communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-03002-x OPEN

Heterogeneous digital biomarker integration outperforms patient self-reports in predicting Parkinson's disease

Kaiwen Deng[1,6], Yueming Li[1,6], Hanrui Zhang[1], Jian Wang[2], Roger L. Albin[3,4] & Yuanfang Guan[1,5]✉

Parkinson's disease (PD) is one of the first diseases where digital biomarkers demonstrated excellent performance in differentiating disease from healthy individuals. However, no study has systematically compared and leveraged multiple types of digital biomarkers to predict PD. Particularly, machine learning works on the fine-motor skills of PD are limited. Here, we developed deep learning methods that achieved an AUC (Area Under the receiver operator characteristic Curve) of 0.933 in identifying PD patients on 6418 individuals using 75048 tapping accelerometer and position records. Performance of tapping is superior to gait/rest and voice-based models obtained from the same benchmark population. Assembling the three models achieved a higher AUC of 0.944. Notably, the models not only correlated strongly to, but also performed better than patient self-reported symptom scores in diagnosing PD. This study demonstrates the complementary predictive power of tapping, gait/rest and voice data and establishes integrative deep learning-based models for identifying PD.

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [2] Eli Lilly and Company, Indianapolis, IN, USA. [3] Department of Neurology, University of Michigan, Ann Arbor, MI, USA. [4] VAAAHS GRECC, Ann Arbor, MI, USA. [5] Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [6] These authors contributed equally: Kaiwen Deng, Yueming Li. ✉email: gyuanfan@umich.edu

Parkinson's disease (PD) is one of the first disorders for which digital biomarkers were explored. The clinical features of PD include movement abnormalities such as tremors, bradykinesia, and rigidity[1]. Mobile devices and built-in sensors like accelerometers and gyroscopes provide the ability to convert the features into digital signals as quantitative surrogates of PD symptoms. The success of distinguishing the person with Parkinson's (PwP) from otherwise healthy individuals in the general population using the digital biomarkers will enable a remote and more convenient way for symptom evaluations and diagnosing, with minimal interruptions in the participants' daily life[2,3].

A multitude of public datasets is available[4] for analyzing various types of digital biomarkers such as voice recordings[2,5,6], movement data[2,7,8], the magnetic resonance imaging (MRI)[9], and the handwriting patterns[9,10]. And the application of machine learning or deep learning techniques like Support Vector Machine[11,12] and the Convolutional Neural Network[13] successfully build diagnostic models on these biomarkers. For example, a community-based challenge benchmarked algorithms using 30-s rest and gait data from cell phones to differentiate self-reported PwP from healthy subjects[14,15].

Despite a considerable number of studies using digital biomarkers for PwP detection, prior studies on the movement data have mainly focused on the evaluation of gross motor skills such as walking and rest with medium sample sizes[4]. In addition to gross motor function impairment, PwP typically experiences difficulty in daily tasks requiring fine-motor skills, *e.g.*, picking up objects, buttoning, tying shoelaces, and writing[16]. Abnormalities of fine-motor coordination are not only characteristic of PwP, but their presence is often a more sensitive indicator of PwP than changes in gait or balance, particularly in early phase PD. As a result, clinicians often use qualitative analysis of simple fine movements, such as finger tapping, to assess patients for possible parkinsonism. Additionally, assessment of fine movements is a component of standard clinical severity rating scales for PwP research[17]. However, there have been few digital biomarker studies focusing on fine-motor skills, and the applications of state-of-the-field machine learning techniques such as deep learning approaches only report a moderate performance[13]. Instead, prior studies with better performances mostly have focused on traditional signal extraction methods[18–20]. In addition, there are no fair comparisons or integration of algorithms across different types of motor assessments aimed at differentiating PwP from healthy individuals.

In this study, we aim at addressing the above open question with an innovative self-reported dataset collected by mPower[2]. mPower is an APP designed for collecting data from PwP and otherwise healthy individuals, including 20–30-s walking data, 10-s voice data, and 20-s finger tapping data. For a substantial number of individuals (2729), all three data types are available, allowing parallel evaluations of the performance of the models.

We first applied the deep learning algorithms on the finger-tapping accelerometer data. Then we construct the models on the coordinates of the tapping positions, under the expectation that PD patients exhibit worse coordination and slower motion during the tapping task compared to healthy individuals. This will lead to inaccurate and/or changing positions of the tapping contact points. The tapping coordinate models outperformed the gait/rest algorithm and the voice algorithm in differentiating PwP from otherwise healthy individuals. We further integrated all three types of models: finger tapping, voice, and rest/gait, and achieved an AUC of 0.944 in diagnosing PD. Importantly, further evaluations with AUCs on the individuals with self-reported UPDRS (The Unified Parkinson's Disease Rating Scale) scores show that our model can significantly outperform the symptom scores. The AUC of our work is 0.949, and the AUCs of UPDRS scores are 0.823, and 0.935 when using the UPDRS motor experience part.

Results

The goal of this study is to explore and integrate digital biomarkers of movement beyond the gross motor skills captured by gait and rest evaluations. We will present the development of the models for tapping records, followed by outlines of models for voice and walking data. The latter have benchmarks established by previous studies[15,21,22]. Finally, we will present the performance of the integrated models and demonstrate the clinical utility of the models by comparing with patient self-reports.

Model built on accelerometer data of tapping performs well in identifying PD. The tapping data was obtained from the mPower portal ([https://www.synapse.org/#!Synapse:syn5511439/tables/]). Data included a total of 8,003 individuals. Among them, 6418 have self-reported diagnosis, and 1060 are self-identified as having a professional diagnosis of PD (Table 1). We obtained a total of 78,879 records (75048 of them have the self-reported diagnosis), ranging from 1 to 522 for each individual (the median record number is 3; 75% of them have less than 6 records, and 95% of them have less than 30). On average, the individuals with PD have 40.2 records, and the individuals without PD have 6.06. Because each individual has multiple records, individual level, instead of record-level cross-validation was carried out in all analyses to prevent overestimation of the performance[23,24]. Specifically, in each round of cross-validation, 75% of the individuals and all their records are used as the training set, and the remaining 25% of individuals are used as the test set (Fig. 1a).

Table 1 Demographics of the individuals used in the tapping study.

