Title

Interpretable video-based tracking and quantification of parkinsonism clinical motor states.

Journal

npj Parkinsons Disease, 10(1)

ISSN

2373-8057

Authors

Deng, Daniel
Ostrem, Jill
Nguyen, Vy
et al.

Publication Date

2024-06-25

DOI

10.1038/s41531-024-00742-x

npj | parkinson's disease

Article

Published in partnership with the Parkinson's Foundation

https://doi.org/10.1038/s41531-024-00742-x

Interpretable video-based tracking and quantification of parkinsonism clinical motor states

Daniel Deng[1], Jill L. Ostrem[1], Vy Nguyen[1], Daniel D. Cummins[1], Julia Sun[1], Anupam Pathak[2], Simon Little[1] ✉ & Reza Abbasi-Asl[1,3,4] ✉

Quantification of motor symptom progression in Parkinson's disease (PD) patients is crucial for assessing disease progression and for optimizing therapeutic interventions, such as dopaminergic medications and deep brain stimulation. Cumulative and heuristic clinical experience has identified various clinical signs associated with PD severity, but these are neither objectively quantifiable nor robustly validated. Video-based objective symptom quantification enabled by machine learning (ML) introduces a potential solution. However, video-based diagnostic tools often have implementation challenges due to expensive and inaccessible technology, and typical "black-box" ML implementations are not tailored to be clinically interpretable. Here, we address these needs by releasing a comprehensive kinematic dataset and developing an interpretable video-based framework that predicts high versus low PD motor symptom severity according to MDS-UPDRS Part III metrics. This data driven approach validated and robustly quantified canonical movement features and identified new clinical insights, not previously appreciated as related to clinical severity, including pinkie finger movements and lower limb and axial features of gait. Our framework is enabled by retrospective, single-view, seconds-long videos recorded on consumer-grade devices such as smartphones, tablets, and digital cameras, thereby eliminating the requirement for specialized equipment. Following interpretable ML principles, our framework enforces robustness and interpretability by integrating (1) automatic, data-driven kinematic metric evaluation guided by pre-defined digital features of movement, (2) combination of bi-domain (body and hand) kinematic features, and (3) sparsity-inducing and stability-driven ML analysis with simple-to-interpret models. These elements ensure that the proposed framework quantifies clinically meaningful motor features useful for both ML predictions and clinical analysis.

Parkinson's disease (PD) is a common neurodegenerative disorder characterized by progressive motor symptoms (e.g., bradykinesia, rest tremor, rigidity, postural instability) that can be disabling and significantly impair quality of life[1–3]. The ability to quantify motor symptom progression in PD patients is crucial for assessing and optimizing therapeutic interventions, such as dopaminergic medications and deep brain stimulation (DBS)[4]. Such quantification requires accurate and continual monitoring of motor symptom severity and fluctuations. Currently this objective is only partially satisfied by the *status quo* strategy of intermittent motor assessments assessed at one time point by a single clinician using the Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[5]. Clinical signs such as finger tapping slowness and decrement (bradykinesia) within the MDS-UPDRS have been discovered by clinical heuristics and codified by expert consensus, precluding well-validated, objective, data-driven and quantifiable assessment of patients. Day-to-day fluctuation of symptoms relies on subjective recall from patients often captured by motor diaries[6]. These approaches are limited by high assessment variance, imperfect recall, and recency bias[7]. To overcome these

[1]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [2]Google Inc., Mountain View, Mountain View, CA, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [4]UCSF Weill Institute for Neurosciences, San Francisco, CA, USA. ✉e-mail: Simon.Little@ucsf.edu; Reza.AbbasiAsl@ucsf.edu

limitations, the field needs more reliable and objective tracking of PD clinical states.

