npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Using AI to measure Parkinson's disease severity at home

Md Saiful Islam[1,2]✉, Wasifur Rahman[1], Abdelrahman Abdelkader[1], Sangwu Lee[1], Phillip T. Yang[3], Jennifer Lynn Purks[3], Jamie Lynn Adams[3], Ruth B. Schneider[3], Earl Ray Dorsey[3] and Ehsan Hoque[1]

We present an artificial intelligence (AI) system to remotely assess the motor performance of individuals with Parkinson's disease (PD). In our study, 250 global participants performed a standardized motor task involving finger-tapping in front of a webcam. To establish the severity of Parkinsonian symptoms based on the finger-tapping task, three expert neurologists independently rated the recorded videos on a scale of 0–4, following the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS). The inter-rater reliability was excellent, with an intra-class correlation coefficient (ICC) of 0.88. We developed computer algorithms to obtain objective measurements that align with the MDS-UPDRS guideline and are strongly correlated with the neurologists' ratings. Our machine learning model trained on these measures outperformed two MDS-UPDRS certified raters, with a mean absolute error (MAE) of 0.58 points compared to the raters' average MAE of 0.83 points. However, the model performed slightly worse than the expert neurologists (0.53 MAE). The methodology can be replicated for similar motor tasks, providing the possibility of evaluating individuals with PD and other movement disorders remotely, objectively, and in areas with limited access to neurological care.

npj Digital Medicine (2023)6:156 ; https://doi.org/10.1038/s41746-023-00905-9

INTRODUCTION

Parkinson's disease (PD) is the fastest-growing neurological disease, and currently, it has no cure. Regular clinical assessments and medication adjustments can help manage the symptoms and improve the quality of life. Unfortunately, access to neurological care is limited, and many individuals with PD do not receive proper treatment or diagnosis. For example, in the United States, an estimated 40% of individuals aged 65 or older living with PD do not receive care from a neurologist[1]. Access to care is much scarce in developing and underdeveloped regions, where there may be only one neurologist per millions of people[2]. Even for those with access to care, arranging clinical visits can be challenging, especially for older individuals living in rural areas with cognitive and driving impairments.

The finger-tapping task is commonly used in neurological exams to evaluate bradykinesia (i.e., slowing of movement) in the upper extremities, which is a key symptom of PD[3]. The task requires an individual to repeatedly tap their thumb finger with their index finger as fast and as big as possible. Videos of finger-tapping tasks have been used to analyze movement disorders like PD in prior research. However, the videos are often collected from a few participants (<20)[4], or the studies only provide binary classification (e.g., slight vs. severe Parkinsonian symptoms; Parkinsonism vs. non-Parkinsonism) and do not measure PD severity[5,6]. Additionally, existing models lack interpretability, making it difficult to use them in clinical settings. Most importantly, the videos are noise-free as they are recorded in a clinical setting with the guidance of experts. Machine learning models trained on clean data may not perform effectively if the task is recorded in a noisy home environment due to the models' susceptibility to data shift[7]. Consequently, these models may not enhance access to care for Parkinson's disease.

Imagine anyone from anywhere in the world could perform a motor task (i.e., finger-tapping) using a computer webcam and get an automated assessment of their motor performance severity. This presents several challenges: collecting a large amount of data in the home environment, developing interpretable computational features that can be used as digital biomarkers to track the severity of motor functions, and developing a platform where (elderly) people can complete the tasks without direct supervision. In this paper, we address these challenges by leveraging AI-driven techniques to derive interpretable metrics related to motor performance severity and apply them across 250 global participants performing the task mostly from home. Three experts and two non-experts rated the severity of motor performance watching these videos, using the movement disorder society unified Parkinson's disease rating scale (MDS-UPDRS). Our proposed interpretable, clinically relevant features highly correlate with the experts' ratings. An AI-based model was trained on these features to assess the severity score automatically, and we compared its performance against both expert and non-expert clinicians. Note that, all the experts are US neurologists with at least 5 years of experience in PD clinical studies and actively consult PD patients. The non-experts are MDS-UPDRS certified raters but do not actively consult PD patients. One of the non-experts holds a medical degree from a non-US institution and has actively engaged in multiple PD clinical studies. The other non-expert is a second-year neurology resident who has been active in movement disorder research for 10 years. Figure 1 presents an illustrative overview of our system.

RESULTS

Data

We obtained data from 250 global participants (172 with PD, 78 control) who completed a finger-tapping task with both hands (see Fig. 2 for examples). Participants used a web-based tool[8] to record themselves with a webcam primarily from their homes.

[1]Department of Computer Science, University of Rochester, 250 Hutchinson Road, Rochester, NY 14620, USA. [2]Department of Computer Science and Engineering, Bangladesh University of Engineering & Technology, West Palashi, Dhaka 1000, Bangladesh. [3]Department of Neurology, University of Rochester Medical Center, 601 Elmwood Avenue, Rochester, NY 14642, USA. ✉email: mislam6@ur.rochester.edu

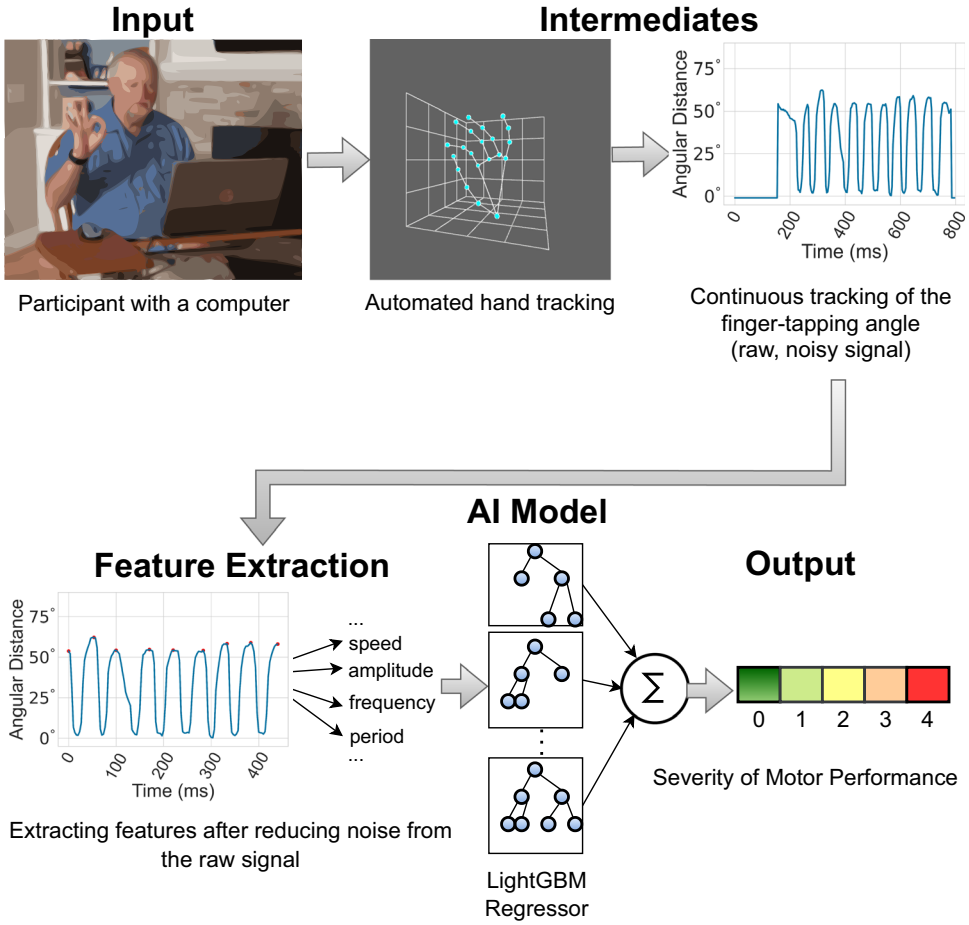

Fig. 1 Overview of the AI-based system for assessing the severity of motor performance. Anyone can perform the finger-tapping task in front of a computer webcam. The system employs a hand-tracking model to locate the key points of the hand, enabling a continuous tracking of the finger-tapping angle incident by the thumb finger-tip, the wrist, and the index finger-tip. After reducing noise from the time-series data of this angle, the system computes several objective features associated with motor function severity. The AI-based model then utilizes these features to assess the severity score automatically. Authors have obtained consent to publish the image of the participant.

Fig. 2 Data collection. The participants, both those with Parkinson's disease (PD) and healthy controls, performed the task primarily in a noisy home environment without any clinical supervision. The dataset includes blurry videos caused by poor internet connection, videos where participants had difficulty following instructions, and videos with overexposed or underexposed backgrounds. These issues are common when collecting data from home, particularly from an aged population that may be less familiar with technology than other age groups. Authors have obtained consent to publish the images of the participants.

Table 1. Demographic characteristics of the participants.

Subgroup	Attribute	With PD	Without PD	Total
	Number of participants	172	78	250
Sex, *n* (%)	Male	109 (63.4%)	28 (35.9%)	137 (54.8%)
	Female	63 (36.6%)	50 (64.1%)	113 (45.2%)
Age in years, *n* (%) (range: 18–86, mean: 62.13)	Below 20	0 (0.0%)	3 (3.8%)	3 (1.2%)
	20–29	0 (0.0%)	10 (12.8%)	10 (4.0%)
	30–39	1 (0.6%)	3 (3.8%)	4 (1.6%)
	40–49	5 (2.9%)	6 (7.7%)	11 (4.4%)
	50–59	34 (19.8%)	14 (18.0%)	48 (19.2%)
	60–69	64 (37.2%)	30 (38.5%)	94 (37.6%)
	70–79	62 (36.0%)	12 (15.4%)	74 (29.6%)
	Above 80	6 (3.5%)	0 (0.0%)	6 (2.4%)
Ethnicity, *n* (%)	white	161 (93.6%)	69 (88.5%)	230 (92%)
	Asian	2 (1.2%)	5 (6.4%)	7 (2.8%)
	Black or African American	1 (0.6%)	2 (2.6%)	3 (1.2%)
	American Indian or Alaska Native	2 (1.2%)	0 (0.0%)	2 (0.8%)
	others	1 (0.6%)	0 (0.0%)	1 (0.4%)
	not mentioned	5 (2.9%)	2 (2.6%)	7 (2.8%)
Recording environment, *n* (%)	Home	140 (81.4%)	59 (75.6%)	199 (79.6%)
	Clinic	25 (14.5%)	17 (21.8%)	42 (16.8%)
	Unknown	7 (4.1%)	2 (2.6%)	9 (3.6%)

With PD column represents the participants with self-reported diagnosis of Parkinson's disease, Without PD column represents the participants who did not self-report to have Parkinson's disease.

