Identification of kinematic biomarkers for self-reported Parkinson's disease symptoms

Ayala Matzner
Bar-Ilan University
Yuval El-Hanany
Bar-Ilan University
Izhar Bar-Gad (izhar.bar-gad@biu.ac.il)
Bar-Ilan University

Article

Keywords: Parkinson's disease, digital biomarker, accelerometer, machine learning, kinematic sensors

Posted Date: December 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2321844/v1

Abstract

Monitoring the motor fluctuations and the severity of symptoms over time in Parkinson's disease (PD) patients is crucial for quantifying the progression of the disease and the adjustment of personalized therapy. The widespread availability of wearable sensors enables remote tracking of patients and the development of digital biomarkers for motor-related symptoms derived from the kinematic data acquired from these devices. Despite the progress in remote monitoring of PD symptoms, most research has been conducted on controlled behavior in the clinic, which departs considerably from individual patients' everyday behaviors and daily routines. This manuscript describes our top-performing algorithm in the Biomarker & Endpoint Assessment to Track Parkinson's Disease DREAM Challenge, funded by the MJFF, for predicting self-labeled PD symptom severity from free-behavior sensor data. To account for the self-labeled nature of the dataset and to capture each patient's subjective perception, we applied personalized automatic prediction algorithms consisting of ensembles of multiple random forest models followed by a predictability assessment of each patient. The results highlight the gradual approach required to develop new solutions in this field and constitute an important step forward in generating automatic and semi-automatic techniques that can facilitate the treatment of PD patients.

Introduction

Parkinson's disease (PD) is a chronic, progressive, neurodegenerative disease. The main motor symptoms of PD are resting tremor, bradykinesia, akinesia, rigidity and postural instability [1–3]. The primary cause of the clinical manifestations of Parkinsonian motor symptoms is the death of midbrain dopaminergic neurons [4]. The cardinal treatment of PD is dopamine replacement therapy (DRT), based on the elevation of dopamine levels throughout the brain. Despite the efficacy of this medication, as the disease progresses, PD patients experience on-off medication cycles consisting of motor fluctuations. During the ON periods, the motor symptoms are controlled, whereas during the OFF periods, these symptoms reemerge [5,6]. The ON states are often complicated by dyskinesia, a severe side effect of DRT characterized by exaggerated and involuntary movements [7]. These motor fluctuations are partially managed with dose adjustment of medications. An accurate assessment of the patient's overall clinical state forms the basis of most clinical evaluations and allows for an efficient adjustable long-term therapy. The current standard for this assessment relies on either a clinician-based assessment, using the Movement Disorder Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[8], or a patient-based approach, using the patient's self-report descriptive diaries (e.g. Hauser Diary [9], Parkinson's symptom diary [10]). The clinician-based approach is performed periodically during clinical visits, which limits its temporal resolution and masks symptom fluctuations in between visits. The patient-based approach may potentially have a better temporal resolution; however, it is subjective, prone to patient bias, and requires the patient's cooperation over a long period of time, which often results in inadequate monitoring of the symptoms [11]. To overcome the limitations of both approaches and to enhance the management of PD symptoms, an objective ongoing measure for monitoring the objective state and the subjective feeling of the patient is required.

The recent advances and the growing prevalence of digital technologies such as wireless devices and wearable sensors can be harnessed to improve individualized health care. These technologies enable the ongoing monitoring of patients' health status during their daily lives over extended periods of time and outside the clinical environment. Continuous remote monitoring with digital health interventions can identify alterations in the patient's state of health which permits faster and better decision making in clinical care [12]. Digitalization in healthcare is evolving in many areas, including cardiac [13], blood pressure [14] and diabetes [15] monitoring, in addition to fall detection [16] and smoking cessation [17]. As a primarily movement disorder, the treatment of PD may significantly benefit from digital health technologies which can capture ongoing motor symptoms and their fluctuations over time [18–20]. Movement can reliably be tracked using common kinematic sensors embedded in numerous wearable devices, such as smartwatches and smartphones. Kinematic sensors use a variety of physical principles to measure linear accelerations (accelerometers), rate of rotation (gyroscopes) and direction (magnetometers) along and around the three spatial axes. These sensors, combined with machine learning algorithms, are increasingly being used for *human activity recognition* in healthy individuals [21–23], as well as for detecting abnormal behavior in various movement disorders [24–26], and particularly in PD [27–31]. However, studies to date have a number of limitations with respect to the devices and methodologies suggested to provide personalized evaluation and tracking of the clinical state of PD. In particular, PD tracking studies tend to be conducted in a controlled environment [29,32], which makes them poorly applicable to everyday settings. Second, the severity of the symptoms during the experiments is usually annotated by the experimenter or an accompanying clinician, which may not reflect the patient's subjective evaluations. Thus, to develop the next generation of clinical interventions for PD patients based on kinematic data from wearable devices motor fluctuations need to be assessed in ecological conditions and take the subjective perception of the patient into account.

The Dialog for Reverse Engineering and Methods (DREAM) challenges, a crowdsourcing effort to examine key questions in biology and medicine, have launched a series of challenges designed to help researchers identify ways to use kinematic data from wearable devices to monitor PD patients, with the end goal of propelling at-home monitoring of disease progression. The first challenge, the Parkinson's Disease Digital Biomarker DREAM challenge, was designed to accurately identify PD status and symptom severity using data collected during the performance of specific tasks, while being monitored by a clinician [33]. Subsequently, a new challenge, the Biomarker and Endpoint Assessment to Track Parkinson’s Disease (BEAT-PD) DREAM challenge, was launched to determine whether PD severity could be evaluated from passively collected kinematic data recorded using consumer wearables during the course of daily life. In this challenge, the participants were provided with raw kinematic data of PD patients recorded at home and were asked to predict the patients' self-reported medication state and symptom severity [34]. This article describes our top-scoring solution for this challenge.

Results

This study was based on the BEAT-PD DREAM challenge dataset, under the auspices of the Michael J. Fox Foundation for Parkinson's Research and Sage Bionetworks [35]. The dataset was sourced from two different studies; the CIS-PD and the REAL-PD (see Methods), and consisted of 28 patients (16 males, 12 females), aged 62.1 ± 9.5 (mean ± STD). Both studies collected mobile sensor data from patients with PD as they went about their daily lives. Each session consisted of the kinematic signals (CIS: accelerometer, REAL: accelerometer and gyroscope) and the patients' subjective scores of their PD symptoms on a severity scale ranging from 0 to 4. The symptoms included the on-off medication state (where a score of 0 refers to the fully ON state, and 4 to fully OFF) and the severity of the dyskinesia and tremor. In this study, not all symptoms' scores were available for all subjects, such that each symptom category contained a different subset of all subjects (on-off: 22 subjects, dyskinesia: 16 subjects, tremor: 19 subjects). The goal of the challenge was to generate a model capable of predicting the patients' subjective assessments of their symptoms using the kinematic data. The challenge was separated into three different sub-challenges, corresponding to the three symptom categories included in this study, and involved predicting the severity of on-off, dyskinesia and tremor, respectively.

Predictions in the BEAT-PD challenge were evaluated by the weighted mean square error (wMSE) score. Due to the wide range in the number of sessions scored by each subject (Fig. 1A), the MSE was first calculated for each separately:

$$(1)\, MSE_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(y_{i,k} - \hat{y}_{i,k})^2$$

where n_k corresponded to the number of test sessions for subject k, $y_{i,k}$ was the *i*th rated score for subject k, and $\hat{y}_{i,k}$ was the corresponding predicted score. The subject-specific MSEs were then combined by weighting by the square root of the number of each subject's sessions to generate the wMSE score:

$$(2)\, wMSE = \frac{\sum_{k=1}^{N}\sqrt{n_k}MSE_k}{\sum_{k=1}^{N}\sqrt{n_k}}$$

The approach presented in this manuscript (*team "HaProzdor"*) won second place in the on-off challenge and fourth place in the dyskinesia and tremor sub-challenges.

In this section, we describe our solution to the challenge. First, we present the dataset, with its limitations and difficulties. Then, we describe our modeling approach, with the rationale for each of the stages in the process. Finally, we discuss the objective performance of our model, and its performance compared to the other winning models.

Data exploration

The distribution of the scores of each of the symptoms was skewed towards lower scores, and the scores of all the sub-datasets in both the CIS-PD and the REAL-PD were highly unbalanced (Fig. 1B). Moreover, different patients displayed very different distributions of subjective symptoms (Fig. 1C). These different symptoms are interrelated in the sense that tremor is expected to occur during the off period, thus leading to a positive correlation between these scores whereas tremor (which is characteristic of reduced dopamine levels) and dyskinesia (which is characteristic of excess dopamine levels) do not typically occur together. However, the correlations between the scores were not as expected, since both the tremor-on-off and tremor-dyskinesia scores displayed mostly positive correlations (Fig. 1D). On the single-subject level, the correlations between the symptom scores varied greatly. For example, subject 1004 exhibited highly positive correlations between all symptom scores (Fig, 1D, red circle). This subject rated the exact same score for all categories in numerous sessions. By contrast, subject 1023 rated different symptoms in an uncorrelated manner (Fig. 1D, blue circle).