Demographic type	Demographic value	Individual number of PD	Individual number of Non-PD	Total individual number
Gender	Male	698 (13.9%)	4329 (86.1%)	5027
	Female	359 (26.1%)	1014 (73.9%)	1373
	Prefer not to answer	0 (0.0%)	6 (100.0%)	6
	(Missing data)	3	9	12
Age	≤35	36 (0.95%)	3753 (99.04%)	3789
	35-50	158 (13.3%)	1034 (86.7%)	1192
	50-65	509 (56.2%)	397 (43.8%)	906
	>65	352 (71.5%)	140 (28.5%)	492
	(Missing data)	5	34	39
Smoke	No	696 (17.4%)	3299 (82.6%)	3995
	Yes	362 (16.6%)	1820 (83.4%)	2182
	(Missing data)	2	239	241

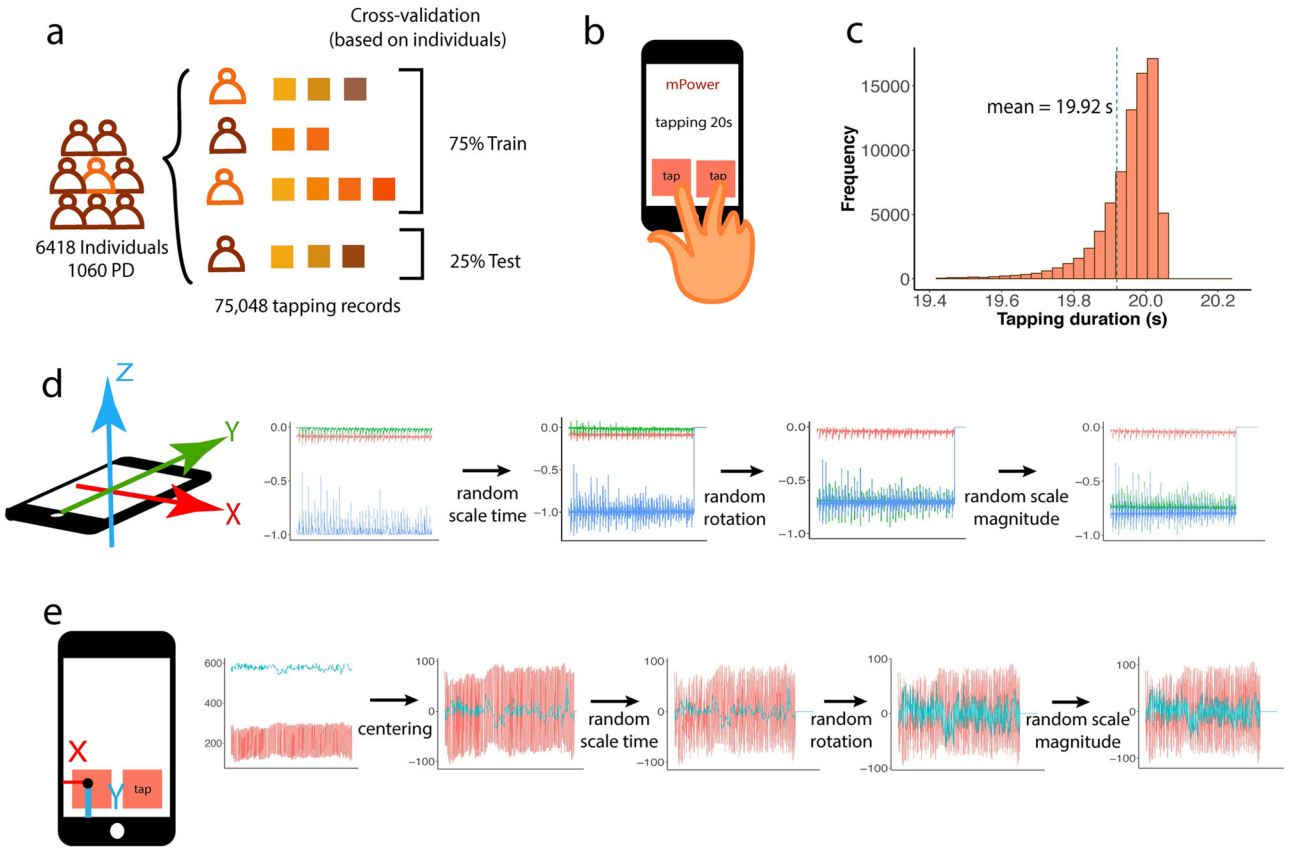

Fig. 1 Summary of tapping data and preprocessing techniques. a Training and test data are separated by individuals to avoid contamination. **b** Participants were asked to tap successively in turns on the screen of the phone for 20 s. Machine learning models are trained for accelerometer and coordinate data separately. **c** Distribution of the length of the records (**d**, **e**) demonstrate the preprocessing and augmentation techniques for raw accelerometer and coordinate data. Normalization/centering, random time scaling, rotation, and magnitude scaling were applied in a sequential order. At the last step, magnitude scaling was applied to each channel of the accelerometer and coordinated data.

During the tapping task, participants were asked to alternately tap two fingers of the dominant hand on the touchscreen of the phone, placed flat on a table, within two contiguous squares (Fig. 1b). Participants were asked to tap as quickly as possible for 20 s. Nearly 99% of the 75048 actual collected data have the duration times located between 19.4 and 20.2 s. The average duration time is 19.92 s, with a standard deviation 0.481 (Fig. 1c). Two types of data are recorded. The first type is the accelerometer, composed of [x, y, z] coordinate values sampled at 100 Hz (Fig. 1d). The second type is the location of the tapping on the screen, recorded as [*x*, *y*] coordinate values (Fig. 1e).

We first developed models that use accelerometer data to predict PD. When participants tap on the screen, the tapping motion induces acceleration of the phone. Thus, accelerometer data is capable of capturing the magnitude, direction and speed for the movement of the phone. A 1D deep learning network architecture was implemented, where the only dimension is time. [x, y, z] values were used as three channels of the input, analogous to color channels in image analysis (Fig. 2a). Timewise perturbation, magnitude augmentation, and spatial augmentation by rotating the record reference frames progressively improved model performance (Fig. 2b, Supplementary Table 1). As each individual had multiple records, we took the maximal or average prediction values across all records for each individual. Taking the maximal values appeared to perform better than taking the average of all records, as it ensures the PD individuals always have as high scores as possible, and generates a good separation with the non-PD individual scores (Fig. 2b). This likely reflects fluctuations in motor performance in PwP with the maximal values capturing peak abnormalities that are most predictive of PD. The performance (mean AUC value) from 5-fold cross-validation is 0.8340 at the record level, and 0.9174 at the individual level.

Deep learning models based on tapping coordinate data predict PD accurately. The mPower tapping coordinate data is composed of *x*, *y* positions on the cell phone screen, and the timestamp of each tap. The average number of taps in a single record is 153, ranging from 2 to 359. Before the deep learning models, we made exploratory experiments on the machine learning techniques using the time-series data feature extraction algorithms. We extracted 1508 features for each record using the Python package tsfresh[25], and fit a LightGBM model[26]. The dataset split and pulling methods are the same as those in the deep learning models. The average AUC is 0.692, and the performances are highly dependent on the training and test data, ranging from 0.6 to 0.92 in the 5-fold cross-validation.