Technology-based objective symptom quantification, such as those supported by wearable tracking devices and their ability to record kinematic (movement) data, introduces a potential solution[8–10]. Unfortunately, due to device expenses and technical limitations, embodiments of these quantification systems have yet to be regularly adopted into clinical practice for PD assessment[11–13]. More recently, advancements in machine learning (ML) and computer vision have yielded accessible solutions for extracting kinematic information at key anatomical positions from video data without the need for physical marker systems[14–17]. These computer vision solutions have the potential to address the shortcomings of existing methods, significantly enhance diagnostic accuracy, and open new avenues for optimizing personalized medical therapy in PD[18–22]. However, to date, many video-based diagnostic tools have implementation challenges of expensive and inaccessible technology, often requiring multi-camera setups, pristine video collection protocols, or additional sensors that are infeasible for conventional use. Additionally, typical "black-box" ML implementations are not tailored to be clinically interpretable, either due to complex and unintuitive algorithms or a lack of analysis on feature stability and optimality. Therefore, they are generally ineffective in generating novel clinical insights and are challenging to integrate into current clinical care or critically, to develop clinical oversight for. Finally, these tools often focus on prolonged videos from a formal clinical examination or features from a single motor modality, increasing the burden of video acquisition and missing the opportunity to integrate different domains (e.g., body posture, hand movement, facial expression), which has the potential to significantly increase the accuracy and robustness of ML predictions[23–26]. They also typically only predict metrics directly corresponding to a single modality (i.e., predict only MDS-UPDRS finger tapping score)[22,27]. A truly valuable video-based solution for tracking PD motor symptom progression would need to be affordable, accessible, automated, transparent, and able to obtain rich and clinically relevant metrics for holistic evaluation of PD symptoms[28,29].

In this study, we address these needs by (1) publicly releasing a comprehensive kinematics dataset from 31 patients all with parkinsonism, and (2) developing a video-based framework to automatically predict PD motor symptom high versus low severity according to the MDS-UPDRS Part III metrics (total score). Following interpretable ML principles[30,31], our primary contribution is to enforce model robustness and interpretability by integrating (1) automatic, data-driven kinematic metric evaluation guided by pre-defined digital features of movement, (2) combination of bi-domain (body and hand) kinematic features, and (3) sparsity-inducing and stability-inducing ML analysis with simple-to-interpret models. We perform a comprehensive kinematic feature stability analysis to identify conserved features across ensemble of models[32–34] and feature contributions to model outputs via tree SHAP (Shapley Additive exPlanations) analysis[35]. These elements in our design ensure that the model quantifies clinically meaningful motor features providing new clinical insights for quantification of PD severity, including gait features and finger features, not specifically examined in a classical movement disorders examination. Our framework is enabled by retrospective, single-view videos recorded on consumer-grade devices such as smartphones and digital cameras, thereby eliminating the requirement for specialized equipment. In addition, the framework has the advantage of being able to extract rich and meaningful features from just three to seven seconds of video for efficient training and accurate prediction. Validation of features by ML models are enabled by a leave-one-subject-out cross-validation (CV) scheme, which selects a different patient each time and isolates all samples related to that patient as test samples. This approach has been mirrored in other PD studies to protect against data leakage and ensure unbiased results[36,37].

Results

Clinical and demographic characteristics

Clinical data were obtained from 31 participants all with parkinsonism and who were evaluated at UCSF as part of a multi-day deep phenotyping cohort study. The protocol included standardized video recordings taken in both "on" and "off" dopaminergic medication states while clinical rating scales were performed. Video kinematic data were retrospectively extracted from the individual subject's clinical videos. These kinematic data is publicly released as part of this study. The video clips were collected with a single tablet camera, which is mounted on a fixed stand, in a clinical setting and not initially formalized or collected with optimized settings for a computer vision - ML pipeline. Disease severity was scored using the MDS-UPDRS at the time of assessment. All patients had Parkinsonian symptoms at the time of evaluation. Seven patients were later determined to have Progressive Supranuclear Palsy (PSP) and one patient did not meet clear diagnostic criteria and was classified as PUCS (Parkinsonism of uncertain clinical significance)[38]. Four participants were not taking dopaminergic medication and were only assessed in the "off" medication state. Ten participants exhibited on-state dyskinesia. In terms of wake time spent with dyskinesias (MDS-UPDRS 4.1 score), seven reported less than 25% of their waking day (score = 1) and three reported between 25% to 50% of their waking day (score = 2). In terms of functional impact of dyskinesias on activities and social interactions (MDS-UPDRS 4.2 score), five reported no impact (score = 0) and five reported slight impact (score = 1).