Demographic information for the participants is presented in Table 1.

Following the MDS-UPDRS guidelines, we considered each participant's left and right-hand finger-tapping as two separate videos. All these $250 \times 2 = 500$ videos are rated by three expert neurologists with extensive experience providing care to individuals with PD and leading PD research studies. However, after undertaking manual and automated quality assessments, we removed 11 videos from the dataset. Ultimately, we had 489 videos for analysis (244 videos for the left hand and 245 for the right hand). We obtained the ground truth severity score (a) by majority agreement when at least two experts agreed on their ratings (451 cases), or (b) by taking the average of three ratings and rounding it to the nearest integer when no majority agreement was found (38 cases). Supplementary Table 2 contains additional information on how the severity scores are distributed across demographic subgroups.

Rater agreement

The three expert neurologists demonstrated good agreement on their ratings, as measured by (a) Krippendorff's alpha score of 0.69 and (b) Intra-class correlation coefficient (ICC) score of 0.88 (95% confidence interval: [0.86, 0.90]). Figure 3 provides an overview of pair-wise agreement between expert raters. All three experts agreed in 30.7% of the videos, and at least two agreed in 93% of the videos. The three raters showed a difference of no more than 1 point from the ground truth in 99.2%, 99.5%, and 98.2% of the cases, respectively. These metrics suggest that the experts can reliably rate our videos recorded from home environments.

Features as digital biomarkers

We quantified 47 features measuring several aspects of the finger-tapping task, including speed, amplitude, hesitations, slowing, and rhythm. We also quantified how much an individual's wrist moves using 18 features. For each feature, Pearson's correlation coefficient (r) is measured to see how the feature is correlated with the ground truth severity score, along with a statistical significance test (significance level $\alpha = 0.01$). We found that 22 features were significantly correlated with the severity scores, which reflects their promise for use as digital biomarkers of symptom progression. Table 2 shows the top 10 features with the highest correlation. These features are clinically meaningful as they capture several aspects of speed, amplitude, and rhythm (i.e., regularity) of the finger-tapping task, which are focused on the MDS-UPDRS guideline for scoring PD severity.

Traditionally, human evaluators cannot constantly measure the finger-tapping speed. Instead, they count the number of taps the participant has completed within a specific time (e.g., three taps per second). However, in our case, the videos were collected at 30 frames per second rate, thus allowing us to track the fingertips 30 times per second and develop a continuous measure of speed. The former approach, the number of finger taps completed in unit time, is termed as "frequency" throughout the paper, and "speed" (and "acceleration") refers to the continuous measure (i.e., movement per frame). Similarly, "period" refers to the time it takes to complete a tap, and thus, is a discrete measure. In addition, finger tapping amplitude is measured by the maximum distance between the thumb and index-finger tips during each tap. Since linear distance can vary depending on how far the participant is sitting from the camera, we approximated amplitude using the maximum angle incident by three key points: the thumb-tip, the wrist, and the index fingertip. As we see in Table 2, several statistical measures of continuous speed are significantly correlated with PD severity. These granular computations are only attainable using automated video analysis, which, to our knowledge, was missing in prior literature.

Performance of non-expert clinicians

When an individual lacks access to a neurologist with expertise in movement disorders, they may consult with a non-specialist

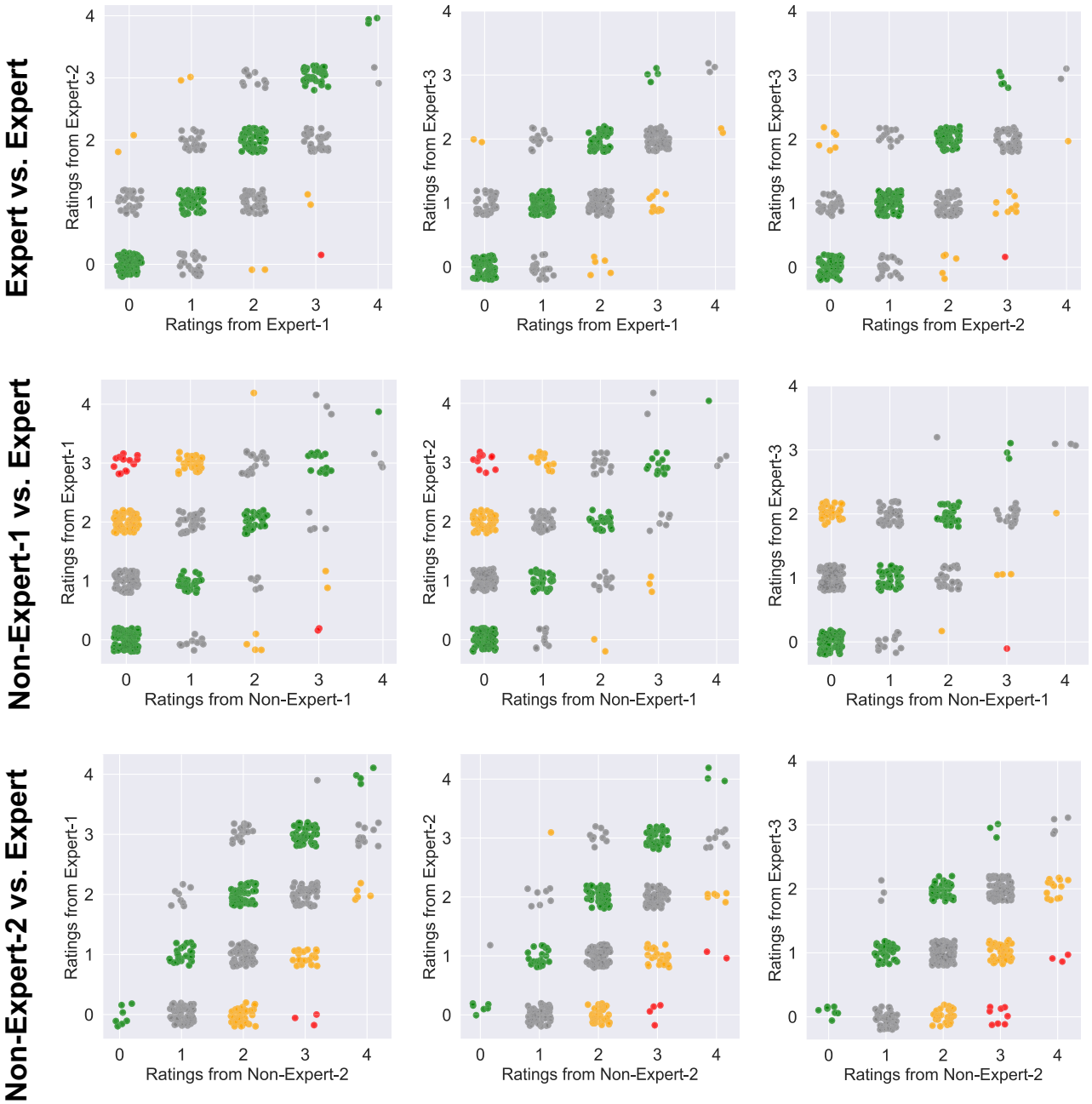

Fig. 3 An overview of how the experts and the non-experts agreed on their ratings. Green dots indicate two raters having a perfect agreement, while gray, orange, and red dots imply a difference of 1, 2, and 3 points, respectively. We did not observe any 4 points rating difference. The high density of green and gray dots and an ICC score of 0.88 verifies that the experts demonstrated high inter-rater agreement among themselves, and the finger-tapping task can be reliably rated when recorded from home. However, the non-experts were less reliable than the experts, demonstrating moderate agreement with the three expert raters (the average ICC of a non-expert's ratings and the ratings from the three experts were 0.72, 0.74, and 0.70, respectively).

clinician. Thus, it is critical to assess how a clinician with limited expertise in movement disorders or Parkinson's disease may perform compared to experts in this field. To this end, we recruited two investigators. The first investigator (referred to as non-expert-1 throughout the manuscript) had completed an MBBS (i.e., Bachelor of Medicine, Bachelor of Surgery) degree but not additional medical training (e.g., residency), was certified to administer the MDS-UPDRS, and had the experience of rating the severity of PD symptoms in multiple research studies. The second investigator (referred to as non-expert-2) is a second-year neurology resident at a reputed medical center in the United States. This investigator has been involved in movement disorders research and clinical trials for 10 years and takes care of less than 20 PD patients a year. We asked the non-expert clinicians to rate the same videos that the three experts had rated. We observed moderate reliability in their ratings, with an average intra-class correlation coefficient (ICC) of 0.75 compared to the ground truth scores (Fig. 3). On average, the non-experts' ratings deviated from the ground truth severity score by 0.83 points, and the average Pearson's correlation coefficient (PCC) between the non-experts' ratings and the ground truth severity scores was 0.61.