We imported these records into a 1D deep learning network. We tested a variety of models including training models with or without adding the timestamps (Supplementary Table 2; Supplementary Fig. 1A), applying diverse augmentation and normalization methods (Supplementary Table 3, 4; Supplementary Fig. 1B, 2), and tuning the models with a series of hyperparameter settings (Supplementary Table 2; Supplementary Fig. 1A). We found that centering timestamp and coordinate data, and using Adabound improved model performance (Fig. 3b). On the other hand, other techniques, including normalizing the coordinates by subtracting the button position and 2D-rotation with a random angle in a

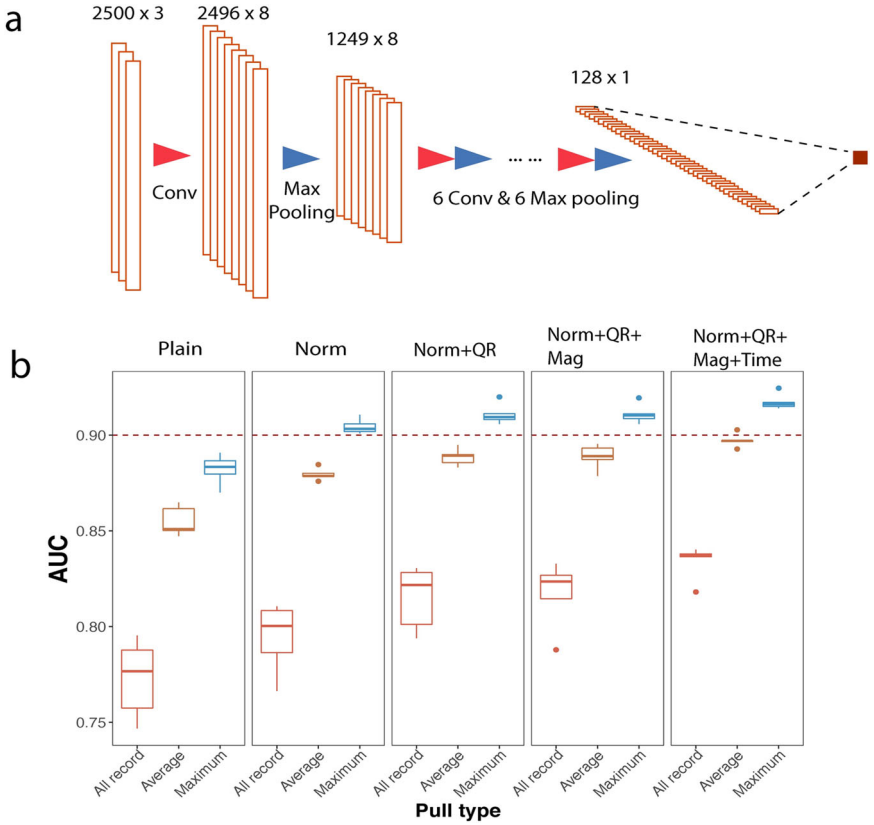

Fig. 2 Augmentation and normalization improve the performance of the tapping accelerometer model. a represents the model structure for the accelerometer data tapping model with seven convolutional layers, seven max-pooling layers, and 1 fully connected layer. **b** "Plain Model" represents the model using raw data without augmentation or normalization. "Norm" represents the model with Z-score normalization. "Quaternion Rotation" (QR), "Magnitude", and "Time" denote the three types of augmentation methods. The data were pre-processed in sequential order of the methods shown above the plot. "All record" indicates the performance evaluated on the record level; "Average" and "Maximum" represent the individual level performance by using the average or the maximum prediction of all records of the same individual. The models with normalization and all augmentation methods showed the top performance.

specific range (−90° to 90° and 0°–360°), impaired performances (Fig. 3b, Supplementary Fig. 2). During cross-validation, the AUC was 0.9352, compared to 0.9174 using accelerometer data ($p < 0.05$ for the cross-validation performances).

Comparison of tapping models with voice and gait/rest models for predicting PD. In order to compare the performance of the tapping models and the fine-motor skills, voice, and gait and rest (accelerometer) models for predicting PD. We identified a total of 2729 individuals (645 PwP) in the mPower dataset who had all three types of data. We retested the above accelerometer and coordinate models based on the tapping records of this set of individuals by cross-validation.

Next, we retrained a model for gait and rest for this set of individuals. This model follows Zhang et al.[15], which was a top-performing model in the DREAM Parkinson's disease challenge. Briefly, the accelerometer data of the cell phone during 20- or 30-s walking and rest activities were input into a 1D deep learning network of three channels, integrating spatial and time augmentation (Fig. 4). The models were trained separately for walking and rest. Then, we took the maximal value of the predictions across all records. The walking data achieved an average AUC of 0.8983.

Additionally, we trained a voice model, using a 1D deep learning network of one channel. The voice data in the mPower dataset are audio recordings of the participants saying 'Ahh..' for 10 s, sampled at a frequency of 44.1 kHz. We tested various modeling techniques and found that time-wise and magnitude-wise augmentation could improve the performance (Supplementary Fig. 3). The voice data achieved an average AUC of 0.8335 at the individual level in this population.

We also trained the accelerometer and coordinate tapping models with this population, which achieved average AUCs of 0.8983 and 0.9236. We assembled the two models by averaging the prediction scores from the two machine learning models for the same individual and predicted PwP status with the assembled score (Fig. 3c). The assembled model produces better performance than either of the single models, with an average AUC of 0.9333 (Fig. 4b). We found that tapping models significantly outperformed voice and gait/rest models in this population ($p < 0.05$ for both of the cross-validation performances, Fig. 4d).

We examined whether assembling the voice and gait/rest models in conjunction with tapping data models can further improve predictive performance (Fig. 4a). After averaging the evaluation scores from the four models (walk, voice, accelerometer tapping, and coordinate tapping) and using the assembled score to predict disease status, the average AUC reached 0.944, better than any single model ($p < 0.05$ for the cross-validation performances, Fig. 4d). This suggests that the evaluation scores of different models may have specific limitations and that assembling all sources of information may alleviate the individual limitations and obtain better performance (Fig. 4c, d).

Robust performance across demographic groups and comparison to patient self-reports. Previous studies of Parkinson's