Patients were dichotomized into two groups associated with low and high Parkinsonian motor symptom severity based on the sample median MDS-UPDRS Part III (motor) score of 32, consistent with the cut-off threshold for mild and moderate-to-severe disease severity recommended in literature[39]. The average UPDRS III scores of 18.5 and 43.4 for the low and high impairment groups fall under the mild and moderate-to-severe categories, respectively. Table 1 summarizes the clinical and demographic characteristics of patients with low ($n = 33$) and high ($n = 25$) severity motor symptoms (see also Supplementary Fig. 1).

The dichotomized groups demonstrated well-balanced characteristics with only age and cognitive profile (MoCA) also showing a difference between the high and the lower severity cohort. The group with high motor symptom severity included patients diagnosed as having PSP, a

Table 1 | Summary of clinical and demographic characteristics in patients and associated video clips dichotomized by MDS-UPDRS Part III motor score

	Low Motor Impairment ($n = 33$)	High Motor Impairment ($n = 25$)	p-value
Disease Type	PD = 33, PSP = 0, PUCS = 1	PD = 17, PSP = 7, PUCS = 1	0.00[c]
Age (yrs)	67.1 ± 7.4	71.4 ± 6.2	0.02[a]
Sex	M = 20, F = 13	M = 15, F = 10	0.96[c]
Handedness	R = 28, L = 3, A = 2	R = 21, L = 3, A = 1	0.89[c]
Education (yrs)	18.3 ± 2.5	17.5 ± 2.3	0.22[a]
Disease duration (yrs)	7.4 ± 4.0	5.9 ± 3.0	0.14[a]
Medication status	On = 18, Off = 15	On = 9, Off = 16	0.16[c]
MDS-UPDRS, Part I score	9.4 ± 5.4	10.4 ± 6.3	0.55[a]
MDS-UPDRS, Part II score	8.8 ± 7.5	18.4 ± 13.7	0.01[b]
MDS-UPDRS, Part III score	18.5 ± 10.3	43.4 ± 8.2	0.00[a]
Hoehn and Yahr scale	1.9 ± 0.8	3.0 ± 1.1	0.00[b]
MoCA score	26.5 ± 3.2	23.5 ± 5.0	0.01[b]

PD Parkinson's disease, *PSP* Progressive supranuclear palsy, *PUCS* Parkinsonism of uncertain clinical significance, *MDS-UPDRS* Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale, *MDS-UPDRS* Part I non-motor experiences of daily life, *MDS-UPDRS* Part II motor experiences of daily life, *MDS-UPDRS* Part III motor examination, *MoCA* Montreal Cognitive Assessment.

[a]Independent 2-sample *t*-test, [b]Mann–Whitney *U*-test, [c]Fisher's exact test.

neurodegenerative disorder with similar Parkinsonian motor symptoms to PD but with more rapid progression[40,41]. PSP may also exhibit subtle differences in its patterns of motor deficits as compared to PD (Supplementary Fig. 2)[42]. Select patients could appear in both high and low severity groups if levodopa medication significantly altered their motor symptom severity to move them from high to low severity. Note that the leave-one-subject-out CV protocol ensures that data from the same subject, irrespective of severity group assignment, may be used only during training or validation and never both. Therefore, the risks of dependency on hidden covariates and potential data leakage are implicitly addressed during training and validation by ensuring that the data used for validation stemmed from patients unseen in training.