Modeling the severity rating

We employed a machine learning model (i.e., LightGBM regressor[9]) that predicts the severity of PD symptoms based on the extracted features from a finger-tapping video. To evaluate the model, we implemented a leave-one-patient-out cross-validation approach, that leaves all data (e.g., videos of both left and right hand) associated with a particular patient as a test set, and the model is trained with the remaining data. The evaluation uses multiple iterations to ensure that the performance is validated once for each patient. The severity rating predicted by the model is a continuous value, ranging from 0 to 4. We employed several standard metrics used to assess the performance of machine learning models in regression tasks: mean absolute error (MAE), mean squared error (MSE), Kendall rank correlation coefficient (Kendall's τ), mean absolute percentage error (MAPE), Pearson's correlation coefficient (PCC), and Spearman's rank correlation coefficient (Spearman's ρ). However, in this manuscript, we primarily focus on MAE and PCC which are the most popular in assessing a model's regression capability. Also, to measure classification accuracy, the continuous prediction is converted to five severity classes (0, 1, 2, 3, and 4) by rounding it to the closest integer. The classification result is reported in Fig. 4, showing that the model's predictions largely agree with the ground truth severity scores. The model's reliability in rating the videos is moderate, as indicated by an ICC score of 0.76 (95% C.I.: [0.71, 0.80]). On average, the model predictions deviated from the ground truth severity scores by 0.58 points, and the Pearson's correlation coefficient between the predictions and the ground truth severity scores was 0.66. Since the ground truth scores were derived from the three experts' ratings, it is natural to find an excellent correlation (PCC = 0.86) and a minimal difference (MAE = 0.27) between an average expert and the ground truth. However, this is an unfair baseline to compare the performance of the model and the non-expert. Instead, we looked at how the

Table 2. Features most correlated with the ground truth severity scores for the finger-tapping task.

Feature	Statistic	r	p-value	Rank
Speed	Inter-quartile range	−0.56	10^{-33}	1
	Median	−0.52	10^{-27}	2
	Maximum	−0.32	10^{-10}	6
Amplitude	Median	−0.50	10^{-25}	3
	Maximum	−0.41	10^{-17}	4
Frequency	Inter-quartile range	0.32	10^{-10}	5
	Standard deviation	0.29	10^{-8}	8
Period	Entropy (i.e., irregularity)	0.32	10^{-10}	7
	Variance (normalized by the average period)	0.28	10^{-8}	9
	Inter-quartile range	0.27	10^{-7}	10

The ground truth scores are obtained from the aggregate of expert ratings. r and p-value indicate Pearson's correlation coefficient and significance level (obtained by correlation test), respectively. Features are ranked based on the correlation coefficients (the rank-1 feature is the most correlated with the ground truth severity scores).

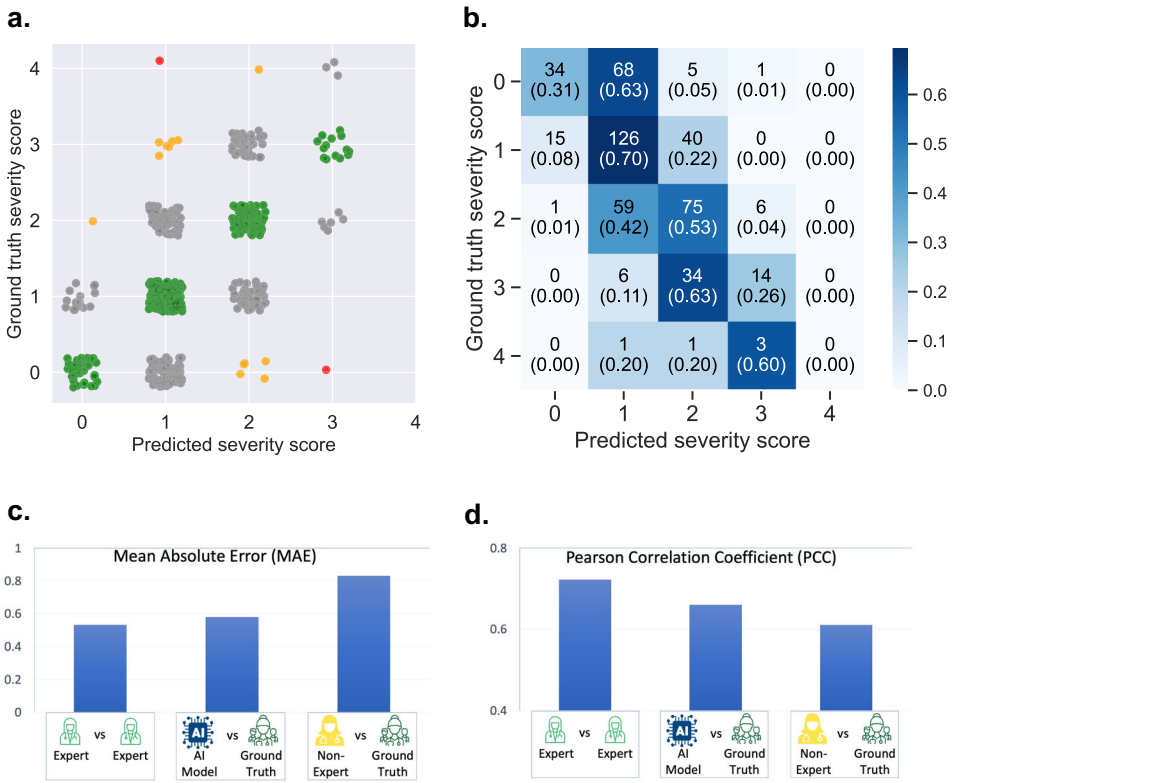

Fig. 4 Model performance. a We observe good agreement between the predicted severity and the ground truth scores. Green dots indicate correct predictions, while gray, orange, and red dots imply a difference of 1–3 points between the predicted and actual scores. We did not observe any 4 points rating difference. **b** The confusion matrix presents the agreement numerically. **c** The mean absolute error (MAE) measures the difference between two ratings. The model incurs slightly higher MAE than an average expert but substantially lower MAE than the non-experts. **d** Pearson correlation coefficient (PCC) measures the correlation between two sets of ratings. The model's predicted severity ratings are more correlated with the ground truth scores than the non-experts' (higher PCC) but less correlated than an average expert's (lower PCC) ratings.

expert neurologists agreed with each other to establish a human-level performance of the rating task. On average, any pair of experts differed by 0.53 points from each other's ratings, and their ratings were correlated with PCC = 0.72. In most of the metrics we tested, the LightGBM regression model outperformed the non-expert clinicians but *was outperformed* by the experts (Fig. 4). Please see Supplementary Table 1 for details.

Interpretability of model predictions

We used SHapley Additive exPlanations (SHAP) to interpret the outputs of the machine learning model. SHAP values provide a way to attribute a prediction to different features and quantify the impact of each feature on the model output[10]. This information can be useful for understanding how a model makes decisions and for identifying which features are most important for making accurate predictions. We found that important features identified by SHAP align well with our previously identified significant features (see Table 2). Top-10 most important features include finger-tapping speed (IQR), freezing (maximum duration and number of freezing), absence of periodicity, period (variance, minimum), wrist-movement (minimum, median), and frequency (IQR), which are all significantly correlated with the ground-truth severity score (at a significance level, $\alpha = 0.01$). The only top-10 feature that does not correlate significantly with the severity score is the median period (average time taken to complete each tap), which is also underscored in the MDS-UPDRS guideline. These results indicate that the model is looking at the right features while deciding the finger-tapping severity scores from the recorded videos, further suggesting the model's reliability.

Analyzing bias

To better evaluate the performance of our model across various demographic groups, we conducted a group-wise error analysis that takes into account factors such as sex, ethnicity, age, and Parkinson's disease diagnosis status. This approach allows us to assess any potential biases or inaccuracies in our model's predictions and make necessary improvements to ensure equitable and accurate results for all users. We combined the model predictions of the samples for each patient when they were in the test set. Additionally, we tracked the demographic attributes associated with each sample, allowing us to evaluate the performance of our model across various demographic groups.

Our model achieved a mean absolute error (MAE) of 0.60 points (standard deviation, std = 0.48) for male subjects ($n = 267$) and 0.55 points (std = 0.39) for female subjects ($n = 222$), indicating relatively accurate predictions for both sexes. Furthermore, we conducted a statistical test (i.e., two-sample two-tailed *t*-test) to compare the errors across the two groups and found no significant difference (p-value = 0.21). We also performed a similar error analysis on our model's predictions for subjects with PD ($n = 333$) and those without PD ($n = 156$). The model had an MAE of 0.57 points for subjects with PD and 0.59 points for those without PD. However, we found no significant difference in the errors between these two groups (p-value = 0.61). These results suggest that our model does not exhibit detectable bias based on sex and performs similarly for PD and non-PD groups.

Age had a slight negative correlation with the error of model predictions (Pearson's correlation coefficient, $r = -0.06$). However, the correlation was not statistically significant at $\alpha = 0.05$ significance level (p-value = 0.20). This suggests that the performance of the model is not significantly different across the younger and older populations.

Finally, we tested whether the model exhibits bias based on an individual's ethnicity. Since we did not have enough representation from all the ethnic groups, the analysis only focused on the white vs. non-white population. Our model had an MAE of 0.57 points (std = 0.45) for white subjects ($n = 452$), while the MAE was 0.65 (std = 0.41) for non-white subjects ($n = 37$). Although the difference was not statistically significant (p-value = 0.29 based on *t*-test) at a 95% confidence level, the model seems to perform slightly worse for the non-white population.

Impact of video quality

Since most of the video recordings are collected from participants' homes, the quality of the videos may vary widely. Several factors, such as the lighting condition of the recording environment, quality of the data capturing devices (i.e., webcams and internet browsers), participants' cognitive ability and understanding of the task, surrounding noise (e.g., multiple persons being visible in the recording frame) can affect the quality of the recorded videos. Hence, it is essential to analyze how video quality influences clinical ratings, the performance of the pose estimation model used in this study, and, eventually, the model performance. When providing finger-tapping severity ratings for each video, each expert neurologist identified cases where the video was difficult to rate due to quality issues or the participant's inability to follow the task appropriately. We grouped the videos into two categories: high-quality videos, which none of the experts had difficulty rating, and low-quality videos, which at least one expert had difficulty rating.