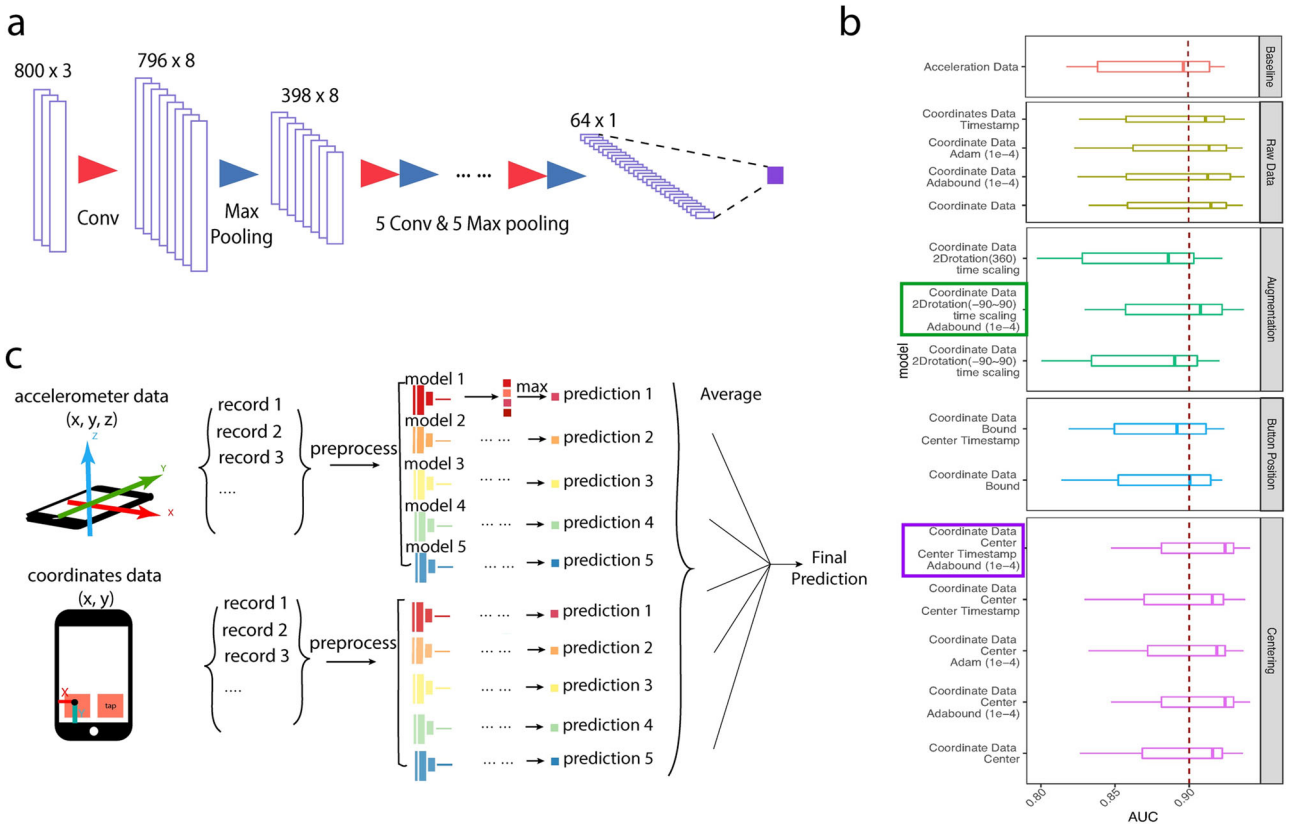

Fig. 3 Model design for tapping coordinate data. a depicts the model structure for the coordinate data tapping model (inputs: *x*, *y* coordinates and timestamp). The model contains six convolutional layers, six max-pooling layers, and one fully connected layer. **b** Experiments were designed to search for the optimal data processing methods and network hyperparameters, including the normalization strategy on the coordinates and the timestamps, augmentations with 2D-rotation and time scale, and the network optimizer. Best performing models by augmentation and centering groups are enclosed in squares. **c** A combined model is generated by averaging the prediction scores from accelerometer and from coordinates.

Disease presented the relationships between the risk, symptom and the demographics, such as the advancing age[27,28], gender[29,30], and the smoking behavior[31,32]. We examined if the performance is robust against these diverse demographic groups. We found that the performance remains similar and strong for different genders, smoking groups, and age groups, other than the age group < = 35 years old (Fig. 5a–c). All of the AUCs of the age groups older than 35 are higher than 0.85. We also evaluate our assembling model on the individuals older than 45 (the average individual number is 520), and have an average AUC of 0.885, 0.25 higher than the previous work on the same age group[13]. Age group ≤35 has an average AUC of 0.8 in the tapping models, and 0.681 in the assembling models. Of note, only 0.8% of this population has a positive PD diagnosis. Additionally, the prediction values of PD and normal groups are well separated into different demographic groups (Fig. 5d–f). As age increases, the overall prediction values increase as expected, due to more PD patients. Prediction values are independent of smoking status (Supplementary Fig. 4).

We further compared the performance of the prediction models against UPDRS (The Unified Parkinson's Disease Rating Scale) patient self-reports, when they are available. The AUC of the combined deep learning model on these shared individuals is 0.9486. UPDRS directly evaluated on the binary label achieved an AUC of 0.8232, and UPDRS part 2 (motor experience of daily living) achieved an AUC of 0.9356 (Fig. 6a, Supplementary Table 5). This result suggests that the digital biomarkers can make more accurate predictions than patient self-reports. Additionally, prediction values from the combined deep learning model showed strong correlation with UPDRS scores and with UPDRS part 2 scores (0.4240 and 0.5378, respectively, Fig. 6b, c, Supplementary Fig. 5, Supplementary Table 6). This result supports the clinical relevance and utility of the deep learning model.

Discussion

In this study, we proved the ability of finger-tapping positions on identifying the PwPs and the reported performances, an average AUC of 0.935, support that digital biomarkers for diagnosing PD can be developed beyond gross motor skills such as gait/rest, the primary focus of prior literature[33–35]. We also identified a digital biomarker approach that successfully predicts PwP status with even higher accuracy. This digital biomarker approach is based on integrative measurements of fine-motor skills and gross motor skills. The combined model achieved an average AUC of 0.944, a superior performance compared to patient self-reports. Compared to previous studies, our work has the largest dataset[4], and presents the ability of deep learning techniques to retrieve a high accuracy on predicting the PD diagnosis[13]. Our movement data model also has a higher performance in AUC than the previous works using traditional machine learning techniques[11,12,36,37].

However, lacking the information about how the diagnoses are determined, and the reliability of the self-reported conditions, our model can be influenced by the biases from different diagnostic criteria and the false-positive self-identifications. Besides, our models still have two potential limitations. First, the number of records per patient is different and the guideline allows the participant to submit data at any time they want. These lead to