Automatic extraction of motor features

We designed a computational framework to automatically extract a large array of features representing movement characteristics in raw, unedited video recordings of Parkinsonian patients performing motor tasks (Fig. 1a). A small but highly predictive subset of these features was then selected for training and validation of our ML to predict motor symptom severity quantified by MDS-UPDRS Part III metrics (Fig. 1b). Associated with each

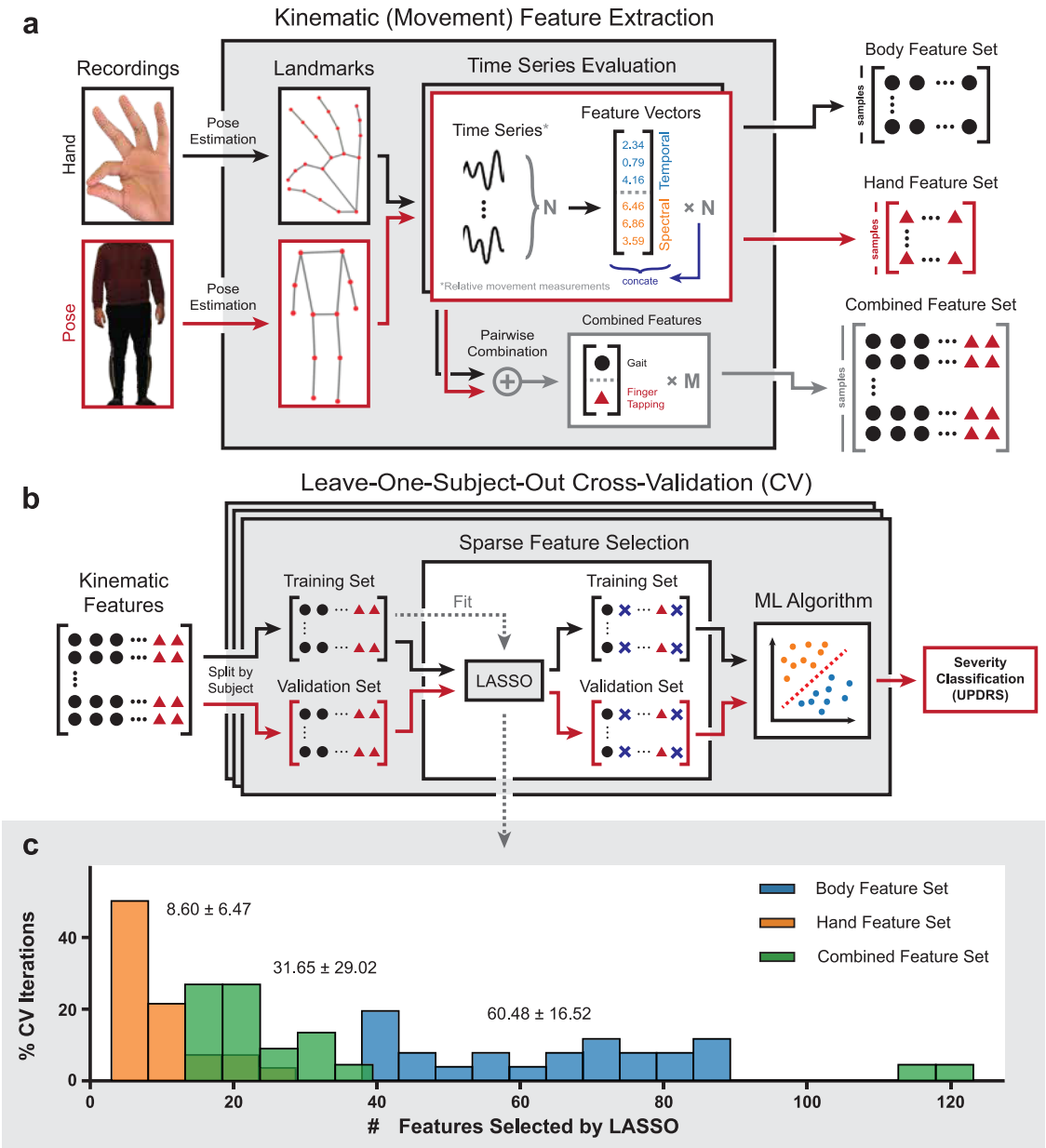

Fig. 1 | Schematic overview of automatic feature extraction from video recordings for classification of Parkinson's disease motor symptom severity. a From recordings of participants performing prescribed motor assessment tasks, we extracted movement (kinematic) time series at key landmarks using the pose estimation library MediaPipe. Relative movement measurements were computed based on the extracted signals, from which various temporal and spectral metrics were computed as features. Pairwise combinations of samples from the same patient under the same medication state were performed on the body and hand feature sets to form the bi-modality combination feature set. **b** To obtain objective measurements of classification performances, we performed a leave-one-subject-out cross-validation (CV), where samples from one patient are held out as the training set for each CV iteration, on each dataset. The CV process is repeated 16 times to account for variabilities in models trained. During CV, a small subset of features with high predictive power with regard to the assigned group labels was selected via least absolute shrinkage and selection operator (LASSO) feature selection based on the training set. The fitted feature selection is then applied to both the training and validation sets. The selected features of the training and validation sets were used to train and validate, respectively, the various machine learning (ML) models. **c** LASSO feature selection consistently identified small sets of salient features during each CV iteration. The average number of features selected at each iteration is reported as mean ± standard deviation.

patient record was a full-body video of walking/gait, a video of the finger-tapping task, or both, collected during the same visit with a standard digital camera. In total, we extracted 40 full-body walking/gait recordings from 25 participants and 48 hand recordings from 27 participants. A comprehensive computer vision pipeline based on deep learning techniques (see Methods) was used to extract kinematic time series from each video recording. The video backgrounds are different but similar clinical scenarios. Since the computer vision pipeline, and CNN-based algorithms in general, is also robust against different backgrounds[43], video backgrounds should not be a significant source of noise variance during extraction. We extracted kinematic time series from 13 major body landmarks[14] from each full-body video and kinematic time-series from 8 major hand landmarks from each hand video recording. The landmarks extracted are 3D, but we chose to discard