Impact on rater agreement. We examined whether the agreement between the expert raters varied depending on the quality of the videos. For the high-quality videos ($n = 385$), the measure of inter-rater reliability, known as the intra-class correlation coefficient (ICC), was found to be 0.879 (95% confidence interval [CI] = [0.86, 0.90]). On the other hand, the ICC for the low-quality videos ($n = 104$) was 0.806 (95% CI = [0.73, 0.86]). Although this difference is not statistically significant (at a 95% confidence level) due to the overlap in the confidence intervals for the two groups, it suggests that the video quality may have slightly impacted the agreement among the expert raters. We also ran a Chi-square test of independence and found no significant association between finding a majority agreement among the experts and the quality of the videos at a 95% confidence interval (test-statistic = 0.9965, p-value = 0.318, degree of freedom = 1). In addition, there were 38 videos (out of 489) where the experts disagreed by at least 2 points. 27 (71%) out of these 38 videos were marked as high-quality by all experts, whereas only 11 videos (29%) were marked as low-quality by one or more experts.

Impact on pose-estimation model. Next, we tried to evaluate whether video quality impacts the performance of the pose estimation model used in this study (i.e., MediaPipe). For each frame in the video, MediaPipe provides a "hand presence score" while estimating the coordinates of the hand key-points. The hand presence score is a measure of confidence in the model's ability to track the hand in the current frame. The score is a number between 0 and 1, where 0 indicates no confidence and 1 indicates high confidence. A high hand presence score indicates that the model is confident in its ability to track the hand. This means that the landmarks are likely to be accurate and that the pose estimation is likely to be correct. A low hand presence score indicates that the model is not confident in its ability to track the hand. Therefore, to estimate the pose estimation performance on a video, we used the average hand presence score across all the frames in the video after dropping the starting and ending frames that do not contain the hand of interest. The mean hand presence scores were 0.967 and 0.962, respectively, across the high-quality ($n = 385$) and low-quality videos ($n = 104$). Based on these results, we did not observe a notable difference in the performance of MediaPipe hand tracking between the high-quality and low-quality videos. This finding was supported by statistical analysis

using a two-sample two-tailed *t*-test, which yielded a *p*-value of 0.36 with test statistic, $t = 0.91$.

Additionally, we manually injected noise to randomly selected 86 "high-quality" (43 videos of right-hand and 43 of left-hand) videos from our dataset to see how different types of noise impact the confidence (hand presence scores) of MediaPipe. Specifically, we applied different levels of blurring by performing a low-pass filter operation on each frame with two different kernel sizes (3×3 (slight blurring) and 9×9 (substantial blurring)) and also injected different levels of Gaussian noise with zero means and two different standard deviations (25 (low amount of noise), and 40 (high amount of noise)). As a result, each video had five distinct versions: the original video, slightly blurred, substantially blurred, low amount of Gaussian noise, and high amount of Gaussian noise. In general, MediaPipe had lower confidence scores for the videos with injected noise. For example, the mean confidence score for the original videos was 0.96 (std = 0.031) while the mean for the slightly blurred and substantially blurred videos were 0.958 (std = 0.033) and 0.935 (std = 0.08), respectively. Similarly, the mean scores for the videos with a low amount of added noise and a high amount of added noise were 0.897 (std = 0.124) and 0.784 (std = 0.170), respectively. The difference in MediaPipe confidence scores was not significant between the original and slightly blurred videos. However, the rest of the group-wise differences were statistically significant as validated by paired sample *t*-tests (please see Supplementary Table 5 for associated test statistics).

Impact on model performance. Finally, we also evaluate whether the model performs poorly for low-quality videos. Specifically, we evaluate the prediction errors of the model across two groups of videos (high-quality and low-quality) and run two-sample *t*-tests to see whether there is any significant difference in prediction errors across these two groups. The mean absolute errors of the model were 0.581 (std = 0.43) and 0.578 (std = 0.506) respectively across the high and low-quality videos. Therefore, we observed no significant difference in performance across these two groups ($t = 0.064$, p-value = 0.95) based on a two-sample *t*-test. Also, we analyzed the cases where the model predictions were off by more than 1.5 points (resulting in at least 2 points difference when the continuous predictions are converted into severity classes). Out of 15 such occurrences, 10 videos were marked as high-quality by all experts and only 5 videos were marked as low-quality by at least one expert.

DISCUSSION

Here we report three significant contributions. First, we demonstrate that the finger-tapping task can be reliably assessed by neurologists from remotely recorded videos. Second, this study suggests that AI-driven models can perform close to clinicians and possibly better than non-specialists in assessing the finger-tapping task. Third, the proposed model is equitable across sex, age, and PD vs. non-PD subgroups. These offer new opportunities for utilizing AI to address movement disorders, extending beyond Parkinson's disease to encompass other conditions like ataxia and Huntington's disease, where finger-tapping provides valuable insights into the severity of the disease.

Our tool can be expanded to enable longitudinal tracking of symptom progression to fine-tune the treatment of PD. People with PD (PwP) often exhibit episodic symptoms, and longitudinal studies require careful management of variables to ensure accurate temporal responses to individual doses of medication. It is best practice to conduct repeated ratings under consistent conditions, such as at the same time of day, the same duration after the last medication dose, and with the same rater[11]. However, the limited availability of neurological care providers and mobility constraints of elderly PwP make this challenging. In the future, we envision extending our platform for other neurological tasks (e.g., postural and rest tremors, speech, facial expression, gait, etc.) so that patients can perform an extensive suite of neurological tasks in their suitable schedule and from the comfort of their homes. For this use case, our tool is not intended to replace clinical visits for individuals who have access to them. Instead, the tool can be used frequently between clinical visits to track the progression of PD, augment the neurologists' capability to analyze the recorded videos with digital biomarkers and fine-tune the medications. In healthcare settings with an extreme scarcity of neurologists, the tool can take a more active role by automatically assessing the symptoms frequently and referring the patient to a neurologist if necessary.

We introduce several digital biomarkers of PD—objective features that are interpretable, clinically useful, and significantly associated with the clinical ratings. For example, the most significant feature correlated with the finger-tapping severity score is the interquartile range (IQR) of finger-tapping speed (Table 2). It measures one's ability to demonstrate a range of speeds (measured continuously) while performing finger-tapping and is negatively correlated with PD severity. As the index finger is about to touch the thumb finger, one needs to decelerate and operate at a low speed. Conversely, when the index finger moves away from the thumb finger after tapping, one needs to accelerate and operate at high speed. A higher range implies a higher difference between the maximum and minimum speed, denoting someone having more control over the variation of speed and, thus, healthier motor functions. Moreover, the median and the maximum finger-tapping speed, as well as the median and the maximum finger-tapping amplitudes have strong negative correlations with PD severity. Finally, IQR and standard deviation of finger tapping frequency, as well as entropy, variance, and IQR of tapping periods, are found to have strong positive correlations with PD severity, as they all indicate the absence of regularity in the amplitude and periods. These findings align with prior clinical studies reporting that individuals with Parkinson's have slower and less rhythmic finger tapping than those without the condition[12]. These computational features can not only be used as *digital biomarkers* to track the symptom progression of PwP but also explain the model's predicted severity score (e.g., an increase in the severity score can be attributed to factors like reduced tapping speed, smaller amplitude, etc.)

The proposed model exhibits an average prediction error of 0.58, indicating that it frequently predicts a level 1 severity for a healthy individual who actually has a severity score of 0. As observed in the confusion matrix, the model misclassifies 63% of the ground-truth zero severity scores as severity 1. Moreover, it is less accurate for videos with severity 3 and 4. We achieved an overall accuracy of 50.92% when utilizing the LightGBM regressor as a classifier, highlighting the need for further enhancement in the model. However, it is important to note that the lack of accuracy is also evident among experts and non-experts. On average, a pair of experts only concurred on the severity of a finger-tapping video 51.35% of the time. Additionally, the non-experts obtained an overall accuracy of 36.03%. This underscores the challenge of precisely assessing symptoms of Parkinson's disease. In clinical settings, although the assessment of motor signs is important, it alone does not determine a Parkinson's disease diagnosis. Clinicians also consider the patient's health history, medications, and non-motor symptoms, such as anxiety, depression, impaired sense of smell, constipation, and changes in sleep among other factors, to determine a diagnosis. However, MDS-UPDRS scores are highly suitable for monitoring individuals already diagnosed with Parkinson's disease. An increase in the severity score from the baseline indicates the manifestation of more Parkinsonian symptoms, while a decrease suggests an improvement in symptoms. Considering this use case, severity assessment is typically regarded as a regression problem rather than a classification problem. Thus, having a strong correlation

(e.g., Pearson's correlation coefficient) and low error (e.g., mean absolute error) with respect to the ground-truth labels are the most desirable metrics.

When developing tools for analyzing data recorded in home environments, it is crucial to consider the various types of noise that can naturally occur in such settings. Home videos may exhibit background noise, inadequate lighting, blurriness, and other artifacts. These factors can pose challenges for both doctors and models in evaluating the videos. In this study, at least one of the experts expressed discomfort in rating 104 out of 489 videos due to quality issues. Although the difference was not statistically significant, there was a slightly lower level of inter-rater agreement among the experts when it came to low-quality videos. However, the model's performance, as indicated by the prediction error, was similar for both good and poor-quality videos. It is possible that the video quality remained sufficient for the pose-estimation model and the feature-extraction framework employed to automatically assess the severity of the finger-tapping task. Nevertheless, further analysis indicates that MediaPipe (the pose-estimation model used in this study) struggles when significant external systematic noise is introduced into the videos. For instance, the confidence of MediaPipe hand-tracking dropped when a subset of the good-quality videos was intentionally blurred or when random Gaussian noise was added. In general, we recommend ensuring a minimum level of video quality to obtain reliable ground truth and facilitate the use of more precise tools for video processing.

