Processing Toolbox for MATLAB*. 2016.

[61] B. M. Bot, *Sage-Bionetworks: mPower-sdata*. Sage Bionetworks, 2020.