the noisier depth (z) information in favor of normalization for consistent frame-by-frame scaling and comparability of features. This was followed by filtering and sub-segmentation of the videos. As a result, all full-length videos were segmented into short, salient, and error-free video segments (for examples, see Supplementary Video 1, 2). From full-body videos, we produced 132 video segments with a mean duration of 5.1 ± 1.1 s. From hands videos, we produced 195 video segments with a mean duration of 5.7 ± 1.0 s (Supplementary Fig. 3). Removing those with improper or missing associated data entries, we retained 126 full-body video segments and 189 hand video segments. Based on the clinical dichotomization, 65 body video segments and 83 hands video segments were labeled as exhibiting less severe motor symptoms, whereas 61 and 106, respectively, were labeled as exhibiting more severe motor symptoms. The class assignments for the video segments were roughly balanced with marginally higher membership in the "more severe" category.

Once the landmarks have been extracted, simple first-order relationships between two or three landmarks (e.g., arm-body lateral angle, thumb-index distance; see Methods for complete list of relationships) were defined, computed, and used for subsequent analysis, provided that they make physiological sense and are not occluded in the videos. The concept behind this design was that some or all of the relative kinematic time series should capture aspects of high-level movement features such as stride patterns or finger-tapping consistency. Then, motor features used in classification were computed based on the relationship time series and their time derivatives based on select temporal and spectral kinematic metrics (see Methods). In total, our framework generated 339 features for each of the 126 video segments from the body and 105 features for each of the 189 video segments from the hands. Moreover, to unify the motor modalities, we also integrated body and hand feature sets via pairwise concatenation of body and hand feature vectors associated with the same patient with the same medication status. This procedure created 895 combined feature vectors with 496 features. 417 of the combined vectors are labeled as "less severe" and 478 as "more severe". In all cases, large numbers of salient features were extracted from each video recording without any manual tracking. However, the high-dimensional nature of the feature sets posed a challenge to feature interpretability. To identify a minimal and optimal feature subset for the classification task, we introduced a sparsity-inducing feature selection module based on the least absolute shrinkage and selection operator (LASSO)[44] technique to the classification framework. This module identified on average 60.48 ± 16.52, 8.60 ± 6.47, and 31.65 ± 29.02 features most important for severity prediction among the body, hand, and combined features, respectively (Fig. 1c). The significantly reduced feature set sizes and enabled further analysis and interpretation of trained models for generating clinically relevant insights.