As we prepare to roll out our AI tool in healthcare settings, we must prioritize ethical considerations such as data security, user privacy, and algorithmic bias. As AI platforms become increasingly integrated into healthcare domains, there is growing emphasis on protecting against data breaches and crafting appropriate regulatory frameworks prioritizing patient agency and privacy[13]. This ever-evolving landscape will have significant implications for our future approach. Additionally, algorithmic bias and its risks in perpetuating healthcare inequalities will present ongoing challenges[14]. Many AI algorithms tend to underdiagnose underserved groups[15], and it is critical to evaluate and report the model performance across age, ethnicity, and sex subgroups. Our proposed model does not demonstrate detectable bias across the male and female populations, people with and without PD, white and non-white populations, and to a particular age group. Ensuring fairness and equitable performance across diverse demographics is essential for an AI model to be ethically sound and applicable in healthcare settings.

The study has some limitations that can be improved in the future. To begin with, some of the objective features examined in this study might be influenced by tremors, a significant symptom of Parkinson's disease that can frequently obscure signs of bradykinesia. For instance, accurately identifying individual finger taps necessitates precise peak detection from the finger-tapping motion over time. When tremors impact the motion, the signal can become unstable, posing a challenge in detecting clear and distinct peaks. Errors in peak detection would consequently impact the assessment of several features employed in this study, including finger-tapping period, frequency, and amplitude. Additionally, it is possible that severe tremors will affect the performance of pose estimation models. Pose estimation models operate by tracking the motion of body parts in a video. Tremors can induce erratic and unpredictable motion of the body parts, which can impede the model's ability to track the motion. This can result in inaccuracies in estimating the pose, ultimately compromising the accuracy of the extracted features. Unfortunately, we do not have the tremor diagnosis for the participants in this study and, therefore, could not provide definitive answers to these concerns. It is worth noting that, due to the jerky and unpredictable movements caused by tremors, doctors also encounter difficulties in assessing the speed of an individual's motion, rendering the diagnosis of bradykinesia challenging. Future studies may further investigate the connection between tremors and bradykinesia.

In addition, the proposed AI-driven model (i.e., LightGBM regressor) was trained with 489 videos from 250 global participants. While this dataset is the largest in the literature in terms of unique individuals, it is still a relatively small sample size for training models capable of capturing the essence of complex diseases such as Parkinson's. Furthermore, it is worth noting that the dataset utilized in this study exhibits class imbalance. More specifically, there is a scarcity of samples for severity classes 3 and 4. While the number of videos for severity classes 0, 1, and 2 amounted to 108, 181, and 141, respectively, there were only 54 and 5 videos available for severity classes 3 and 4. This class imbalance may have contributed to the model's inability to predict the most severe class (i.e., severity 4) correctly. Also, there is room for improvement in developing a fair and equitable model. Although not statistically significant, the proposed model is slightly more inaccurate for the male (vs. female) and non-white (vs. white) populations. In our dataset, videos of male subjects had significantly higher severity ($n = 267$, mean severity $= 1.43$, std $= 0.98$) than female subjects ($n = 222$, mean severity $= 1.19$, std $= 1.94$), as indicated by p-value of 0.003 obtained by one-tailed t-test. As we had less data for modeling the higher severity classes, this could have contributed to a higher average error for male individuals compared to females. Also, higher errors for non-white populations could be due to having less data to represent them. Notably, 92% of participants in this study self-reported as white. Diversifying our training data and gathering feedback from critical stakeholders (especially from traditionally underrepresented and underserved communities) will be important first steps for us to take toward building fair, high-performance algorithms in the future. To that end, we will diversify our dataset to be representative of the general population. Individuals with non-white ethnicity are typically underrepresented in clinical research[16]. Thus, emphasizing the recruitment of these populations through targeted outreach will be essential, especially considering the risks that homogeneous training data can pose in furthering healthcare inequalities[17]. In the future, we plan to improve our model's performance by building (i) a larger dataset with a better balance in severity scores and (ii) a gatekeeper to improve data quality. Additional data will be essential in building more powerful models with potentially better performance. Furthermore, we can improve our data quality and model performance by developing "quality control" algorithms to provide users with real-time feedback on capturing high-quality videos: such as adjusting their positioning relative to the webcam or moving to areas with better lighting. Developing user-friendly features to strengthen data collection and ensure minimum video quality will be crucial for collecting videos remotely without direct supervision.

Finally, it is worth noting that the symptoms of Parkinson's disease can vary depending on whether an individual is in the ON-state (under the effect of PD medication) or in the OFF-state (not under the effect of PD medication). It would be intriguing to investigate whether the model can effectively detect differences in symptom severity between participants who are ON or OFF PD medication in future studies.

METHODS

Data sources

Participants' data were collected through a publicly accessible web-based tool (https://parktest.net/). This tool allows individuals to contribute data from the comfort of their homes, provided they have a computer browser, internet connection, webcam, and microphone. In addition to the finger-tapping task, the tool also

gathers self-reported demographic information such as age, sex, ethnicity (i.e., white, Asian, Black or African American, American Indian or Alaska Native, others) and whether the participant has been diagnosed with Parkinson's disease (PD) or not. Moreover, the tool records other standard neurological tasks involving speech, facial expressions, and motor functions, which can help to extend this study in the future.

We collected data from 250 global participants who recorded themselves completing the finger-tapping task in front of a computer webcam. Data was collected primarily at participants' homes; however, a group of individuals (48) completed the task in a clinic using the same web-based tool. Study coordinators were available for the latter group if the participants needed help. The demographic characteristics of the study participants are provided in Table 1.

Clinical ratings

The finger-tapping task videos were evaluated by a team of five raters, including two non-specialists and three expert neurologists. The expert neurologists are all associate or full professors in the Department of Neurology at a reputable institution in the United States, possess vast experience in PD-related clinical studies, and actively consult PD patients. Both of the non-specialists are MDS-UPDRS certified independent raters. One of the non-specialists holds a non-U.S. bachelor's degree in medicine (MBBS) and has experience conducting PD clinical studies. The other non-specialist is a second-year neurology resident with 10 years of experience in clinical research related to movement disorders. The first non-specialist does not consult PD patients and the second non-specialist takes care of <20 PD patients per year.

The raters watched the recorded videos of each participant performing the finger-tapping task and rated the severity score for each hand following the MDS-UPDRS guideline. The MDS-UPDRS guideline is publicly available at https://www.movementdisorders.org/MDS-Files1/Resources/PDFs/MDS-UPDRS.pdf. The finger-tapping task is discussed in Part III, Section 3.4. The severity rating is an integer ranging from 0 to 4 representing normal (0), slight (1), mild (2), moderate (3), and severe (4) symptom severity. The rating instructions emphasize focusing on speed, amplitude, hesitations, and decrementing amplitude while rating the task. In addition to providing the ratings, the raters could also mark videos where the task was not properly performed or when a video was difficult to rate. We excluded the difficult-to-rate videos marked by any of the experts when analyzing the performance of the raters.

To compute the ground-truth severity scores, we considered only the ratings the three expert neurologists provided. If at least two experts agreed on the severity rating for each recorded video, this was recorded as the ground truth. If the experts had no consensus, their average rating rounded to the nearest integer was considered the ground truth. The ratings from the non-specialists were used solely to compare the machine-learning model's performance.

Feature extraction

We developed a set of features by analyzing the movements of several key points of the hand. The feature extraction process is comprised of five stages: (i) distinguishing left and right-hand finger-tapping from the recorded video, (ii) locating the target hand for continuous tracking, (iii) quantifying finger-tapping movements by extracting key points on the hand, (iv) reducing noise, and (v) computing features that align with established clinical guidelines, such as MDS-UPDRS. We implemented these stages using Python. Supplementary Note 1 lists the exact version for each Python package used in this study.

The finger-tapping task is performed for both hands, one hand at a time. However, to rate each hand independently, we divided the task video into two separate videos, one featuring the right hand and the other featuring the left hand. We manually reviewed each video and marked the transition from one hand to the other. The data collection framework will be designed to record each hand separately to avoid manual intervention in the future. After hand separation, the video V (of the hand-separated finger-tapping task) and the hand category (i.e., left or right) are provided as inputs to the feature extraction pipeline. If the hand category is specified as left, the analysis focuses solely on the movements of the left hand, and the same applies to the right-hand category.

After separating the left and right-hand finger-tapping videos, we applied MediaPipe Hands (https://google.github.io/mediapipe/solutions/hands.html) to detect the coordinates of 21 key points on each hand. MediaPipe is an open-source project developed by GoogleAI that provides a public API of a highly accurate state-of-the-art model for hand pose estimation. In addition, the hand pose estimation model is very fast, easy to integrate into a machine learning framework, and supports various platforms, including Android, iOS, and desktop computers. Furthermore, GoogleAI consistently updates the pose estimation models and seamlessly integrates these updates into the public API. We selected MediaPipe over other pose estimation platforms due to these compelling reasons.

Although each video contained finger-tapping of either the left or right hand (after hand separation), both hands might remain visible in the recording frame. Therefore, we needed to identify the target hand out of multiple visible hands before extracting key points. Initially, we processed the finger-tapping task video V in a frame-by-frame manner using MediaPipe. For each frame V_i, we first tried to locate the specified hand category h. Using the `multi_handedness` output from MediaPipe, we can identify how many different hands MediaPipe has detected and the handedness category, and the confidence score of detection for each detected hand. To identify the detected hand(s) that match the specified hand category h, we used the following heuristics:

```
hands_found ← []
for all j in range(len(multi_handedness)) do
  if (multi_handedness[j].classification[0].label = h) & (multi_
  handedness[j].classification[0].score > 0.9) then
  hands_found. append(j)
  end if
end for
```

We identified three possible scenarios:

- *MediaPipe did not detect any hand with matching hand category*: In this case, we simply assigned the frame to have missing hand.
- *MediaPipe detected a single hand matching the hand category*: We considered this to be the hand we will analyze.
- *Mediapipe detected multiple hands matching the hand category*: When multiple persons are visible in the frame, MediaPipe can detect multiple right hands or multiple left hands. In such situations, we made an assumption that the subject performing the task will have a larger hand compared to the other individual(s) in the background, as the study subject is likely to be the closest to the camera. To identify the largest hand, we compared the Euclidean distance between the coordinates of the wrist (`landmark[0]`) and the thumb-tip (`landmark[4]`) of the detected hands. The hand with the greatest distance was selected for further analysis.