The kinematic data were collected passively throughout the course of the subjects' daily lives. Hence, a single session could contain intermingled periods of different dynamic and static activities (Fig. 2A). These periods were not pre-categorized into different behaviors; however, some actions were easy to identify. These included periods of static activity, which appear as an almost flat accelerometer signal, and walking, which has a typical pattern of 1–2 Hz oscillations. Although the kinematic signals during different behaviors could sometimes be differentiated, signals recorded during sessions with significant differences in their symptom ranking could exhibit similar behavior patterns, thus blurring the distinction between the sessions (Fig. 2B-C). To test the use of “canonical” knowledge about different symptoms, we focused on the tremor category, whose kinematic properties are known. Tremor is typically associated with 4–6 Hz oscillations[36], which typically appear as a peak in this frequency band during rest (i.e. with smaller lower frequency components typical of movement). The tri-axial raw accelerometer signals were merged using the Euclidean norm of the three single-axis signals minus 1 (ENMO) [37].

The power spectral density (PSD) of the ENMO in sessions with different tremor scores was compared. Some high tremor sessions exhibited a peak at 4–6 Hz (Fig. 3A right), whereas other sessions with a high tremor score had no corresponding peak in their PSD (Fig. 3A, middle). Moreover, some sessions ranked by this subject as having no tremor displayed an unexpected peak in the 4–6 Hz band (Fig. 3A, left). Consequently, we measured the tremor oscillatory fraction of the session, which we defined as the fraction of time in the session in which the ratio between the power in the 4–6 Hz (tremor) band and the surrounding 1–9 Hz band exceeded a threshold of 0.7 (Fig. 3B). The oscillatory fraction was then compared to the tremor score in the session (Fig. 3C), and for each subject, we measured the overall correlation between all their sessions and their corresponding oscillatory fractions (Fig. 3D). Unexpectedly, most of the subjects had low correlations, with some subjects (8 of 19, 42%) displaying negative correlations, suggesting that at least in some cases the reported scores represented the overall state as perceived by the subject rather than the score corresponding to the specific symptom.

The preliminary stage of exploration of the raw data highlighted important points that guided the way we developed the predictive models. First, the diversity of subjective scoring between patients emphasized

the need to formulate personalized models for each subject, rather than a single model for the whole cohort. Second, given the lack of stable correlations or consistency between the scores for different symptoms, we generated independent models for each symptom. Third, the scores of the sessions could not be interpreted simply based on the tailor-made feature arising from the raw kinematic signals. Thus, the extraction of complex logical features, such as the tremor oscillatory fraction, were not sufficient for the current dataset. Since simpler features (e.g., range, standard deviation, and mean values of the signal) are meaningless when extracted from a whole session, we divided each session into shorter segments prior to the extraction of a large collection of simple features as the input to the model. In addition, because the data were obtained during free behavior, there were major changes in movement patterns as the subject performed various tasks throughout the sessions. Thus, extracting features from semi-stable short segments better reflected individual activities. During the long 20-minutes session, symptoms could wax and wane, such that short stretches of a symptom could affect the final rating ascribed by the subject. This resulted in feature values that differed throughout the session, which were crucial to the overall prediction. To account for this, our initial preprocessing step involved splitting each session into 10 second segments with an overlap of 5 seconds (see Methods; different segment sizes and overlaps were tested). Fourth, the kinematic data were segmented such that the symptom report was placed in the middle of the session (approximately 10 minutes). After the report time, the session continued for another 10 minutes, during which the severity of the symptoms could change. Given that the second half of the session might provide misleading information about the session score, most of our models only used the first half of the session.

Overview of the experimental design for predicting symptom severity

The general framework of our analysis is shown in Fig. 4 which illustrates the sequential process containing the nested verification (Fig. 4). Sessions were first segmented into short overlapping segments. A large set of features was calculated on each of the segments using temporal, spectral and dimensionality reduced information, and the top ranked features were selected for the model. An ensemble of different variants of random forest (RF) classifiers and regressors were used on the segment features using multiple training/test splits. The best model in each split was used to predict the scores of the test data, and a final prediction score was calculated as the average of all the split predictions. Finally, each subject's predictability was assessed, and the test scores of subjects with low predictability scores were replaced by the subject's naïve mean score.

Feature selection

A large set of features was constructed from each segment (CIS-PD: 368 features, REAL-PD: 469 features. see Methods). This large number of features could have led to an overfitting problem which would have degraded the performance of the model [21,38]. Thus, we performed a feature selection process on the subset that was utilized by the model (see Methods). The top selected features for individual subjects varied across symptoms (Fig. 5A). This is because a feature may be predictive of one category, in terms

of distinguishing between different scores (Fig. 5B, left), whereas for a different category, this same feature will not necessarily differentiate scores well (Fig. 5B, right). Thus, to obtain predictive features, we carried out the selection process separately for each symptom. We next examined commonly selected features across subjects for the same symptom category. We ran the feature selection process for each subject over multiple training/test splits. The ranking of all features for a given symptom were similar between all splits of the same subject, such that the same features tended to be ranked with high values over the different sub-datasets (Fig. 5C), and there was a substantial overlap between the top selected features in each (Fig. 5D). By contrast, the ranking of the features varied between subjects (Fig. 5E), and the overlap between the top selected features was significantly smaller (Fig. 5F). For example, the mean percent of the common tremor features between every ten splits of each CID-PD subject (71 ± 15%, mean ± STD) was significantly higher than the mean of the common tremor features between all pairs of CIS-PD subjects (59 ± 10%, mean ± STD, $p < 0.001$, Mann-Whitney U test). Therefore, as the base method, we conducted the selection process for each subject separately.

Ensemble- based approach to symptom severity prediction

The symptom score prediction approach consisted of the integration of predictions from multiple RF models that differed in terms of their hyperparameters. The optimal variants of the hyperparameters. Each subject/symptom data combination was split into 32 different training/test splits, and on each of the training splits, a set of RF models, differing in their hyperparameters, were trained (see Methods). The performance criterion was the fraction of splits leading the best performing model for all subjects (Fig. 6A). The results showed that the "best-combination" model was the most frequently best performing model in all categories but constituted only 30% of the cases. Analyzing the best performing models in individual cases revealed different distributions of selected models for different subjects (Fig. 6B) suggesting that there was no single optimal model that could be used for prediction, but rather that this depended on the specific subject/symptom combination. However, selecting a single model based on the given available training data may lead to overfitting and does not guarantee optimal generalization on the test data. Thus, we combined the predictions from the 32 best performing models generated for each train/test split into model was chosen based on the assessment of different an ensemble. The final prediction was calculated by averaging the ensemble predictions for each of the sessions, followed by a mean adjustment (Fig. 6C). We assumed that the usage of a multiple-model would optimize the output of single-type classifiers. However, a post-challenge analysis revealed that the use of the same single type model in all the ensembles was not necessarily worse than multiple-type models (Fig. 6D). Although the differences in performance were minor, in some cases, a single specific model could outperform multiple-type models probably due to reduced overfitting.

Cherry-picking: assessing the predictability of individual subjects

Validation of our models using nested 5-fold cross-validation (see Methods) indicated that the predictability of some subjects was low (Fig. 7A-C). Subjects whose scores could not be predicted reliably were thus defined as naïve subjects, and their final score prediction was replaced by the naïve mean. The proportion of naïve subjects varied across symptoms (on-off: 16/22, dyskinesia: 8/16, tremor: 8/19), and different subsets of naïve subjects were found in each of the categories (on-off: 1006, 1044, hbv013, hbv051, hbv077, hbv043; dyskinesia: 1007, 1023, 1034, 1043, 1044, 1048, hbv054, hbv018; tremor: 1007, 1023, 1034, 1046, 1048, hbv013, hbv054, hbv012). However, despite the division into naïve and non-naïve subjects, the mean improvement in the model relative to the naïve MSE of non-naïve subjects was small (on-off: 0.08 ± 0.07, dyskinesia: 0.07 ± 0.06, tremor: 0.07 ± 0.06, mean ± STD).

BEAT-PD Dream challenge results

Forty-three different teams submitted their predictions to this challenge. The teams were ranked according to their wMSE scores (Fig. 8A) and compared to the null model (on-off: 0.967, dyskinesia: 0.4373, tremor: 0.4399). Our submission was awarded second place in the on-off sub-challenge (wMSE = 0.8793), and fourth place in the dyskinesia (wMSE = 0.4205) and the tremor (wMSE = 0.404) sub-challenges. Although the final wMSE scores of the top-performing teams (HaProzdor, dbmi, ROC BEAT-PD, yuanfang.guan, hecky and Problem Solver) were very close, the correlations between the test predictions of these groups were low (supplementary Table 1). Though our models were top performing in this challenge in terms of low wMSE scores, the variance explained by the models for each subject in each of the sub-challenges was low (Fig. 8B).

Discussion

Advances in wearable sensors have laid the groundwork for developing digital biomarkers and measures for the remote monitoring of motor fluctuations during PD. In this study, we presented an ensemble-based approach for predicting PD patients' subjective perception of their symptoms based on kinematic sensor data embedded in smartwatches. The behavioral data were unstructured and collected passively from the patients during their everyday lives in their natural environments. The data were segmented into overlapping 10-second segments, and a large set of simple features combining temporal, spectral and dimensionality-reduction features were extracted from each segment. The features were then fed into an ensemble of RF models, and the ensemble predictions were combined and averaged to generate a final integrated prediction. Due to the diversity in subjective scoring across subjects, the ensemble models were generated individually for each subject. Using nested validation, we showed how the predictability level varied across subjects. For this reason, the test scores of subjects with low predictability were replaced by their naïve scores. The test predictions produced by our models achieved low wMSE scores in all three symptoms categories. However, their predictive values were comparatively small and patient-specific.