Classification of motor symptom severity based on extracted features

To demonstrate that the automatically extracted and sparsified motor features have high predictive power in discriminating between low and high parkinsonian motor symptom severity states, we trained seven different ML models with the generated features (see Methods). We then quantified the classification performances using classification accuracy and average area under receiver operating characteristics curve (AUC) scores estimated with leave-one-subject-out cross-validation (CV) (Fig. 2, Supplementary Fig. 4; see Methods)[45]. In the interest of retaining sufficient sample sizes, we chose to retain data from patients with clinical parkinsonism secondary to PSP and PUCS as well. We also repeated the analysis with only PD patients, which demonstrated only mildly reduction of classification performances across the board (Supplementary Fig. 5), which could be the combined effects of (1) less statistical power from fewer samples; (2) removing significant influence of PSP and PUCS motor patterns; (3) increased imbalance in low versus high motor symptom groups.

When the models were trained and evaluated on the body features, logistic regression (LR), support vector machine (SVM), and Gaussian naive Bayes (GNB) classifiers achieved the highest average AUC of 0.76, 0.75, and 0.75, respectively. They were closely followed by linear discriminant analysis (LDA) and k-nearest neighbors (KNN) classifiers, at 0.72 and 0.71 average AUC scores, respectively. The remaining ensemble models, random forest (RF) and adaptive boosted trees (AB), had the chance-level average AUC scores of 0.53 and 0.49. Performance measured by classification accuracy were similar in relative ranking, with SVM and LR achieving the highest average classification accuracy of 71% and 70%, respectively. Lower but comparable classification accuracies of 68%, 67%, and 67% were observed in LDA, KNN, and GBN classifiers, respectively. RF and AB achieved near-random accuracies of 54% and 51%. Overall, linear models trained with body features demonstrated decent classification performances. This suggests that the body features, parsed from only a few seconds of walking footage, formed an acceptable representation of the motor characteristics of the corresponding patients.

For the models trained and evaluated using hand features, most classifiers achieved average AUC scores between 0.67 and 0.69, with LR being the most performant at 0.69, closely followed by AB, SVM, and RF at 0.68, 0.67, and 0.67 respectively. LDA and KNN were the least performant at 0.62 and 0.61 AUC. Similarly, most classifiers achieved the average classification accuracy of 66–67%, with the exception of LDA and KNN at 63% and 61%, respectively. Similarly, overall, the classification performances of most models trained with hand features were slightly poorer compared to models trained with body features. This may be explained by the fact that movement characteristics measurable by the finger-tapping task are specific and limited, whereas whole-body motor assessments, such as the walking task used in this investigation, may contain richer information for diagnosing the overall level of motor impairment.

Integrating both body and hand features into a single model sustained or improved the classification performances. SVM and LR achieved the highest average AUC scores at 0.78 and 0.79, outperforming the single modality models. LDA, KNN, and GNB achieved lower AUC scores of 0.71, 0.68, and 0.68, consistent with or slightly lower than their counterparts trained with only body or hand features. RF and AB achieved the lowest scores of 0.63 and 0.61. In terms of classification accuracy, SVM and LR achieved an average accuracy of 72%, with LDA, KNN, and GNB following closely at 68%. The least performant RF and AB classifiers achieved average accuracies of 61% and 58%, which still offered an improved lower bound on the accuracies compared to training only with body features. Overall, integrating body and hand features led to mildly improved performances in the most accurate models and similar performances in the remaining models. The addition of hand features reduced the size of the feature set to evaluate but may or may not have contributed significantly to the overall additional variance explained in the motor characteristics.