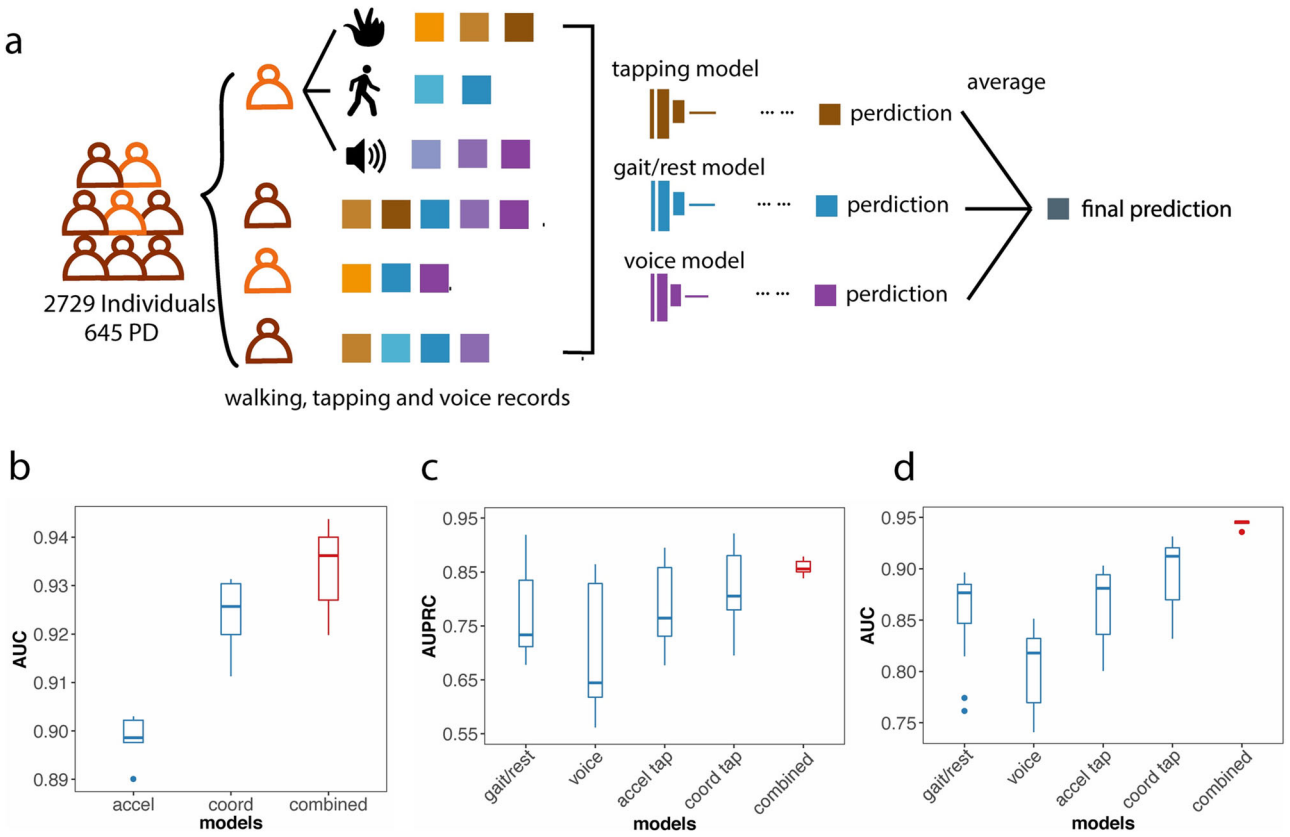

Fig. 4 Combining tapping, voice and gait/rest data improves performance in diagnosing PD. a A total of 2729 individuals (645 PD) who had all three types of data were split into the same cross-validation folds and to train gait/rest, voice, and tapping models. The models are combined together by averaging the prediction scores from each model. **b** Assembling accelerometer and coordinate tapping data models exceeds the performance of either single model. (**c**) and (**d**) depict the comparison among gait/rest, voice, tapping, and assembled models. The combined model achieves the best performance.

potential biases in analysis. We select the maximum scores for each individual to maximize the PD detection, but it may impair the classification for the non-PDs. Second, we include the data with abnormal tapping durations and tapping times. Although only less than 0.1% of the data have these abnormalities, it is possible for the model to learn the biases from them.

PD is characterized by motor abnormalities, specifically tremor and bradykinesia, with an impaired performance of fine-motor skills as a common early manifestation of PD. Micrographia, for example, is often present before the development of overt changes in gait, posture, or voice. Subtle changes in motor performance likely precede the emergence of overt PD[38]. Although the traditional motor-based diagnosis of PD, like the UPDRS, has already achieved a high accuracy, an easily accessible and robust method differentiating PwP from controls would be useful for screening the population, potentially including identification of prodromal PD. A complete MDS-UPDRS test requires the participant to fill a 33-page survey[39], which is time consuming and hard especially for a PD patient with motor disorder. Our work provides a possible solution for the convenient, in-home PD assessment and progression follow-up. It also may provide useful biomarkers for the evaluation of interventions.

These methods might also be useful for evaluating more advanced PD subjects. Identification of digital biomarkers independent of walking/rest data is a potentially useful approach for evaluating important motor functions in more advanced PD patients. As PD progresses, patients often lose their ability to think and reason, along with walking, but will maintain the ability to carry out tapping tests for much longer.

In summary, our results indicate an integrative digital biomarker approach with several potential applications for PD detection and monitoring of disease activities. Our study also emphasizes the advantages of utilizing different types of biomarkers. Previous studies have indicated the relationships between the PD and the other modalities like the rapid eye movement (REM) sleep behavior disorder (RBD)[40], the genotypes[41], and their abilities on PD predictions[42,43]. Combining the digital biomarkers from these sources may further improve the accuracy of the models[4], and may help distinguish PD with other types of tremulous movement disorders like SWEDD (Subjects Without Evidence of Dopaminergic Deficit)[44–46]. Possible future work also include a collection of long-term follow-up data to evaluate the model ability of predicting PD before diagnosis, and a collection of other movement disorders data like the cerebellar ataxia[47], and the Alzheimer's disease-associated movement disorders[48], for distinguishing PD from them. Our work can easily transfer to these classification tasks with the transfer learning strategy[49].

Methods

Ethics. The study data was downloaded from the mPower portal ([https://www.synapse.org/#!Synapse:syn5511439/tables/]). mPower study participants have agreed with secondary analysis of the data when signing on the APP. Additionally, all researchers that have access to the data in this study have obtained mPower permission.

General mPower participant guideline. The data collection was opened to the individuals diagnosed with PD as well as anyone interested in participating as a control[2]. They must be 18 years of age or older, living in the United States, and