Many studies of movement disorders have used signal processing techniques to extract features from kinematic data, where different machine-learning classifiers were then applied to the extracted features to

predict movement-related characteristics [29,30,39,40]. In line with these studies, five out of the top six performing teams, including ours, approached the current challenge by applying machine learning models on the extracted features, whereas only one of the top teams applied a featureless deep learning model. Similar to our approach, all of the five machine learning solutions were based on a large set of simple features and decision tree algorithms, and most of them used RF models. Another characteristic shared by most of the top solutions was the generation of personalized models for each subject, rather than a global model. The differences between the final wMSE scores of the top-performing solutions were minimal. However, unexpectedly, the correlations between the test predictions of these teams were low. This suggests a possible lower bound on performance in the current dataset. The decision of most teams to use tree-based models is not surprising for this challenge. Tree-based methods are well-suited for complex or non-linear datasets like the current one, and the combination of multiple trees using RF reduces the variance and avoids overfitting [41]. We tested other approaches, such as replacing the simple features with hand-crafted features, or using machine-learning models other than RF, but these led to higher wMSE results. Moreover, in the two of the four intermediate rounds preceding the final submission, we included deep neural network models in our solution ensemble, but pure RF ensembles based on the simple features outperformed all of them.

Although our approach compared favorably to the other submissions in this challenge, its objective predictive value was not high. We faced multiple hurdles in the current dataset originating from both the nature of the dataset, and the configurations of the challenge. The dataset was sparse, containing a relatively small number of scored sessions per subject. This impeded the usage of deep learning algorithms, which have been implemented successfully for capturing motor states during PD [42–44]. However, these models require a large amount of data and thus were not effective in this study. Moreover, the data were highly imbalanced, providing a small representation for most high severity scores. One of our RF variants included random under-sampling for class-imbalance correction. However, this model was inferior to most of the other models which did not include a rebalancing step (Fig. 6). Moreover, in intermediate rounds of the challenge, we also tested over-sampling techniques for balancing the classes, such as SMOTENN (data not shown), which did not improve the performance either. An additional challenge was the lack of a "ground truth" due to the dataset's self-labeled nature. An objective assessment by a PD specialist of at least a subset of the sessions alongside the subject-based scoring could improve the models' predictability considerably. This could be done by either the presence of a clinician during the session, or offline using video devices to monitor the patients and establish their symptom severity baseline. The configurations of the challenge also made the prediction difficult since some existing metadata were hidden. This included the sessions' temporal order, time of day, the sensor location and other kinematic signals, such as gyroscope data missing in the CIS-PD dataset, and magnetometer data. All these data and metadata could significantly improve the performance of the predictors [45,46]. In order to enhance predictability in future studies, these should be utilized. Collecting robust amounts of balanced scored sessions per subject accompanied by a clinician assessment, and in conjunction with richer metadata would provide the basis for a more accurate assessment of PD symptoms.

The subjective subject-based ratings in this study were highly specific. In the absence of associated expert baseline ratings, the data reflected the subjective feelings and individual perceptions, which differed substantially across subjects [47,48]. Thus, algorithms transferring information across subjects were not useful, and the generation of personalized models was required. Beyond capturing the symptom-related features, the models "learned" the rating strategies of each subject separately. However, the predictive value of these personalized models varied across subjects: for some subjects, the performance of all of our tested models was below the naïve mean prediction. At least in some cases, this was due to low reliability which was manifested as low self-consistency. For example, some subjects tend to rate all symptoms equally, or subjects graded the symptoms according to their overall feeling; thus, the rating was not always consistent with the symptoms' objective severity. This variability in self-consistency is not necessarily related to PD: differences in self-reporting consistency have been explored in the field of physical activity, and were shown to be associated with multiple factors such as gender, age, educational and marital status [49,50]. In this study, however, the models' success was not only specific to the subject but was also attributed to the specific symptom category. Although some subjects had low predictability in two out of the three symptoms, no subject was unpredictable for all three symptoms. The ranking formulation assessment in this challenge, which used MSE, led to the prioritization of scores at or near the naïve mean. Thus, to avoid costly losses, we implemented cherry picking, by replacing the low-predictability subjects' predictions with their naïve mean. This strategy positioned us among the top-performing teams in all three categories. Nevertheless, the application of cherry-picking is clearly not a practical approach in applications beyond this competition.

The data in this study were acquired from patients engaged in their everyday lives in their natural environment. This is a significant advance over most previous PD tracking studies that have been conducted in a controlled environment. In these studies, in addition to the symptom rating by an expert, the subjects typically performed a specific subset of motor tasks. However, these well-controlled tasks do not always reflect motor fluctuations and impairment during everyday life activities. Thus, collecting data passively in ecological conditions is an essential step towards the remote monitoring of PD progression [33,51]. This study attempted to make this huge leap from the classification of a small set of behaviors to complete unrestricted free behavior. However, as our results show, success is partial, and we assume that this is, in part, inherent to the nature of the rapid shift to uncontrolled data. PD-related motor fluctuations are embedded in the overall kinematic signal, which varies enormously as a function of different activities such as walking, running and sitting. The models are blind to the type of behavior, and their attempt to generalize across activities resulted in a relatively low predictability value. Thus, a more gradual research progression is needed to bridge the gap between clinic-based and free-living PD tracking studies, and an intermediate stage of uncontrolled PD activity identification accompanied by careful objective behavior annotation either online or using video offline is essential.

Overall, this study constitutes an important step in the development of remote monitoring of everyday PD symptom severity through wearable sensors. In-home monitoring can improve the effectiveness of PD therapy, wherein a relatively short time, clinicians can access reliable accurate data on the severity of the

patient's symptoms. Progress in remote monitoring can also provide useful information for PD researchers. Our results indicate some ability to predict the subjective severity of symptoms during uncontrolled daily activities. However, a more gradual approach is required before there is widespread clinical adoption of wearable digital technologies.

Material And Methods

Data Collection

The data were obtained from the Biomarker & Endpoint Assessment to Track Parkinson's disease DREAM Challenge. The kinematic data were collected from two different studies: The Clinician Input Study (CIS-PD)[52] and the REAL-PD (Parkinson@Home) validation study [53]. The CIS-PD dataset was generated by researchers at Northwestern University, the University of Rochester, the University of Alabama, and the University of Cincinnati. During the 6-month long study PD patients wore Apple watch devices during clinic visits and for at-home monitoring. The subjects were asked to report their PD symptoms at 30-minute intervals for the 48 hours before each clinic visit. The symptoms included on-off medication state, dyskinesia, and tremor, and were rated on a severity scale ranging from 0 to 4, with 4 as the most severe. The accelerometer data from the Apple watches were segmented into 20-minute sessions, corresponding to the symptom report, where the report time was situated in the middle of the session.

The REAL-PD study of wearable tracking for PD patients was generated by researchers at the Radboud University Medical Center (NL). In this 2-week long study the patients were provided with a Motorola watch, in addition to their own Android phone, both of which were used to monitor their activity at home. For two days during the study, the subjects were asked to report their symptoms in 30-minute time blocks throughout the day. The subjects rated their tremor on a severity scale from 0–4, and reported their medication status, which was then transformed into an on-off scale of 0–1 and a dyskinesia scale of 0–2.

The kinematic data were segmented relative to the report times into 20-minute recording sessions by taking the central 20 minutes of the time range. To simplify utilizing the two datasets together, we only considered the smartwatch data from the REAL-PD study.

The two studies provided 3-axis linear acceleration signals collected from the smartwatches worn by the patients. The REAL-PD study also included 3-axis rotation rate information collected by the smartwatches' gyroscope sensor. All the data were sampled at 50 samples/second. The subjects contributed a total of 3255 scored sessions to the dataset, most of which were 20 minutes long with a few shorter sessions (272 sessions, 8%). The session data files were standardized to a start time of 0 to remove time-of-day information as well as information about the temporal order of the sessions. Out of the whole sessions, 2791 (86%) were rated with on-off scores, 1926 (59%) with dyskinesia scores and 2362 (72%) with tremor scores. The dataset were partitioned into training (2449 sessions) and testing (806 sessions) datasets. The training/test partitions were sampled randomly for each subject, thus

allowing for the construction of personalized symptom severity models. The training data scores were provided to generate a supervised model to predict the severity of symptoms in the test data.

For up-to-date information on the study, visit:

https://www.synapse.org/#!Synapse:syn20825169/wiki/600405 .

Ethics

The CIS-PD study was sponsored by the Michael J. Fox Foundation for Parkinson's Research and conducted across four US sites: Northwestern University, the University of Cincinnati, the University of Rochester, and the University of Alabama at Birmingham. Each site had local Institutional Review Board (IRB)/Research Ethics Board (REB) approval, and all participants signed informed consent. The REAL-PD study was sponsored by the Michael J. Fox Foundation for Parkinson's Research. The study protocol was approved by the local medical ethics committee (Commissie Mensgebonden Onderzoek, region Arnhem-Nijmegen, the Netherlands, file number 2016 – 1776). All participants received verbal and written information about the study protocol and signed a consent form prior to participation, in line with the Declaration of Helsinki (see also [34]) .