After we identified the target hand from the MediaPipe detected hands for frame V_i, we extracted the `x` and `y` coordinates of four landmarks (i.e., hand key points): center of wrist joint (`WRIST, landmark[0]`), thumb carpometacarpal joint (`THUMB_CMC, landmark[1]`), tip of the thumb (`THUMB_TIP, landmark[4]`), and tip of the index finger (`INDEX_FINGER_TIP, landmark[8]`). These coordinates were used to track the finger-tapping angle and movement of the wrist over time.

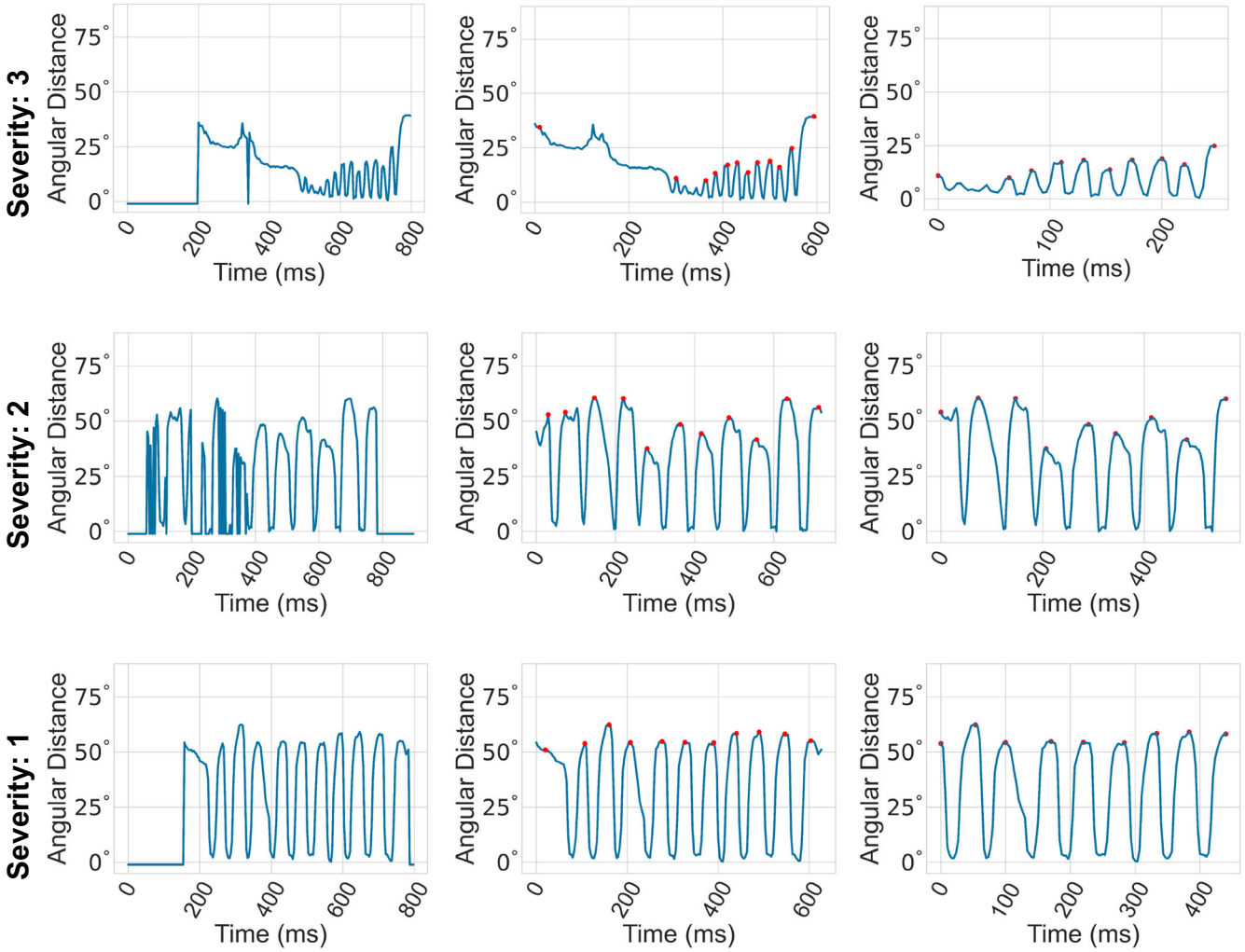

Fig. 5 Data pre-processing. Finger-tapping angles incident by three hand key points (thumb-tip, wrist, index finger-tip) plotted as a time series. Figures on the left show the noisy raw signals directly extracted using MediaPipe. After the noise reduction step, we identified peak angles (red dots) using a custom peak detection algorithm. Finally, trimming the signal by removing the first and last tap yields the cleanest signal used for analysis, as shown on the right. The top figures depict a person with severe tapping difficulty (severity: 3), resulting in low and irregular amplitudes. The central figures show a person with moderate tapping ability (severity: 2), with slow and interrupted tapping and irregular amplitudes. Finally, the bottom figures show a person with good rhythmic tapping ability, albeit with a slower tapping speed (severity: 1).

Instead of using Euclidean distance between the thumb-tip and index finger-tip to measure the amplitude, speed, and other metrics to quantify finger-tapping movements, we used the angle incident by three key points: `THUMB_TIP`, `WRIST`, and `INDEX_FINGER_TIP`. This helped us to deal with participants sitting at a variable distance from the camera since the angle is invariant to the camera distance. We computed the angle for each frame of the recorded video (i.e., if a video was recorded at 30 frames/s, we computed the angle 30 times per second). This helped us assess the speed and acceleration of the fingers in a continuous manner.

The computed finger-tapping angles are plotted as a time-series signal in Fig. 5 (left). Negative values of the angle are indicative of a missing hand in the captured frame. For example, at the beginning or at the end of the recording, the hand might not be visible in the recording frame, as the participant needs to properly position their hand. We detected missing hands using the handedness outputs from MediaPipe, which include information about the left/right hand and hand presence confidence score. In particular, if the target hand (left/right) is not recognized as a handedness category or if the hand presence confidence score is less than 0.90, we labeled the frame as having a missing hand. In these cases, we assign a finger-tapping angle of −1.0. However, MediaPipe can also inaccurately miss the hand in many frames, resulting in a negative value for the angle. To address this issue, we implemented a strategy to interpolate missing angle values when the majority of neighboring frames have non-negative values. Specifically, we looked at the five frames before and after the missing value. If the majority had non-negative angles, we interpolated the value using a polynomial fit on the entire signal. Then, we found the largest consecutive segment of frames where the hand is visible (i.e., the finger-tapping angle was non-negative) in the signal and remove the frames before and after that. This helps us to remove the pre and post-task segments where the participants were not tapping their fingers and ensure that the analysis is focused on the relevant segment of the signal. Figure 5 (middle) shows how the raw, noisy signals were converted to cleaner signals after performing this step.

The participants need to adjust the positioning of their hands before starting to tap their fingers, and they also need to move their hands after completing the task, which introduces further noise to the signal. Specifically, it can impact the first and last tap they undergo. As the task instructs the participants to tap their

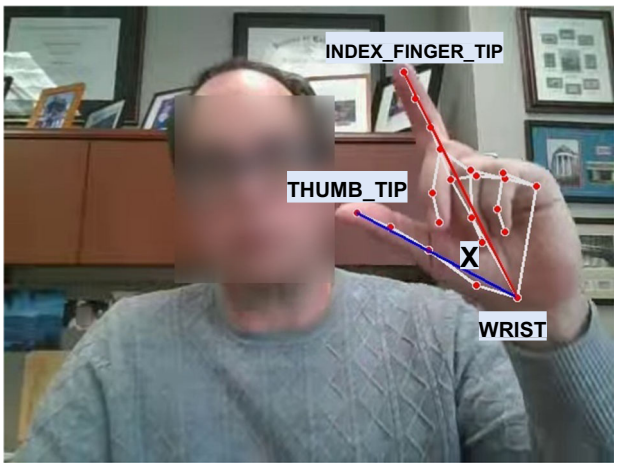

Fig. 6 Hand key points extracted by MediaPipe. All the extracted key-points are displayed as red dots. The three key-points `WRIST`, `THUMB_TIP`, and `INDEX_FINGER_TIP` were used to compute the finger-tapping angle *X*. Authors have obtained consent to publish the images of the participant.

fingers 10 times, we decided to remove the first and last tap, hoping to obtain the cleanest signal to analyze. To accomplish this, we ran a custom peak detection algorithm to find the peaks of the finger-tapping angle, and we removed the portion of the signal before the second peak and after the second-last peak. The peak detection algorithm utilizes some of the unique properties of the task. For example, a peak must be followed by a bottom (i.e., low value of finger-tapping angle) as the tapping is repetitive, the duration between two subsequent taps cannot be too small (i.e., determined by the fastest finger-tapping speed recorded), and the peaks must be bigger than the smallest 25 percentile values of the signal. Figure 5 (right) demonstrates the effectiveness of this step, as it helps to obtain a clean signal that can be used to develop objective measures of the finger-tapping task.