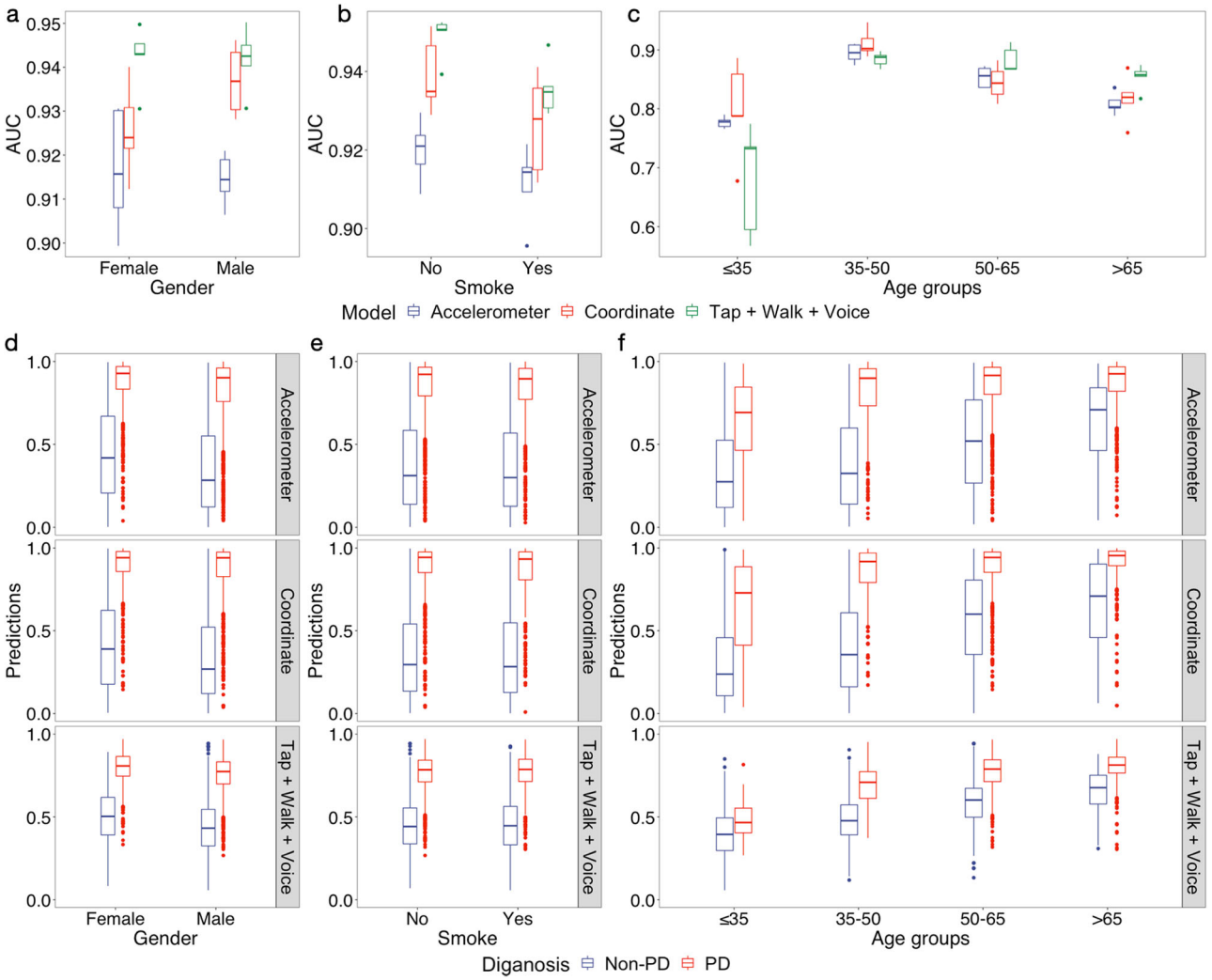

Fig. 5 Performance comparison of different demographic groups. a–c Performances in AUC and the predictions in different groups of data, separated by gender, smoking and age. **d–f** Distribution of prediction scores of Non-PD and PD for different gender, smoking, and age groups.

comfortable reading and writing on an iPhone in English. Participants needed to take a quiz of the study aims, participants right and data sharing, and were required to complete an e-consent process and sign. They were also asked for an email for verification.

Participants can submit their records three times a day. Participants with PD were asked to finish the tests in three scenarios: (1) immediately before taking their medication; (2) after taking their medication (when they are feeling at their best); (3) at some other time. And those with non-PD can complete at any time.

Deep learning architecture and training procedure. Similar neural network structures were built for accelerometer and coordinate data models except for channel lengths according to different sizes of input data: 2500 for the accelerometer model and 800 for the coordinate model. The network contains seven convolutional layers, seven max-pooling layers, and one fully connected layer activated by a sigmoid function for the output (Figs. 2a, 3a).

We performed cross-validation by partitioning the mPower samples into the training set (75%) and the testing set (25%) at the individual level with a random seed. Furthermore, half of the training samples were used to train the model and the other half were used in the validation process for hyperparameter tuning. During each training process, the binary cross-entropy $H_b(p)$ was applied as the loss function for selecting the best epoch, the parameters used in that epoch would then be stored for validation. The loss function is defined as:

$$H_b(p) = -p \times \log(\hat{p}) - (1-p) \times \log(1-\hat{p}) \quad (1)$$

where p is the ground truth (1 or 0 in our case) and $\hat{p}$ is the prediction value. To evaluate the performances of different models, Area Under Receiver Operating Characteristic Curve (AUC) and Area Under Precision-Recall Curve (AUPRC) scores were calculated and compared. Notably, AUC would not be affected by the baseline accuracy (number of PD individuals/number of all individuals), thus was more stable in our imbalanced data[50]. Predicting most individuals as non-PD would retrieve a relatively high accuracy, while the AUC could be nearly random.

Adabound was an variant of the Adam optimizer employing dynamic bounds on the learning rate. It was claimed to have both a rapid training process and good generalization ability[51]. We implemented this algorithm in Theano and applied it to our coordinate data model, along with a series of hyperparameter tunings on input length, batch size, and learning rate. The model with Adabound optimizer, 800 input length, 8 batch size, and 1e−4 learning rate reached the best performance (Fig. 3b).

Data normalization and augmentations. In order to avoid the potential data leakage, both the normalization and the augmentations are applied on each batch separately during the training process[23].

Normalization methods in our experiments include a z-score normalization $(X_b - \underline{X_b})/\mathrm{std}(X_b)$, a centering $X_b - \underline{X_b}$ (X_b is the data in a batch), and a boundary normalization using the bounds of the tapping areas. The boundary normalization is inspired by the fact that the resolutions of different devices may also contribute to the tapping position variances among the individuals. Calculating the relative positions by subtracting the bounds from the raw coordinates can eliminate the bias from devices.

The augmentations are scalings and rotations. To scale the time, we resize the lengths of the raw inputs with a coefficient randomly selected from (0.8, 1.2) using the OpenCV[52]. To scale the magnitude, we multiply the raw signals with coefficients randomly selected from (0.8, 1.2) on each channel. The rotations are implemented based on the quaternion rotation matrix[53] for the 3-D (the accelerometer data) and the 2D (the coordinates data) objects. The alternative rotation ranges in our experiments are $(0, 2\pi)$ and $(-\pi/2, \pi/2)$.

Pulling predictions of multiple records of a single individual and aggregation of the two models. Since mPower collected data from voluntary participants in an