Data preprocessing

Data processing and analysis were performed using Python 3.6, except for the feature extraction which was done with Matlab 2017a (Mathworks). All the sensor data (CIS-PD - smartwatch accelerometer, REAL-PD - smartwatch accelerometer and gyroscope) were segmented into 10-second segments with 50% overlap. A total of 547,122 training and 181,582 testing segments were generated from this process. Accelerometer segments were separated into gravity and free acceleration components by applying a low/high pass filter (0.3 Hz fourth order Butterworth) respectively on the raw signals [54,55]. Each of the three axes in the free acceleration and angular velocity segments was derived temporally to obtain jerk signals. Next, the magnitude of each 3-dimensional signal was calculated using the Euclidian norm. These steps resulted in six time-domain separate signals obtained from each accelerometer segmented signals, and four separate time-domain signals obtained from the gyroscope signals. Each of these time-domain signals, excluding the gravity signals, was Fourier transformed to obtain the corresponding frequency-domain signals for each of the sessions.

Feature extraction and selection

A set of base features was calculated on each of the segment signals, resulting in the assignment of temporal and spectral features to each segment (Supplementary table 2). The CIS-PD dataset was composed of 343 base features, whereas the REAL-PD dataset was composed of 544 base features because of the additional gyroscope data. A second set of dimensionality reduction features was generated from the base features. These consisted of 5 PCA and 20 autoencoder new features that were added to each segment's features list. Before modeling, all the features were scaled to a range of [-1 1].

Feature selection was performed to keep the most relevant features. The selection was done by either F-value ranking, or random forest (RF) variable importance [56].

Random forest models

Seven variants of RF classifiers and regressors were used as prediction models for each subject/symptom combination data. RF models were constructed using the Scikit-learn Python package [57], with default settings. Note that these models outperformed other models tested by our group, including support vector machine, logistic regression and deep learning models. As the models were trained on segmented data windows, an aggregate prediction for the entire session was calculated by averaging the segments' predictions for that session. The models differed in their hyperparameters: (1) RF type: (a) classifier (b) regressor, (2) Model trained on features extracted from (a) full session or (b) first half session prior to the subject scoring, (3) Feature selection method: (a) F-value ranking (b) RF variable importance, (4) Number of top selected features: (a) 100 or (b) 50, and the (5) Class imbalance correction (a) none (b) random under-sampling (Supplementary table 3).

All models were trained on each training dataset (subject/symptom combination), and their weighted mean squared error (wMSE) scores were calculated. From these results, another model variant was trained, which constituted the best combination of each of the hyperparameters in terms of the wMSE scores. Finally, the wMSEs of all seven model variants were compared, and the model that minimized the wMSE score was chosen as the best model type. This model was then used to predict the test data.

In order to reduce overfitting, each training data was split into 32 different 80/20 training/test splits, and the set of RF models was trained separately on each of the splits. For each split, the best performing model was chosen and used to predict the test data as described above. This resulted in an ensemble of 32 different predictions for each session, which were averaged to produce a single prediction for each session. The final predictions of the test sessions of each subject were then calculated by adjusting their mean to the naïve mean score of the subject's training data scores.

Nested K-fold cross validation predictability analysis

We generated 5-fold cross validation training/test pairs from the whole training data, while stratifying the subjects. Each validation set was nested within 10 sub- training/test splits and RF prediction ensembles were generated as described above. The MSE scores of each subject were calculated using the test data of the set. As a null model, we used the subject-specific naïve mean, calculated as the mean of the training set scores for each of the symptoms separately. Each subject's MSE score was compared to its naïve mean MSE score, and the predictability of a single subject for a specific symptom was calculated as the mean differences between the MSEs over all five validation sets. The predictability threshold was set to zero, and a subject who did not cross this threshold for a specific symptom was defined as a naïve subject for that symptom. The test predictions of a naïve subject for a specific symptom were replaced by the subject's naïve mean calculated from the whole training data.

Declarations

Acknowledgments

These data were generated by participants of The Michael J. Fox Foundation for Parkinson's Research Mobile or Wearable Studies. They were obtained as part of the Biomarker & Endpoint Assessment to Track Parkinson's Disease DREAM Challenge (through Synapse ID syn20825169) made possible through partnership of The Michael J. Fox Foundation for Parkinson's Research, Sage Bionetworks, and BRAIN Commons. The Biomarker & Endpoint Assessment to Track Parkinson's disease DREAM Challenge is funded by the Michael J. Fox Foundation for Parkinson's Research.

Data availability

The data sets used for the challenge are available through the BEATPD challenge site**:** https://www.synapse.org/beatpdchallenge

Code availability

Our entire code is available on GitHub: https://github.com/ibglab/BEATPD-HaProzdor

References

1. Dauer, W. & Przedborski, S. Parkinson's Disease. *Neuron* vol. 39 889–909 (2003).
2. Mink, J. W. The basal ganglia: Focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* 50, 381–425 (1996).
3. Jackson-Lewis, V., Blesa, J. & Przedborski, S. Animal models of Parkinson's disease. *Parkinsonism Relat. Disord.* 18 Suppl 1, S183-5 (2012).
4. Langston, J. W., Ballard, P., Tetrud, J. W. & Irwin, I. Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis. *Science* 219, 979–980 (1983).
5. Limousin, P. *et al.* Electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease. *N.Engl.J.Med.* 339, 1105–1111 (1998).
6. Rascol, O. *et al.* Limitations of current Parkinson's disease therapy. *Annals of Neurology* vol. 53 at https://doi.org/10.1002/ana.10513 (2003).
7. Bezard, E., Brotchie, J. M. & Gross, C. E. Pathophysiology of levodopa-induced dyskinesia: potential for new therapies. *Nat. Rev. Neurosci.* 2, 577–588 (2001).
8. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. *Mov Disord.* 18, 738–750 (2003).
9. Hauser, R. A. *et al.* A home diary to assess functional status in patients with Parkinson's disease with motor fluctuations and dyskinesia. *Clin. Neuropharmacol.* 23, 75–81 (2000).

10. Montgomery, G. K. & Reynolds, N. C. Compliance, reliability, and validity of self-monitoring for physical disturbances of parkinson's disease: The parkinson's symptom diary. *J. Nerv. Ment. Dis.* 178, 636–641 (1990).
11. Reimer, J., Grabowski, M., Lindvall, O. & Hagell, P. Use and interpretation of on/off diaries in Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 75, 396–400 (2004).
12. Murray, E. *et al.* Evaluating Digital Health Interventions: Key Questions and Approaches. *American Journal of Preventive Medicine* vol. 51 843–851 (2016)
13. Widmer, R. J. *et al.* Digital health interventions for the prevention of cardiovascular disease: A systematic review and meta-analysis. *Mayo Clin. Proc.* 90, 469–480 (2015).
14. Uhlig, K., Patel, K., Ip, S., Kitsios, G. D. & Balk, E. M. Self-measured blood pressure monitoring in the management of hypertension: A systematic review and meta-analysis. *Annals of Internal Medicine* vol. 159 185–194 (2013).
15. Liang, X. *et al.* Effect of mobile phone intervention for diabetes on glycaemic control: A meta-analysis. *Diabet. Med.* 28, 455–463 (2011).
16. He, Y., Li, Y. & Bao, S. Di. Fall detection by built-in tri-accelerometer of smartphone. in *Proceedings - IEEE-EMBS International Conference on Biomedical and Health Informatics: Global Grand Challenge of Health Informatics, BHI 2012* 184–187 (2012).
17. Whittaker, R., Mcrobbie, H., Bullen, C., Rodgers, A. & Gu, Y. Mobile phone-based interventions for smoking cessation. *Cochrane Database of Systematic Reviews* vol. 2016 (2016).
18. Pasluosta, C. F., Gassner, H., Winkler, J., Klucken, J. & Eskofier, B. M. An emerging era in the management of Parkinson's disease: Wearable technologies and the internet of things. *IEEE J. Biomed. Heal. Informatics* 19, 1873–1881 (2015).
19. Mirelman, A., Giladi, N. & Hausdorff, J. M. Body-fixed sensors for Parkinson disease. *JAMA - Journal of the American Medical Association* vol. 314 873–874 at https://doi.org/10.1001/jama.2015.8530 (2015).
20. Klucken, J., Kruger, R., Schmidt, P. & Bloem, B. R. Management of Parkinson's disease 20 years from now: Towards digital health pathways. *Journal of Parkinson's Disease* vol. 8 S85–S94 (2018).
21. Ahmed, N., Rafiq, J. I. & Islam, M. R. Enhanced human activity recognition based on smartphone sensor data using hybrid feature selection model. *Sensors (Switzerland)* 20, (2020).
22. Lima, W. S., Souto, E., El-Khatib, K., Jalali, R. & Gama, J. Human activity recognition using inertial sensors in a smartphone: An overview. *Sensors (Switzerland)* 19, (2019).
23. Ravi, N., Dandekar, N., Mysore, P. & Littman, M. L. Activity Recognition from Accelerometer Data. in *Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence* 1541–1546 (2005).
24. Grunerbl, A. *et al.* Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients. *IEEE J. Biomed. Heal. Informatics* 19, 140–148 (2015).