We used the `WRIST`, `THUMB_TIP`, and `INDEX_FINGER_TIP` coordinates to compute the finger-tapping angle X_i for the *i*th frame as shown in Fig. 6. Specifically, for this frame, we first identified two vectors $\overrightarrow{\mathrm{WT}_i}$ and $\overrightarrow{\mathrm{WI}_i}$, which represent the lines connecting the wrist to the thumb-tip and the wrist to the index finger-tip, respectively. The finger-tapping angle X_i was then computed using the following formula:

$$X_i = \cos^{-1}\left(\frac{\overrightarrow{\mathrm{WT}_i}\cdot\overrightarrow{\mathrm{WI}_i}}{|\overrightarrow{\mathrm{WT}_i}|\times|\overrightarrow{\mathrm{WI}_i}|}\right)\times 180^\circ \tag{1}$$

In this context, the symbol $\cdot$ represents the dot product operation performed on two vectors, while $|.|$ denotes l^2 norm. The computed finger-tapping angles were then regarded as a time-series $(X_1, X_2, \cdots, X_n$ where n is the number of frames in the video), which was further processed to reduce noise (as mentioned above.) Let t_{frame} be the average duration of one frame for the video being analyzed (computed by dividing the entire video duration by the number of frames in the video). From the entire time-series, we computed the following metrics for each frame V_i (where $i > 1$):

- *Finger-tapping speed* is a continuous measurement of an individual's tapping speed. For each recorded frame V_i, we quantify speed s_{V_i} as the change in the finger-tapping angle compared to the previous frame and divide this by the duration of one frame (so the speed is measured in degree/second unit):

$$s_{V_i} = \frac{|X_i - X_{i-1}|}{t_{\text{frame}}} \tag{2}$$

The average frame rate of the recorded videos was 30, meaning that we can measure an individual's tapping speed 30 times a second.

- *Acceleration* is also a continuous measurement, which is the derivative of finger-tapping speed. Specifically, for each frame V_i, we measure the change in speed compared to the previous frame to quantify acceleration a_{V_i} in degree/second-square unit:

$$a_{V_i} = \frac{|s_{V_i} - s_{V_{i-1}}|}{t_{\text{frame}}} \tag{3}$$

In order to obtain additional important metrics, we utilized a custom peak detection algorithm to identify the peaks in the finger-tapping angles throughout the duration of the video. Let us denote the frame numbers at which the peaks occur as $P_1, P_2, \cdots, P_k$, where k represents the total number of peaks in that specific video. Consequently, the corresponding peak values of the finger-tapping angles can be represented as $X_{P_1}, X_{P_2}, \cdots, X_{P_k}$. Due to the periodic and repetitive nature of the finger-tapping task, each peak can be used to separate one tap from the others. Utilizing these peak values, we proceeded to calculate the following metrics at each peak P_i (where $i > 1$):

- *Finger-tapping period* is measured as the time (in seconds) it took for a participant to complete a tap. We approximate the period T_{P_i} at the *i*th peak as

$$T_{P_i} = (P_i - P_{i-1}) \times t_{\text{frame}} \tag{4}$$

- *Finger-tapping frequency* is the inverse of the finger-tapping period, measuring the number of taps completed per second. Thus, the frequency f_{P_i} at the *i*th peak is calculated as:

$$f_{P_i} = \frac{1}{T_{P_i}} \tag{5}$$

- *Amplitude* is defined as the maximum angle (in degree) made by the thumb-tip, wrist, and index finger-tip while completing a tap. To simplify, finger-tapping amplitudes (A) are the peak values of finger-tapping angles computed at each peak. Thus, $A_{P_i} = X_{P_i}$.

To track wrist movement throughout the duration of the task, we recorded $W_i = (W_i.x, W_i.y)$, the x and y coordinates of the `WRIST` for each frame V_i. Note that, to account for the variable distance of the hand from the camera, the coordinates were normalized by the Euclidean distance between the `WRIST` and `THUMB_CMC`. For the *i*th frame ($i > 1$), we computed three metrics for capturing the wrist movement:

- Absolute wrist movement along *X*-axis,

$$\Delta WX_i = |W_i.x - W_{i-1}.x| \tag{6}$$

- Absolute wrist movement along *Y*-axis,

$$\Delta WY_i = |W_i.y - W_{i-1}.y| \tag{7}$$

- Cartesian distance of wrist movement,

$$\Delta W_i = \sqrt{(W_i.x - W_{i-1}.x)^2 + (W_i.y - W_{i-1}.y)^2} \tag{8}$$

Please note that while speed, acceleration, and wrist movement metrics were computed for each frame, the other metrics were computed for each detected peak. For each of the features above, we measured the median, inter-quartile range (IQR), mean, minimum, maximum, standard deviation, and entropy (a measure of uncertainty or randomness in a signal, calculated using Shannon's formula), and used them as separate features.

Additionally, we measured the following aggregate features based on the entire signal to capture the rhythmic aspects of the finger-tapping task:

- *Aperiodicity*: Periodicity is a concept borrowed from signal processing that refers to the presence of a repeating pattern or cycle in a signal. For example, a simple sinusoidal signal (e.g., $f(t) = \sin(t)$) will have higher periodicity compared to a signal that is a combination of several sinusoidal signals (e.g., $f(t) = \sin(t) + \sin(2t)$). Aperiodicity measures the absence of periodicity (i.e., the absence of repeating patterns). To measure aperiodicity, signals are transformed into the frequency domain using fast Fourier transformation (FFT). The resulting frequency distribution can be used to calculate the normalized power density distribution, which describes the energy present at each frequency. The entropy of the power density distribution is then computed to measure the degree of aperiodicity in the signal. A higher entropy value indicates greater uncertainty in the frequency distribution, which corresponds to a more aperiodic signal. A similar measure was found to be effective in measuring the symptoms of Alzheimer's disease[18].
- *Number of interruptions*: Interruption is defined as the minimal movement of the fingers for an extended duration. We calculated a distribution of continuous finger-tapping speeds across the study population. Our analysis revealed that over 95 percent of the tapping speeds exceeded 50°/s. As a result, any instance where an individual's finger-tapping speed was less than 50°/s for at least 10 ms was marked as an interruption. The total number of interruptions present in the recorded video was then computed using this method.
- *Number of freezing*: In our study, we considered freezing as a prolonged break in movement. Specifically, any instance where an individual recorded <50°/s for over 20 ms was identified as a freezing event, and we counted the total number of such events.
- *Longest freezing duration*: We recorded the duration of each freezing event and calculated the longest duration among them.
- *Tapping period linearity:* We recorded the tapping period for each tap and evaluated the possibility of fitting all tapping periods using a linear regression model based on their degree of fitness (R^2). Additionally, we determined the slope of the fitted line. The underlying idea was that if the tapping periods were uniform or comparable, a straight line with slope = 0 would adequately fit most periods. Conversely, a straight line would not be an appropriate fit if the periods varied significantly.
- *Complexity of fitting periods*: The complexity of fitting finger-tapping periods can provide insights into the variability of these periods. To measure this complexity, we used regression analysis and increased the degree of the polynomial from linear (degree 1) up to 10. We recorded the minimum polynomial degree required to reasonably fit the tapping periods (i.e., $R^2 \geq 0.9$).
- *Decrement of amplitude*: Decrement of amplitude is one of the key symptoms of Parkinsonism. We measured the finger-tapping amplitude for each tap and quantified how the amplitude at the end differed from the mean amplitude and amplitude at the beginning. Additionally, we calculate the slope of the linear regression fit to capture the overall change in amplitude from start to end.

The abovementioned measurements and some of their statistical aggregates result in 65 features used to analyze the recorded finger-tapping videos. We performed a cross-correlation analysis among the features to identify the highly correlated pairs (i.e., Pearson's correlation coefficient, $r > 0.85$) and dropped one feature from each pair. This helps to remove redundant features and enables learning the relationship between the features and the ground truth severity scores using simple models. This is important, as simple models tend not to over-fit the training data and are more generalizable than complex models (commonly known as Occam's razor[19]). After this step, the number of features was reduced to 53.

For each of the 53 features, we perform a statistical correlation test to identify the features significantly correlated with the ground-truth severity score. Specifically, for each feature, we take the feature values for all the recorded videos in our dataset and the associated ground-truth severity scores obtained by the majority agreement of three expert neurologists. We measure the Pearson's correlation coefficient (r) between the feature values and severity scores and test the significance level of that correlation (i.e., p-value). We found 18 features to be significantly correlated (at a significance level, $\alpha = 0.01$). The significant features include finger-tapping speed (inter-quartile range, median, maximum, minimum), acceleration (minimum), amplitude (median, maximum), frequency (inter-quartile range, standard deviation), period (entropy, inter-quartile range, minimum), number of interruptions, number of freezing, longest freezing duration, aperiodicity, the complexity of fitting periods, and wrist movement (minimum Cartesian distance). Table 2 reports the correlation's direction, strength, and statistical significance level for the most correlated ten features. It is important to acknowledge that certain features may exhibit a strong non-linear correlation that is not adequately captured by Pearson's correlation coefficient. Due to this reason, we retained all 53 features as candidates for training machine learning models, even if some of them did not exhibit significant correlations.

Feature processing, model training, evaluation, and explanation

With a small dataset, there is an increased risk of overfitting, where the model learns the noise and specific patterns of the training data rather than generalizing well to unseen data. Feature selection helps mitigate this risk by reducing the complexity of the model and focusing on the most informative features, which reduces the likelihood of overfitting. We used the Boosted Recursive Feature Elimination (BoostRFE) method implemented in the shap-hypertune Python package with the LightGBM base model to reduce the number of features fed to the machine learning model (Supplementary Table 3 contains all the feature selection approaches we tried and the corresponding model performance measures). BoostRFE combines the concepts of boosting and recursive feature elimination. It identifies and ranks the most informative features in a dataset. After selecting the feature set, we scaled all the features based on training data to bring them onto the same scale. We used the "number of top features" to be selected by the BoostRFE method and the scaling method as hyper-parameters, and the best model picked the top-22 features out of 53 candidate features and StandardScaler (implemented in Python sklearn package) as the scaling method.

To model our dataset, we applied a standard set of regressor models (i.e., XGBoost, LightGBM, support vector regression (SVR), AdaBoost, and RandomForest), as well as shallow neural networks (with 1 and 2 trainable layers). We ran an extensive hyper-parameter search for all of these models using the Weights and Biases tool (https://wandb.ai) and found LightGBM[9] to be the best-performing model (Table 3). LightGBM is a gradient-boosting framework like XGBoost[20] that works by iteratively building a predictive model using an ensemble of decision trees. It uses a leaf-wise tree growth strategy where each tree is grown by splitting the leaf that offers the most significant reduction in the loss function. LightGBM is a popular choice in modeling structured

Table 3. Performance of different regressor models we tested.