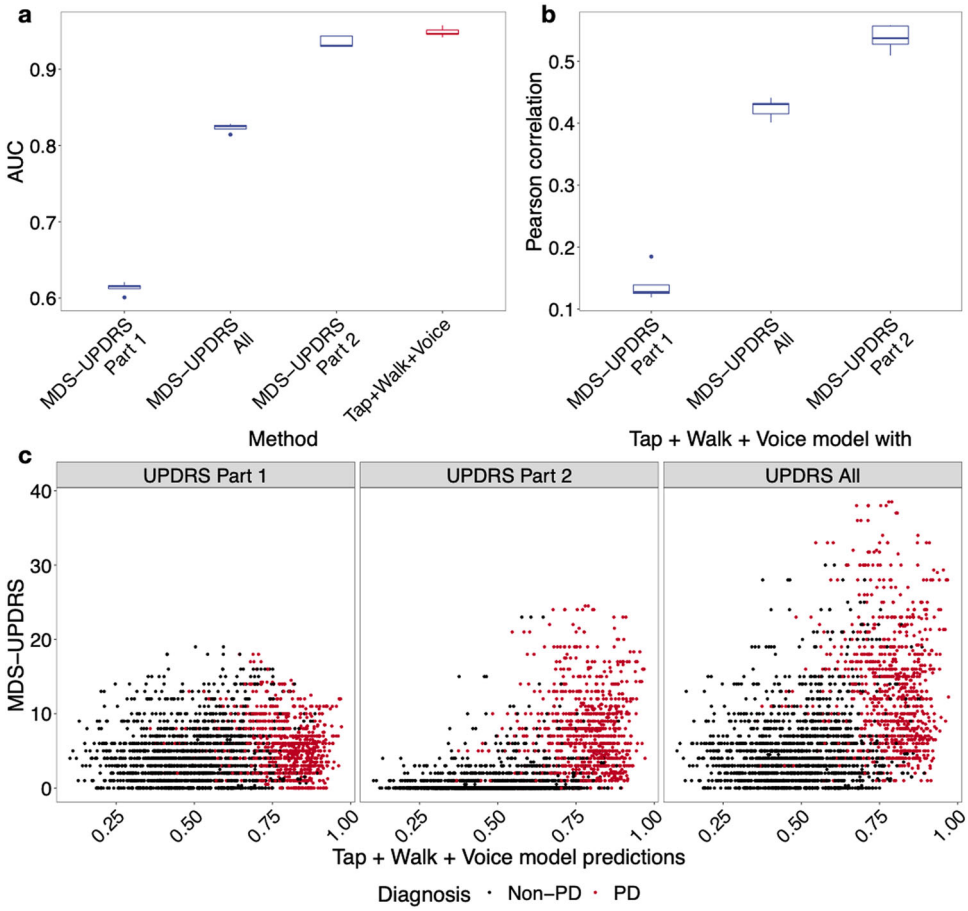

Fig. 6 Correlation with patient self-reports. a AUCs of evaluating self-reported MDS-UPDRS against diagnosis and deep learning models against diagnosis. **b** Pearson correlations between self-reported MDS-UPDRS scores and deep learning-based predictions. **c** Visualization of the relationship between the MDS-UPDRS and the prediction values.

uncontrolled environment, each individual might perform one task multiple times and generate more than one record. We used the same pulling prediction strategy to deal with multiple records from the same individual as models we built on mPower walking data and voice data[15]. Since we applied a 5-fold cross-validation method during model training, we had five evaluation scores for the same record in each fold. We first averaged these five scores to one mean score and used this score in pulling. We tried two pulling methods—average pulling and maximum pulling. In average pulling, the mean of all the records from a single individual was calculated. In maximum pulling, the maximum evaluation score of an individual was picked. The final evaluation score was then used to predict the PwP status of that individual.

Since each tapping test collected both acceleration and tapping coordinate data, we were able to assemble the results of these two models to see if model performance improved. Since the pulling method improved the AUC score for each model, we used the pulled score for each individual during aggregation. The evaluation scores from the two models of the same individual were collected, which is similar to adding another feature in the single model. The mean of the two scores was calculated as the final score for that individual. We used this final score to predict the PwP status.

Assembling gait and voice models on top of the tapping model. In order to compare among models built on different types of data, we identified the individuals with all types of these records. We used records from these individuals to train all four models (gait, voice, and tapping models) with the same split of 5-fold cross-validation. After comparing the performance of these single models, we tried to assemble gait and voice models on top of the best-performed tapping model (coordinate model). Similar strategies as aggregating two tapping models were used—the pulled score for each individual from the three models was collected together and averaged to a final score for the following evaluation.

Statistics and reproducibility. Statistics were performed in the R Studio (R version 4.0.2). The p-values were calculated by the two-sided t-test on the AUCs from the 5-fold cross-validations. The models were constructed using Theano (version 1.0.2) and the Lasagne (version 0.2.dev1) in Python (version 2.7.5). More information about the environment and implementation details can be found in https://github.com/GuanLab/PDTap.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets used in our works are available through the mPower Public Researcher Portal (https://www.synapse.org/mpower). Researchers who are interested in accessing these data should follow the mPower Data Governance (https://github.com/Sage-Bionetworks/mPower-sdata) (Bot et al. 2016). All the data behind the figures are available in the zip file of the Supplementary Data.

Code availability

Code is available at https://github.com/GuanLab/PDTap.

Received: 26 August 2021; Accepted: 23 December 2021;
Published online: 17 January 2022

References

1. Postuma, R. B. et al. MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* **30**, 1591–1601 (2015).
2. Bot, B. M. et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016).
3. Trister, A. D., Dorsey, E. R. & Friend, S. H. Smartphones as new tools in the management and understanding of Parkinson's disease. *NPJ Parkinson's Dis.* **2**, 16006 (2016).