25. Arifoglu, D. & Bouchachia, A. Activity Recognition and Abnormal Behaviour Detection with Recurrent Neural Networks. *Procedia Comput. Sci.* 110, 86–93 (2017).
26. Zhang, S., Sneddon, A., Poon, S. K., Vuong, K. & Loy, C. T. A Deep Learning-Based Approach for Gait Analysis in Huntington Disease. *Stud. Health Technol. Inform.* 264, 477–481 (2019).
27. Pérez-López, C. *et al.* Assessing motor fluctuations in parkinson's disease patients based on a single inertial sensor. *Sensors (Switzerland)* vol. 16 at https://doi.org/10.3390/s16122132 (2016).
28. Salarian, A. *et al.* Gait assessment in Parkinson's disease: Toward an ambulatory system for long-term monitoring. *IEEE Trans. Biomed. Eng.* 51, 1434–1443 (2004).
29. Patel, S. *et al.* Monitoring motor fluctuations in patients with parkinsons disease using wearable sensors. *IEEE Trans. Inf. Technol. Biomed.* 13, 864–873 (2009).
30. Cancela, J. *et al.* A comprehensive motor symptom monitoring and management system: The bradykinesia case. in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10* vol. 2010 1008–1011 (Conf Proc IEEE Eng Med Biol Soc, 2010).
31. Salarian, A. *et al.* Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system. *IEEE Trans. Biomed. Eng.* 54, 313–322 (2007).
32. Hoff, J. I., Van Der Meer, V. & Van Hilten, J. J. Accuracy of Objective Ambulatory Accelerometry in Detecting Motor Complications in Patients with Parkinson Disease. *Clin. Neuropharmacol.* 27, 53–57 (2004).
33. Sieberts, S. *et al.* Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. *bioRxiv* 2020.01.13.904722 (2020).
34. Sieberts, S. K. *et al.* Developing better digital health measures of Parkinson's disease using free living data and a crowdsourced data analysis challenge. *medRxiv* 2021.10.20.21265298 (2021) doi:10.1101/2021.10.20.21265298.
35. BEAT-PD DREAM Challenge - syn20825169. https://www.synapse.org/#!Synapse:syn20825169/wiki/596118.
36. Parkinson, J. *An essay on the shaking palsy.* (Sherwood, Neely and Jones, 1817).
37. van Hees, V. T. *et al.* Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity. *PLoS One* 8, e61691 (2013).
38. Guyon, I. & Elisseeff, A. *An Introduction to Variable and Feature Selection*. *Journal of Machine Learning Research* vol. 3 (2003).
39. Zheng, H., Yang, M., Wang, H. & Mcclean, S. Machine learning and statistical approaches to support the discrimination of neuro-degenerative diseases based on gait analysis. *Stud. Comput. Intell.* 189, 57–70 (2009).
40. Najafi, B. *et al.* Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly. *IEEE Trans. Biomed. Eng.* 50, 711–723 (2003).
41. Breiman, L. Random forests. *Mach. Learn.* 45, 5–32 (2001).

42. Sigcha, L. *et al.* Deep learning approaches for detecting freezing of gait in parkinson's disease patients through on-body acceleration sensors. *Sensors (Switzerland)* 20, (2020).
43. Zhang, A. *et al.* Automated tremor detection in Parkinson's disease using accelerometer signals. in *Proceedings - 2018 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2018* 13–14 (Institute of Electrical and Electronics Engineers Inc., 2019).
44. Pfister, F. M. J. *et al.* High-Resolution Motor State Detection in Parkinson's Disease Using Convolutional Neural Networks. *Sci. Rep.* 10, 1–11 (2020).
45. Nyholm, D., Lennernäs, H., Johansson, A., Estrada, M. & Aquilonius, S. M. Circadian rhythmicity in levodopa pharmacokinetics in patients with parkinson disease. *Clin. Neuropharmacol.* 33, 181–185 (2010).
46. Suzuki, M., Mitoma, H. & Yoneyama, M. Quantitative Analysis of Motor Status in Parkinson's Disease Using Wearable Devices: From Methodological Considerations to Problems in Clinical Applications. *Parkinson's Disease* vol. 2017 (2017).
47. Ishihara, L. S. *et al.* Self-reported parkinsonian symptoms in the EPIC-Norfolk cohort. *BMC Neurol.* 5, 15 (2005).
48. Leritz, E., Loftis, C., Crucian, G., Friedman, W. & Bowers, D. Self-awareness of deficits in Parkinson disease. *Clin. Neuropsychol.* 18, 352–361 (2004).
49. Gorzelitz, J. *et al.* Predictors of discordance in self-report versus device-measured physical activity measurement. *Ann. Epidemiol.* 28, 427–431 (2018).
50. Dyrstad, S. M., Hansen, B. H., Holme, I. M. & Anderssen, S. A. Comparison of self-reported versus accelerometer-measured physical activity. *Med. Sci. Sports Exerc.* 46, 99–106 (2014).
51. Del Din, S., Godfrey, A., Mazzà, C., Lord, S. & Rochester, L. Free-living monitoring of Parkinson's disease: Lessons from the field. *Mov. Disord.* 31, 1293–1313 (2016).
52. Elm, J. J. *et al.* Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson's disease data. *npj Digit. Med.* 2, 1–6 (2019).
53. Raykov, Y. P. *et al.* Probabilistic modelling of gait for remote passive monitoring applications. 1–8 (2018).
54. Anguita, D., Ghio, A., Oneto, L., Parra, X. & Reyes-Ortiz, J. L. A Public Domain Dataset for Human Activity Recognition Using Smartphones. in *ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* 437–442 (2013).
55. Lara, Ó. D. & Labrador, M. A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutorials* 15, 1192–1209 (2013).
56. Rogers, J. & Gunn, S. Identifying feature relevance using a random forest. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 3940 LNCS 173–184 (Springer Verlag, 2006).

57. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).

Figures

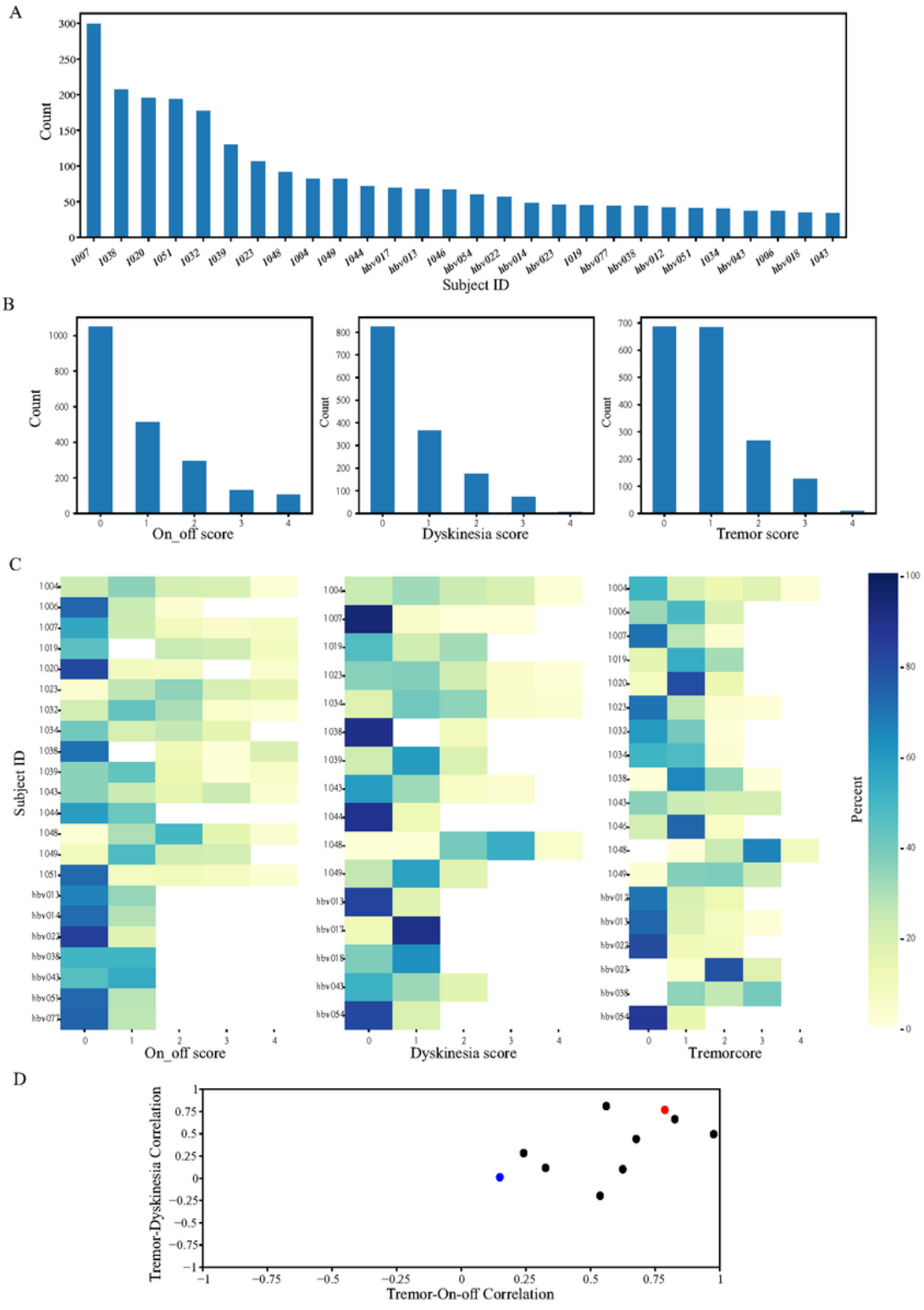

Figure 1

Data exploration: score distributions. (A) Number of scored training sessions provided by each subject. (B) Score distributions across all subjects for the on-off (left), dyskinesia (middle) and tremor (right) categories. (C) Score distributions for single subjects for on-off (left), dyskinesia (middle) and tremor (right) categories. White values represent zero values (the subject did not rate this score). (D) The relation between correlations of tremor-on-off and tremor-dyskinesia scores for single subjects. Only subjects having scores for all three categories in all sessions are displayed. Red circle corresponds to CIS-PD subject 1004, and blue circle to CIS-PD subject 1023.