Model	MAE	MSE	Accuracy (%)	Kendall's τ	MAPE (%)	PCC	Spearman's ρ
SVR	0.5861	0.5586	**51.94**	0.5044	32.14	0.6388	0.6329
Random forest regressor	0.5920	0.5482	49.28	0.5116	33.5	0.6518	0.6389
AdaBoost regressor	0.5926	0.5499	45.6	0.5038	33.75	0.6219	0.6317
XGBoost regressor	0.5904	0.5553	51.33	0.4989	32.57	0.6417	0.6282
LightGBM regressor	**0.5802**	**0.5364**	50.92	**0.5147**	**32.01**	**0.6563**	**0.6429**
Shallow neural network-I (one trainable layer)	0.6154	0.6007	46.83	0.4810	33.9	0.6097	0.6100
Shallow neural network-II (two trainable layers)	0.6069	0.6162	51.33	0.4813	32.42	0.6004	0.6044

Performance was reported based on leave-one-patient-out cross-validation. For each performance metric, the best value is highlighted in bold text. Some of the metrics are abbreviated for the simplicity of presentation.
MAE mean absolute error (points), *MSE* mean squared error (points), *MAPE* mean absolute percentage error (%), *PCC* Pearson's correlation coefficient.

data due to its fast training speed, low memory usage, and high performance. To run the hyper-parameter search for the LightGBM regressor, we experimented with different learning rates, maximum depth of the tree, number of estimators to use, etc. The list of all hyper-parameters, their search space, and the corresponding value for the best model is reported in Supplementary Table 4.

The dataset we analyzed exhibited class imbalance. Specifically, we had 108, 181, and 141 videos with severity scores of 0, 1, and 2, respectively, while the number of videos with severity scores of 3 and 4 was significantly lower, with only 54 and 5 videos, respectively. To address this issue, we experimented with a technique called Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla et al.[21]). SMOTE generates synthetic data samples for the minority classes by selecting an instance from each minority class and choosing one of its k-nearest neighbors (where k is a user-defined parameter). It then creates a synthetic instance by interpolating between the selected instance and the chosen neighbor. However, the model performance degraded (i.e., Pearson's correlation coefficient decreased from 0.6563 to 0.6422, and the mean absolute error increased from 0.5802 to 0.5807) after integrating SMOTE. Therefore, we ended up not using minority oversampling.

To assess the model's performance, we employed leave-one-patient-out cross-validation (LOPO-CV) technique. LOPO-CV involves partitioning the dataset in a manner where each patient's data is treated as a distinct validation set, while the remaining data is utilized for training the model. This process entails multiple iterations, with each iteration excluding the data samples of a specific patient (both left and right-hand videos in our case) as the test set, while the machine learning model is trained on the remaining data. LOPO-CV ensures that the model's performance is evaluated on unseen patients, resembling real-world scenarios where the model encounters new patients during deployment. This approach is particularly well-suited for machine learning applications in the healthcare domain.

We used seven metrics to evaluate the performance of several regression models attempted to measure the severity of the finger-tapping task: mean absolute error (MAE), mean squared error (MSE), classification accuracy, Kendall rank correlation coefficient (Kendall's τ), mean absolute percentage error (MAPE), Pearson's correlation coefficient (PCC), and Spearman's rank correlation coefficient (Spearman's ρ). MAE is a metric used to measure the average magnitude of the errors in a set of predictions without considering their direction. It is calculated by taking the absolute differences between the predicted and actual values and then averaging those differences. The smaller the MAE, the better the model is performing. MSE is also a measure of the model's error, however, it penalizes the bigger errors more as the errors are squared. Instead of capturing errors on an absolute scale, MAPE measures errors on a relative scale. For example, making a one-point error when the ground-truth value is 4 will be considered a 25% error using MAPE. Both Kendall's τ and Spearman's ρ are popularly used in statistics to measure the ordinal association between two measured quantities. PCC measures the strength and direction of the relationship between two variables and ranges from -1 to $+1$. A correlation of -1 indicates a perfect negative relationship, a correlation of $+1$ indicates a perfect positive relationship, and 0 indicates no relationship between the variables. Finally, accuracy captures the percentage of time the model is absolutely correct (when the regression values are converted to five severity classes). Note that, unlike the other metrics, accuracy does not distinguish between making a one-point error and larger errors, and thus it is not commonly used in regression tasks. In general, these broader sets of metrics provide a more detailed and diverse picture of the model performance.

To explain the model's performance, we used SHapley Additive exPlanations (SHAP). SHAP is a tool used for explaining the output of any supervised machine learning model[10]. It is based on Shapley values from cooperative game theory, which allows us to assign an explanatory value to each feature in the input data. The main idea behind SHAP is to assign a contribution score to each feature that represents its impact on the model's output. These contribution scores are calculated by considering each feature's value in relation to all possible combinations of features in the input data. By doing this, SHAP can provide a detailed and intuitive explanation of why a particular model made a certain decision. This makes it easier for human decision-makers to trust and interpret the output of a machine-learning model.

Use of large language models

ChatGPT (https://chat.openai.com/chat)—a large language model developed by OpenAI (https://openai.com/) that can understand natural language prompts and generate text—was used to edit some parts of the manuscript (i.e., suggest improvements to the language, grammar, and style). All suggested edits by ChatGPT were further verified and finally integrated into the manuscript by an author. Please note that ChatGPT was used only to suggest edits to existing text, and we did not use it to generate any new content for the manuscript.

Ethics

The study was approved by the Institutional Review Board (IRB) of the University of Rochester, and the experiments were carried out following the approved study protocol. We do not have written consent from the participant as the study was primarily administered remotely. However, participants provided informed consent electronically for the data used for analysis and photos presented in the figures.

Reporting summary

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

DATA AVAILABILITY

The recorded videos were collected using a web-based tool. The tool is publicly accessible at https://parktest.net. Unfortunately, we are unable to share the raw videos due to the Health Insurance Portability and Accountability Act (HIPAA) compliance. However, the extracted features and clinical ratings are publicly available: https://github.com/ROC-HCI/finger-tapping-severity. The features are provided in a structured format that can be easily integrated with existing machine-learning workflows.

CODE AVAILABILITY

The codes for video processing and feature extraction, as well as for model training, are publicly available: https://github.com/ROC-HCI/finger-tapping-severity.

Received: 10 April 2023; Accepted: 10 August 2023;
Published online: 23 August 2023

REFERENCES

1. Willis, A., Schootman, M., Evanoff, B., Perlmutter, J. & Racette, B. Neurologist care in Parkinson disease: a utilization, outcomes, and survival study. *Neurology* **77**, 851–857 (2011).
2. Kissani, N. et al. Why does Africa have the lowest number of neurologists and how to cover the gap? *J. Neurol. Sci.* **434**, 120119 (2022).
3. Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **55**, 181–184 (1992).
4. Khan, T., Nyholm, D., Westin, J. & Dougherty, M. A computer vision framework for finger-tapping evaluation in Parkinson's disease. *Artif. Intell. Med.* **60**, 27–40 (2014).
5. Williams, S. et al. Supervised classification of bradykinesia in Parkinson's disease from smartphone videos. *Artif. Intell. Med.* **110**, 101966 (2020).
6. Nunes, A. S. et al. Automatic classification and severity estimation of ataxia from finger tapping videos. *Front. Neurol.* **12**, 2587 (2022).
7. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N.D. *Dataset Shift in Machine Learning*, Vol. 1, 5 (MIT Press, 2008).
8. Langevin, R. et al. The park framework for automated analysis of Parkinson's disease characteristics. In *Proc. ACM Interactive Mobile, Wearable and Ubiquitous Technology*, Vol. 3, 1–22 (Association for Computing Machinery, New York, NY, USA, 2019).
9. Ke, G. et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).
10. Lundberg, S. M. et al. From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Intell.* **2**, 2522–5839 (2020).
11. Perlmutter, J. S. Assessment of Parkinson disease manifestations. *Curr. Protoc. Neurosci.* **49**, 10–1 (2009).
12. Kim, J.-W. et al. Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with Parkinson's disease. *Med. Biol. Eng. Comput.* **49**, 365–371 (2011).
13. Murdoch, B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med. Eth.* **22**, 1–5 (2021).
14. Mhasawade, V., Zhao, Y. & Chunara, R. Machine learning and algorithmic fairness in public and population health. *Nat. Mach. Intell.* **3**, 659–666 (2021).
15. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
16. Clark, L. T. et al. Increasing diversity in clinical trials: overcoming critical barriers. *Curr. Probl. Cardiol.* **44**, 148–172 (2019).
17. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **178**, 1544–1547 (2018).
18. Sharma, R. & Nadkarni, S. Biophysical basis of alpha rhythm disruption in Alzheimer's disease. *Eneuro* **7**, 2 (2020).
19. Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. Occam's razor. *Inf. Process. Lett.* **24**, 377–380 (1987).
20. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proc. 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016).
21. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

ACKNOWLEDGEMENTS

This work was supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under grant W911NF19-1-0029, National Science Foundation Award IIS-1750380, National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number P50NS108676, and Gordon and Betty Moore Foundation.

AUTHOR CONTRIBUTIONS

M.S.I., W.R., P.T.Y., J.L.P., J.L.A., R.B.S., E.R.D., and E.H. conceptualized and designed the experiments. A.A. and S.L. implemented the web-based data collection framework, M.S.I. designed and implemented the feature extraction techniques, and W.R. worked on the machine learning model. M.S.I. worked on data analysis and visualization. M.S.I., W.R., P.T.Y., and E.H. wrote the manuscript. All the authors read the manuscript and provided valuable suggestions for revising it. All authors accept the responsibility to submit it for publication.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00905-9.

Correspondence and requests for materials should be addressed to Md Saiful Islam.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.