4. Mei, J., Desrosiers, C. & Frasnelli, J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Front. Aging Neurosci.* **13**, 633752 (2021).
5. Frank, A. & Asuncion, A. *UCI Machine Learning Repository* (University of California, 2010).
6. Orozco-Arroyave, J. R., Arias-Londono, J. D., Vargas-Bonilla, J. F., Gonzalez-Rativa, M. C. & Noth, E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2014/Orozco14-NSS.pdf.
7. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).
8. Fernandez, K. M. et al. Gait initiation impairments in both Essential Tremor and Parkinson's disease. *Gait Posture* **38**, 956–961 (2013).
9. Marek, K. et al. The Parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).
10. Pereira, C. R. et al. A step towards the automated diagnosis of Parkinson's disease: analyzing handwriting movements. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems* 171–176 (IEEE Computer Society Publications, 2015).
11. Alaskar, H. & Hussain, A. Prediction of Parkinson disease using gait signals. In *2018 11th International Conference on Developments in eSystems Engineering (DeSE)* 23–26 (IEEE Computer Society Publications, 2018).
12. Pham, T. D. Pattern analysis of computer keystroke time series in healthy control and early-stage Parkinson's disease subjects using fuzzy recurrence and scalable recurrence network features. *J. Neurosci. Methods* **307**, 194–202 (2018).
13. Prince, J. & de Vos, M. A deep learning framework for the remote detection of Parkinson's disease using smart-phone sensor data. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* https://doi.org/10.1109/embc.2018.8512972 (2018).
14. Sieberts, S. K. et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. https://doi.org/10.1101/2020.01.13.904722.
15. Zhang, H., Deng, K., Li, H., Albin, R. L. & Guan, Y. Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson's Disease. *Patterns (N Y)* **1**, 100042 (2020).
16. Gokcal, E., Gur, V. E., Selvitop, R., Yildiz, G. B. & Asil, T. Motor and non-motor symptoms in Parkinson's disease: effects on quality of life. *Noro Psikiyatri Arsivi* **54**, 143–148 (2017).
17. Goetz, C. G. et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov. Disord.* **23**, 2129–2170 (2008).
18. Stamatakis, J. et al. Finger tapping clinimetric score prediction in Parkinson's disease using low-cost accelerometers. *Comput. Intell. Neurosci.* **2013**, 717853 (2013).
19. Stamatakis, J., Cremers, J., Macq, B. & Garraux, G. Finger Tapping feature extraction in Parkinson's disease using low-cost accelerometers. In *Proc. 10th IEEE International Conference on Information Technology and Applications in Biomedicine* https://doi.org/10.1109/itab.2010.5687769 (2010).
20. Viteckova, S. et al. Maximal velocity and amplitude decrement angle: a novel parameter for finger tapping instrumental evaluation in Parkinson disease. *Gait Posture* **73**, 474–475 (2019).
21. Sriram, T. V. S., Rao, M. V., Narayana, G. V. S. & Kaladhar, D. S. V. G. K. Diagnosis of Parkinson disease using machine learning and data mining systems from voice dataset. In *Proc. 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)* 2014 (eds Satapathy, S. C., Biswal, B. N., Udgata, S. K., & Mandal, J.K.) 151–157 (Springer International Publishing, 2015).
22. Hariharan, M., Polat, K. & Sindhu, R. A new hybrid intelligent system for accurate detection of Parkinson's disease. *Comput. Methods Programs Biomed.* **113**, 904–913 (2014).
23. Wen, J. et al. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Med. Image Anal.* **63**, 101694 (2020).
24. Chaibub Neto, E. et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit. Med.* **2**, 99 (2019).
25. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh—A Python package). *Neurocomputing* **307**, 72–77 (2018).
26. Ke, G. et al. in *Advances in Neural Information Processing Systems* 30 (eds. Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).
27. Reeve, A., Simcox, E. & Turnbull, D. Ageing and Parkinson's disease: why is advancing age the biggest risk factor? *Ageing Res. Rev.* **14**, 19–30 (2014).
28. Van Den Eeden, S. K. et al. Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *Am. J. Epidemiol.* **157**, 1015–1022 (2003).
29. Haaxma, C. A. et al. Gender differences in Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **78**, 819–824 (2007).
30. Picillo, M. et al. The relevance of gender in Parkinson's disease: a review. *J. Neurol.* **264**, 1583–1607 (2017).
31. Mappin-Kasirer, B. et al. Tobacco smoking and the risk of Parkinson disease: A 65-year follow-up of 30,000 male British doctors. *Neurology* **94**, e2132–e2138 (2020).
32. Savica, R., Grossardt, B. R., Bower, J. H., Ahlskog, J. E. & Rocca, W. A. Time trends in the incidence of Parkinson disease. *JAMA Neurol* **73**, 981–989 (2016).
33. Rehman, R. Z. U. et al. Comparison of walking protocols and gait assessment systems for machine learning-based classification of Parkinson's disease. *Sensors* **19**, 5363 (2019).
34. Morris, R. et al. A model of free-living gait: a factor analysis in Parkinson's disease. *Gait Posture* **52**, 68–71 (2017).
35. Abdulhay, E., Arunkumar, N., Narasimhan, K., Vellaiappan, E. & Venkatraman, V. Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease. *Future Gener. Comput. Syst.* **83**, 366–373 (2018).
36. Yang, M., Zheng, H., Wang, H. & McClean, S. Feature selection and construction for the discrimination of neurodegenerative diseases based on gait analysis. In *2009 3rd International Conference on Pervasive Computing Technologies for Healthcare* 1–7 (IEEE Computer Society Publications, 2009).
37. Oung, Q. W. et al. Wearable multimodal sensors for evaluation of patients with Parkinson disease. In *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* 269–274 (IEEE Computer Society Publications, 2015).
38. Fereshtehnejad, S.-M. et al. Evolution of prodromal Parkinson's disease and dementia with Lewy bodies: a prospective study. *Brain* **142**, 2051–2067 (2019).
39. MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS.htm.
40. Tekriwal, A. et al. REM sleep behaviour disorder: prodromal and mechanistic insights for Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **88**, 445–451 (2017).
41. Angeli, A. et al. Genotype and phenotype in Parkinson's disease: lessons in heterogeneity from deep brain stimulation. *Mov. Disord.* **28**, 1370–1375 (2013).
42. Ruffini, G. et al. Deep learning with EEG spectrograms in rapid eye movement behavior disorder. *Front. Neurol.* **10**, 806 (2019).
43. Prashanth, R., Roy, S. D., Mandal, P. K. & Ghosh, S. Parkinson's disease detection using olfactory loss and REM sleep disorder features. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2014**, 5764–5767 (2014).
44. Bajaj, N. P. S. et al. Accuracy of clinical diagnosis in tremulous parkinsonian patients: a blinded video study. *J. Neurol. Neurosurg. Psychiatry* **81**, 1223–1228 (2010).
45. Kim, M. & Park, H. Using Tractography to Distinguish SWEDD from Parkinson's Disease Patients Based on Connectivity. *Parkinson's Dis.* **2016**, 1–10 (2016).
46. Choi, H., Ha, S., Im, H. J., Paek, S. H. & Lee, D. S. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *NeuroImage: Clinical* **16**, 586–594 (2017).
47. Yang, Z., Zhong, S., Carass, A., Ying, S. H. & Prince, J. L. Deep learning for cerebellar ataxia classification and functional score regression. *Mach. Learn. Med. Imaging* **8679**, 68–76 (2014).
48. Kurlan, R., Richard, I. H., Papka, M. & Marshall, F. Movement disorders in Alzheimer's disease: more rigidity of definitions is needed. *Mov. Disord.* **15**, 24–29 (2000).
49. Tan, C. et al. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning—ICANN 2018* (eds Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L. & Maglogiannis, I.) 270–279 (Springer International Publishing, 2018).
50. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
51. Luo, L., Xiong, Y., Liu, Y. & Sun, X. Adaptive gradient methods with dynamic bound of learning rate. *Proceedings of the 7th International Conference on Learning Representations.* Preprint at https://arxiv.org/abs/1902.09843 (ICLR, 2019).
52. Howse, J. *OpenCV Computer Vision with Python* (Packt Publishing Ltd, 2013).
53. Morais, J. P., Georgiev, S. & Sprößig, W. in *Real Quaternionic Calculus Handbook* (eds. Morais, J. P., Georgiev, S. & Sprößig, W.) 35–51 (Springer, 2014).

Acknowledgements

The voice model of the study is supported by American Parkinson's Disease Association (APDA). Gait/rest part of the study is supported by the Michael J. Fox Foundation. The Tapping model of this study was supported by Eli Lilly and Company internal fund. RLA is supported by R21 NS114749, P50NS123067, and the Parkinson's Foundation.

Author contributions

Study design: Y.G., J.W. Data analytics: Y.L., K.D., H.Z. Manuscript: Y.G., R.A. Figures: Y.L. All authors read and approved the manuscript.

Competing interests

The authors declare the following competing interests: J.W. is a current Eli Lilly and Company employee. Y.G. serves as the scientific advisor for Eli Lilly and Company on the tapping part of this study. R.L.A. serves on the DSMBs for the COMPASS and PASSPORT trials (Biogen), the M-STAR trial (Biohaven), and the Signal-AD trial (Vaccinex). R.L.A. has received consulting fees from the Michael J. Fox Foundation and Takeda. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03002-x.

Correspondence and requests for materials should be addressed to Yuanfang Guan.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Eirini Marouli and Karli Montague-Cardoso. Peer reviewer reports are available.