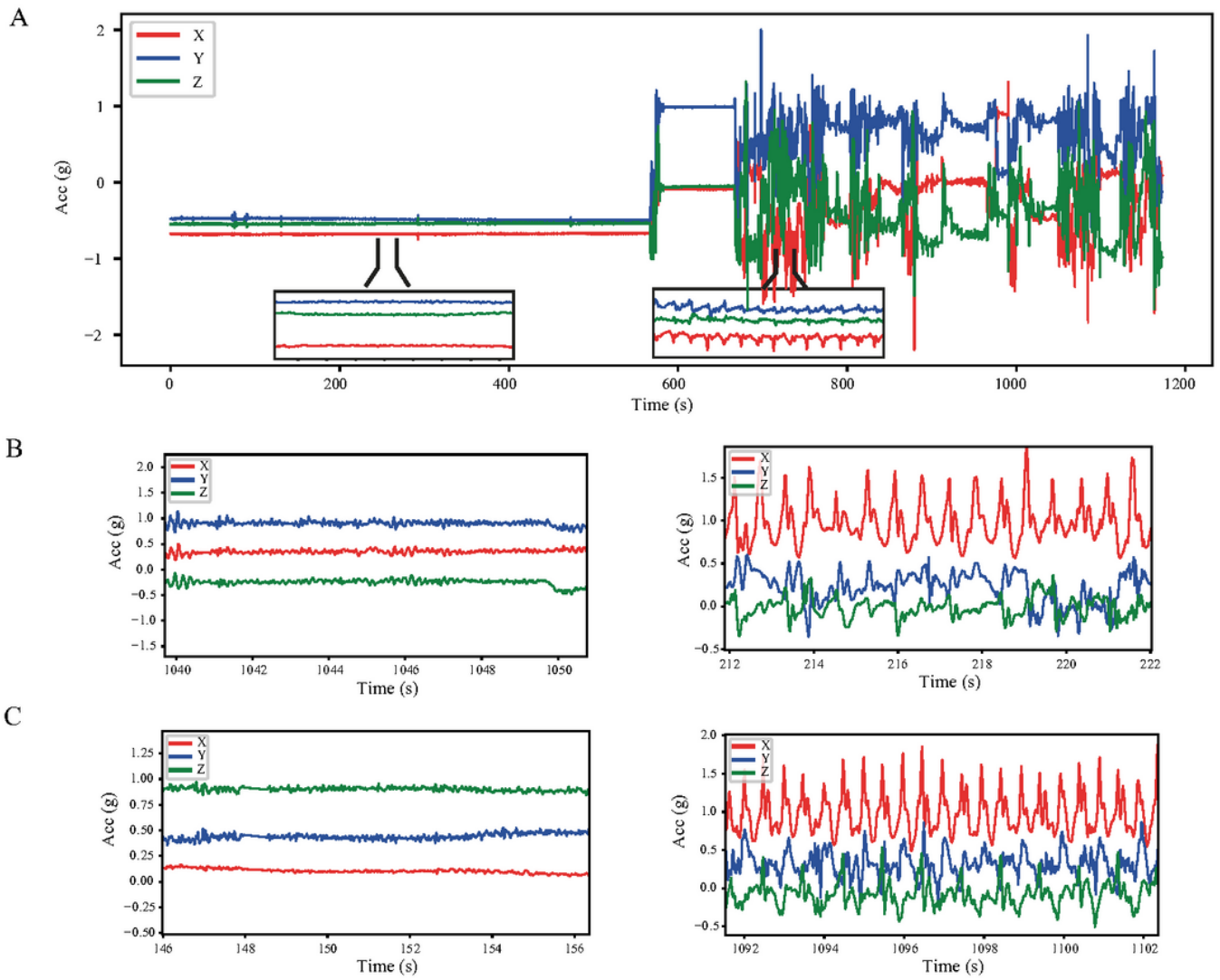

Figure 2

Data exploration: kinematic signals. (A) An example of raw acceleration data from three orthogonal axes during a free behavior session. Insets: enlargements of traces during rest (left) and walking (right). (B-C) Examples of 10 second accelerometer recording segments depicting rest (left) and walking (right) taken from a single session with an on-off score of (B) 0 and (C) 3.

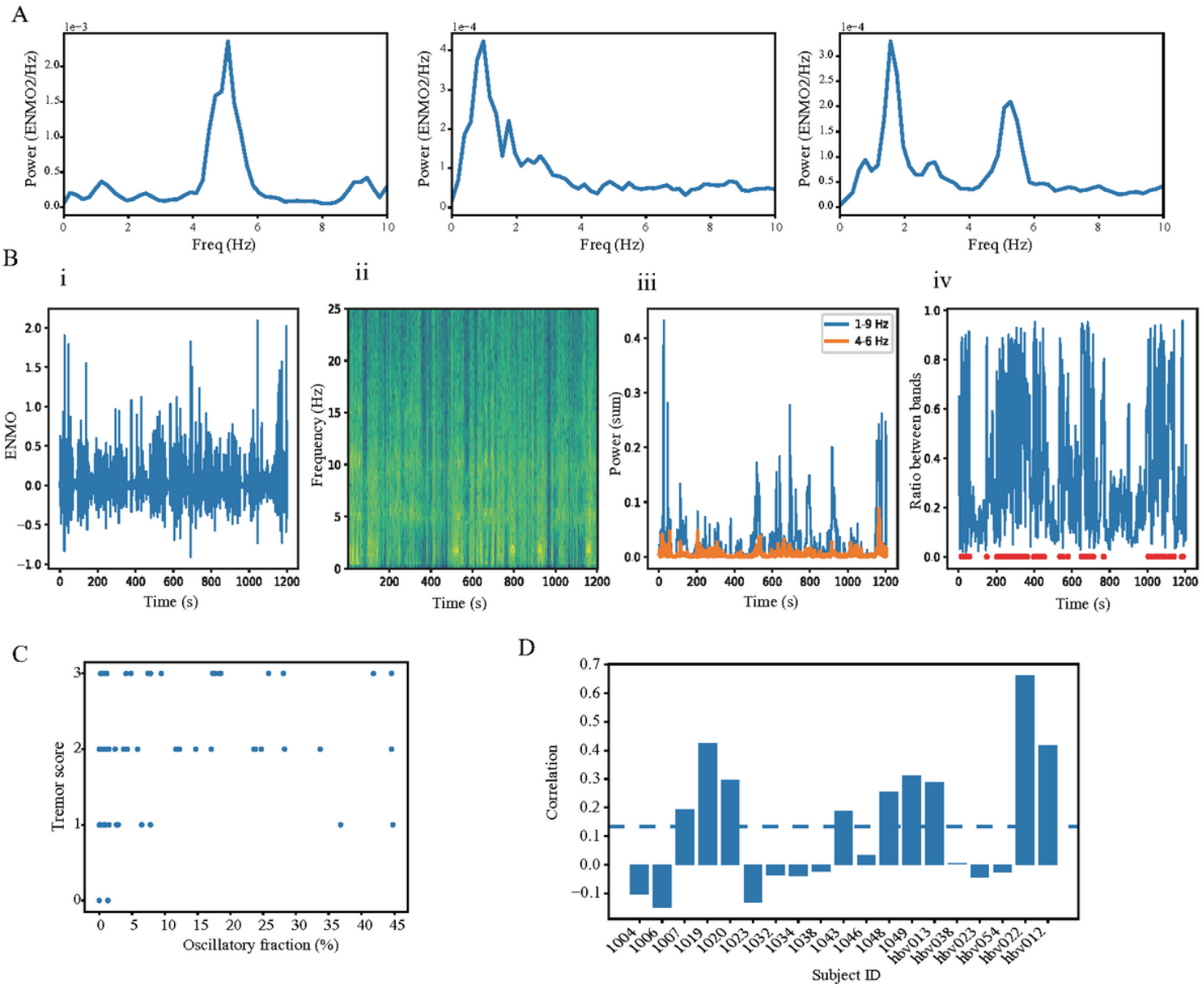

Figure 3

Complex features. (A) Examples of the PSD of the accelerometer signals (ENMO) during sessions rated with a tremor of 3 (left), 4 (middle) and 0 (left). (B) The process of extracting a tremor oscillatory fraction in a single session: the raw ENMO signal is extracted from the 3-axis accelerometer signal (i). The spectrogram of the ENMO signal is calculated (ii), and the power in the 1-9Hz band and 4-6Hz band are summed (iii). The ratio of the 4-6Hz band to the 1-9Hz band is calculated for each timestamp (iv, solid blue line), and oscillatory timestamps are defined as the time when the ratio crosses the threshold of 0.7 (iv, red dots). The oscillatory fraction is then defined as the fraction of time in the session in which the ratio crossed the threshold. In this example, the oscillatory fraction was 17.29%. (C) An example of the relationship between the tremor oscillatory fractions and tremor patient scores in all sessions for a single

subject. (D) Correlation between tremor scores and oscillatory fractions across subjects. The horizontal dashed line indicates the mean correlation across subjects.

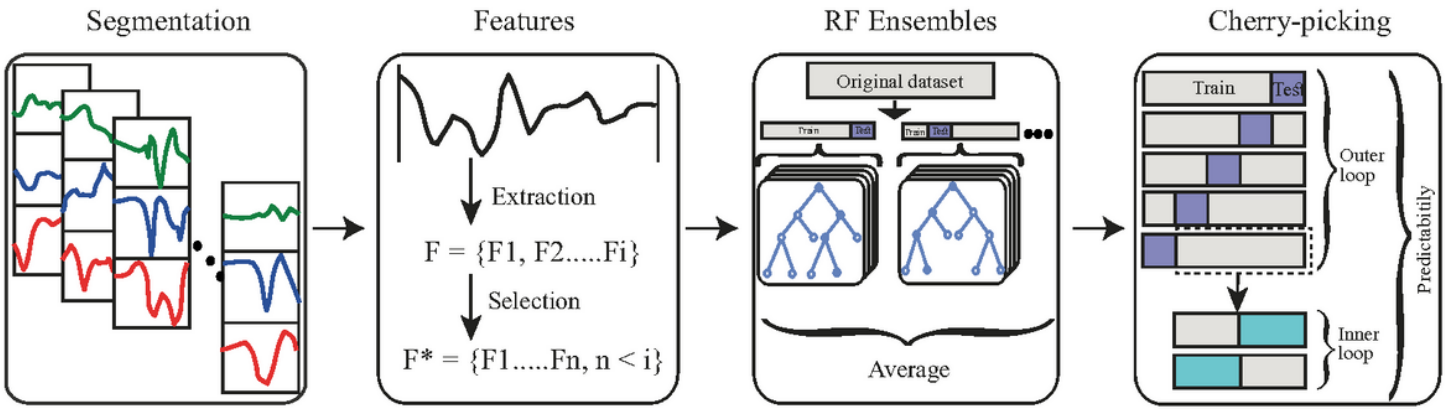

Figure 4

Overview of the model pipeline. Each session is segmented into 10-second overlapping segments, and multiple features are calculated on each of the segments. The top ranked features are selected for the model. These features are fed into an ensemble of RF classifiers using multiple training/test splits, and the ensemble predictions are averaged to generate a single final prediction. Finally, a cherry-picking process is applied based on the predictability assessment of each subject.

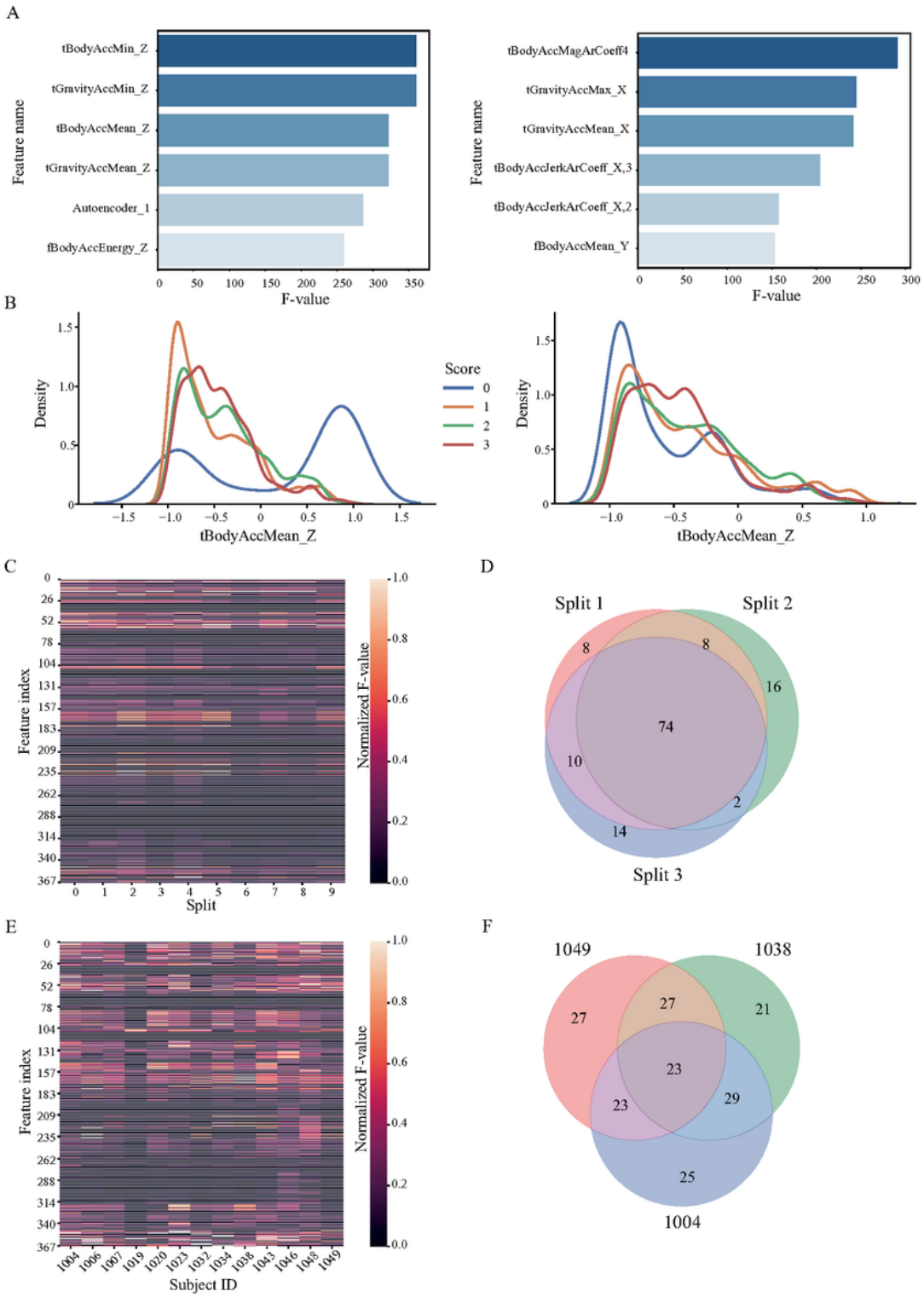

Figure 5

Feature selection. (A) Feature importance map showing the top 6 features of a single subject (CIS-PD 1049), selected according to the highest F-values. The features are graded separately for tremor (left) and on-off (right) symptoms. (B) Distribution of the values of the top-ranked feature for tremor from the example in *A*(mean of Z-axis body accelerometer) over the different scores for tremor (left) and on-off (right) of a single subject. (C) Normalized F-values of all features for the tremor category of a single

subject. The features are rated separately over 10 different randomly split sub-datasets, each containing 80% of the training data. (D) Overlap of the top 100 features, graded with the F-value, in three different splits. (E) Normalized F-values of all features for the tremor category of all CIS-PD subjects. The different colors represent individual subjects. (F) Overlap of the top 100 features for three different subjects.

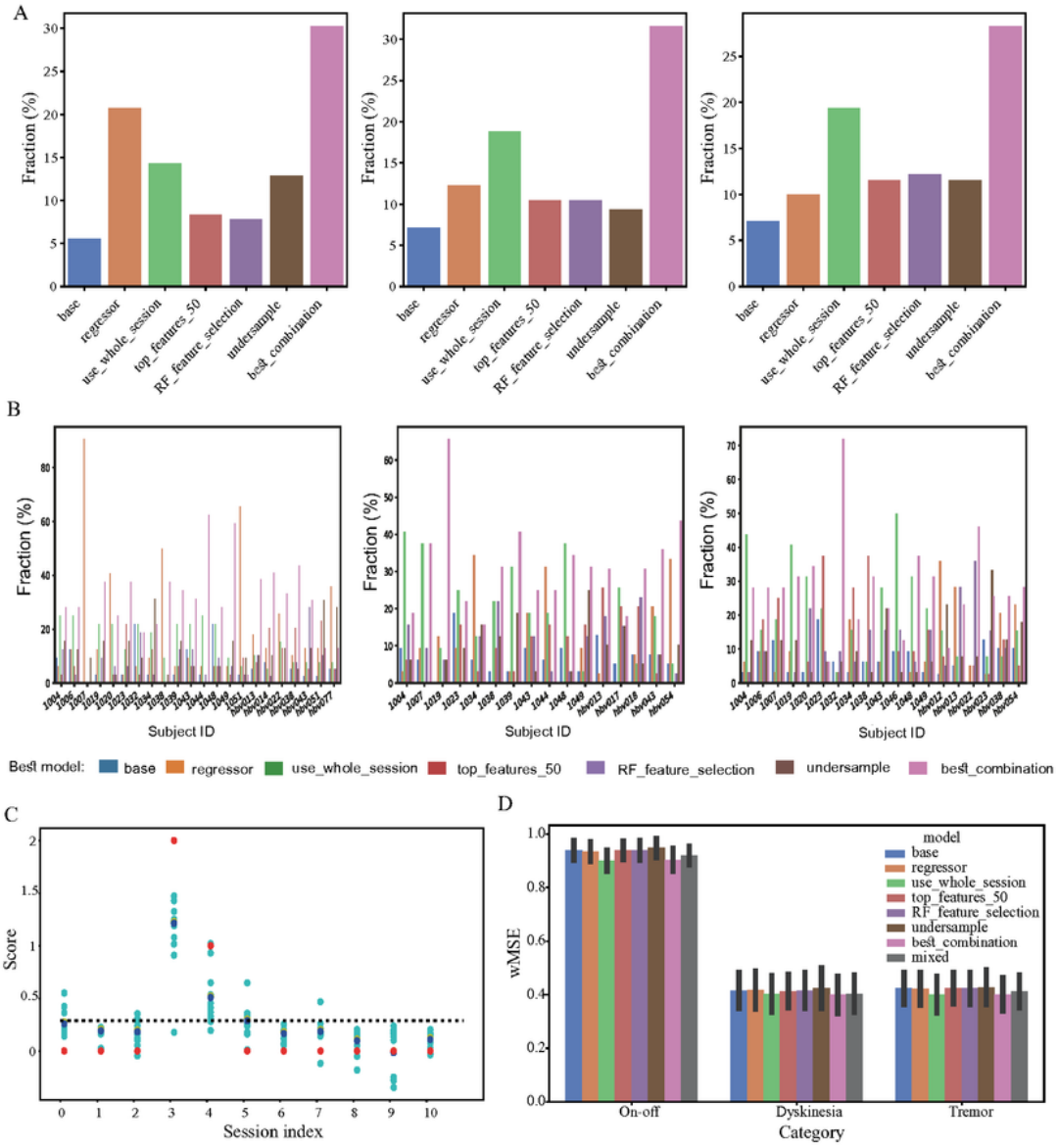

Figure 6

Ensembles of random forest models. (A) Fraction of the selected models in the different splits of the training data of all subjects, for on-off (left), dyskinesia (middle) and tremor (right) symptoms. (B) Same as *A*, divided into single subjects. (C) An example of the session prediction process for the tremor category of a single subject in a validation set. For each session, the ensemble process generates multiple score predictions (light blue circles). The predictions are averaged (yellow circles), and the final predictions (blue circles) are calculated by adjusting the average predictions to the mean of the subject (dashed line). In the validation set, the predictions can be compared to the true scoring of the subject (red circles). (D) Comparison of mean wMSE scores (± STD) of all validation sessions, across the 5 folds of the cross-validation datasets, using ensembles of predictions generated from different model variants (termed "mixed", grey bar), chosen separately in each training split, to ensembles generated from the same model in each of the splits.

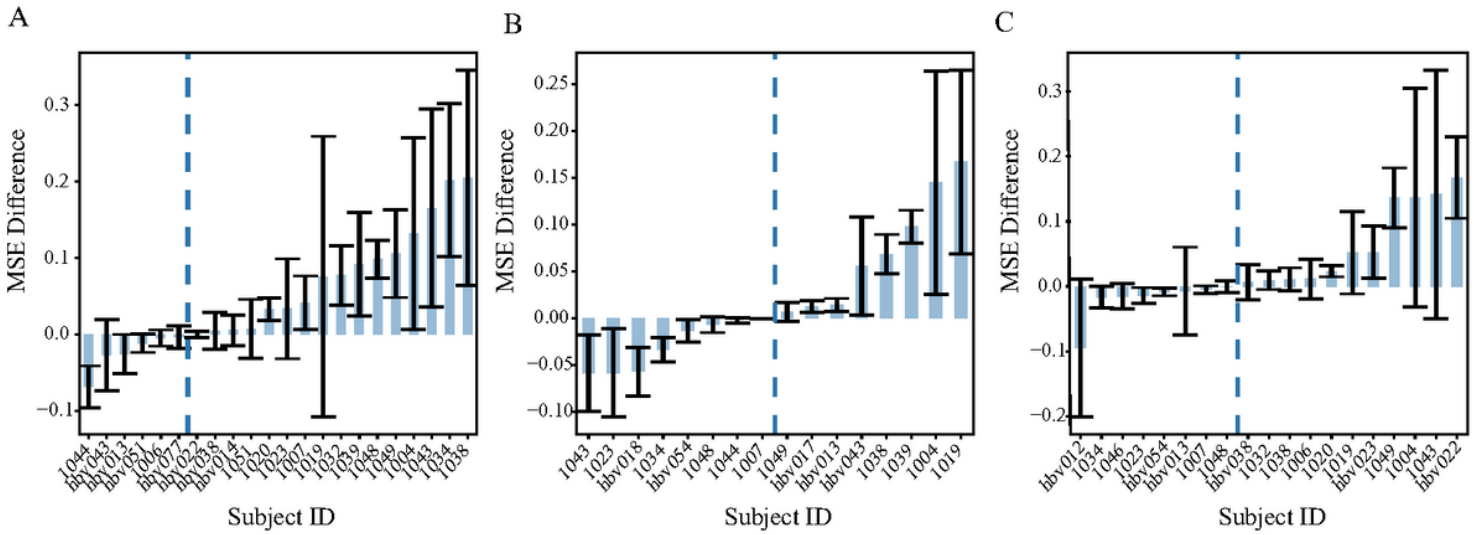

Figure 7

Cherry-picking: **classifying naïve subject.** (A-C) The mean differences (±SEM) between the naïve score and the MSE of each subject, across all five validation sets, for the on-off (A), dyskinesia (B) and tremor (C) categories. Vertical dashed lines represent the separation between naïve subjects, who did not cross the threshold of 0, and non-naïve subjects.

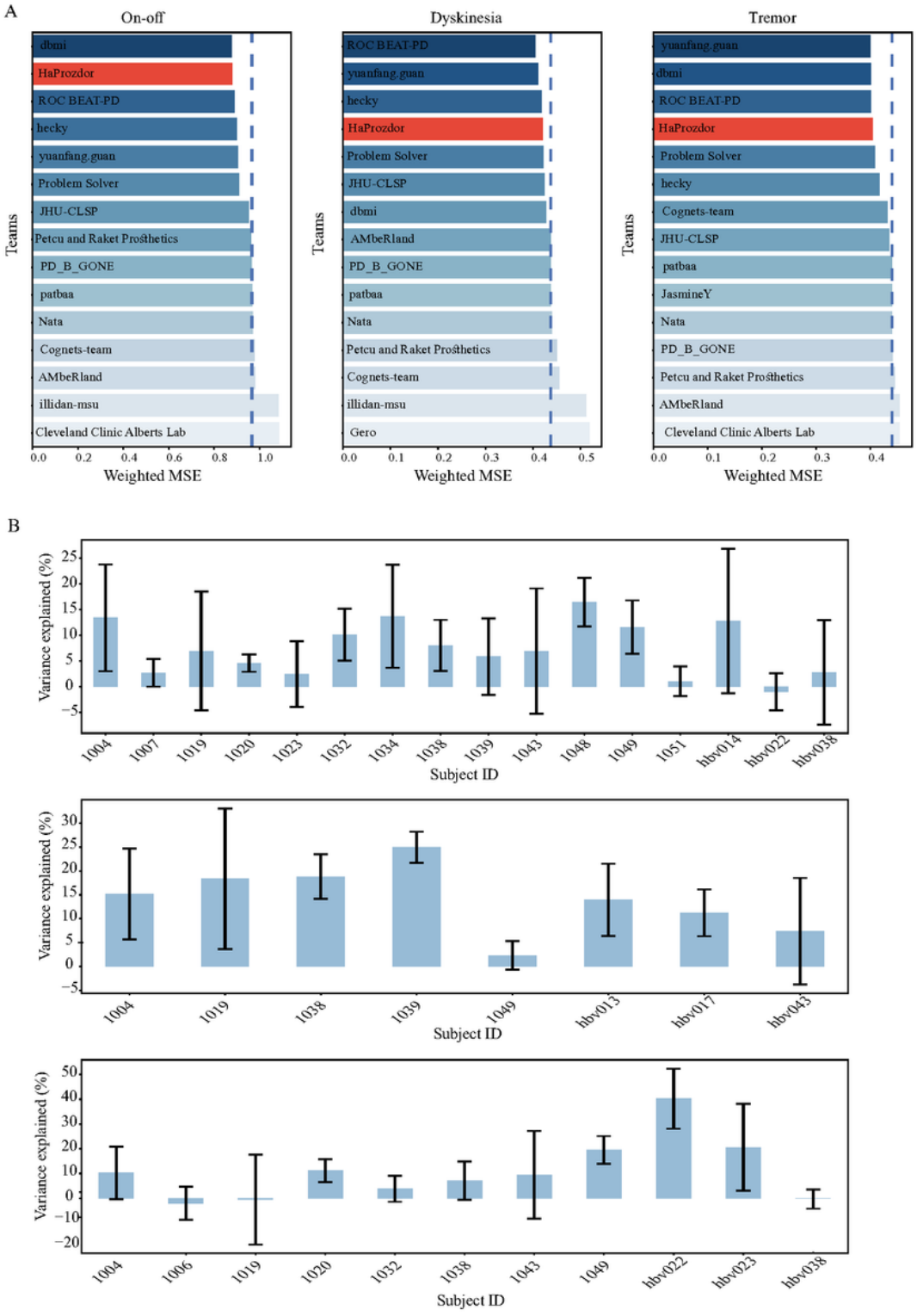

Figure 8

Challenge results. (A) Final round wMSE scores of the top fifteen teams in all the on-off (left), dyskinesia (middle) and tremor (right) sub-challenges (the full results can be accessed through the official challenge website [35]). The red bars represent our scores, and the dashed vertical lines represent the null models for each of the sub-challenges. (B) Mean fraction of variance explained by the models over all five validation

sets ± 1 SEM, for non-naïve subjects in the on-off (top), dyskinesia (middle) and tremor (bottom) categories.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BEATPDSuppTablesSciRep.docx