Clinical insights from feature stability analysis

The classification performances demonstrated that our framework is capable of extracting optimal features for discriminating low and high motor symptom severity. However, direct interpretation of trained ML models is usually challenging due to variance in LASSO feature selection. The variance was a consequence of distinct data partitioning during CV. To allow insight into the most important features and their contributions to model outputs, we performed ensemble feature stability analysis, aggregating selection

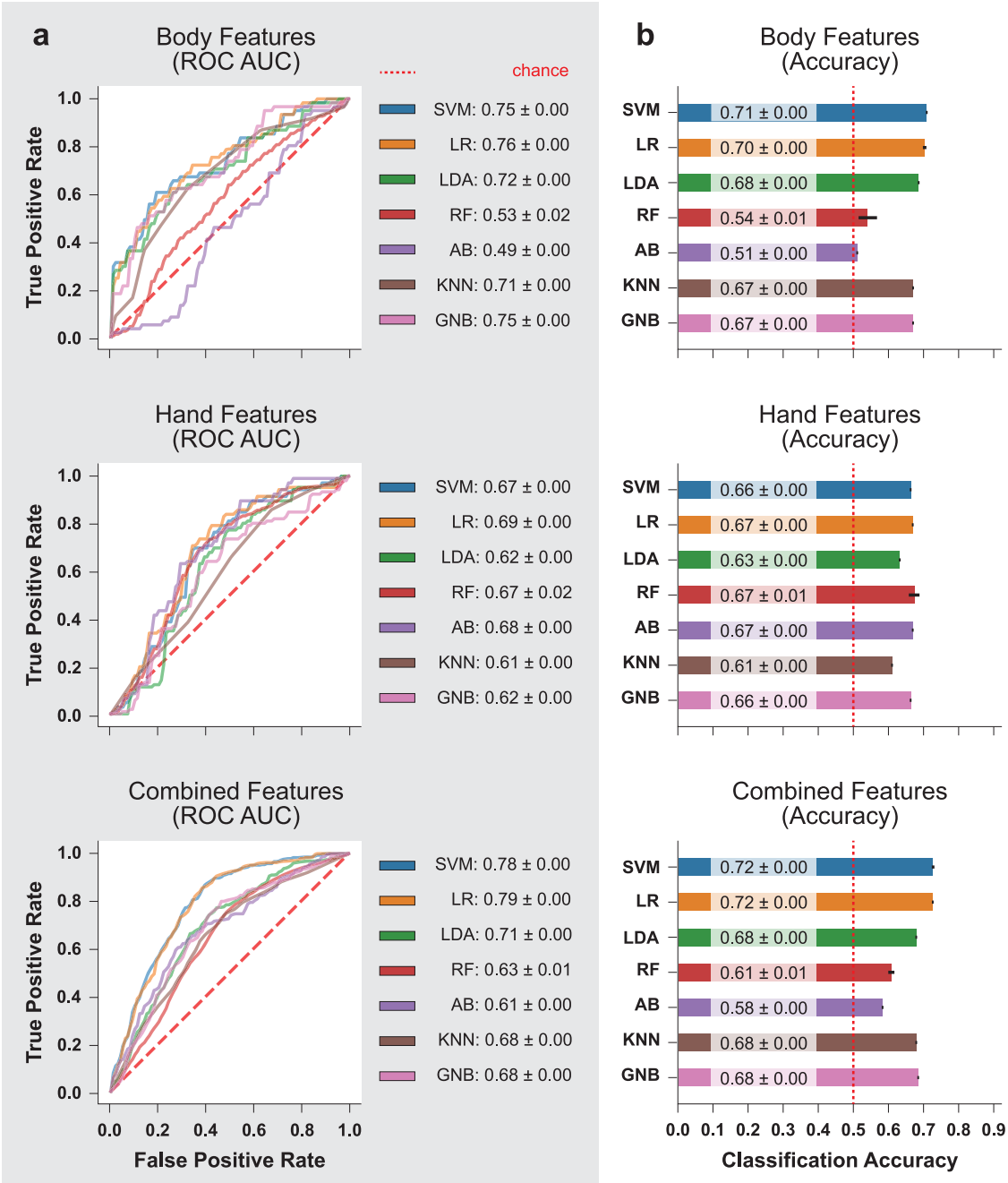

Fig. 2 | Classification performances of seven selected ML classification models. SVM support vector machine, LR logistic regression, LDA linear discriminant analysis, RF random forest, AB adaptive-boosted trees, KNN K-nearest neighbors, GNB Gaussian naive Bayes ROC AUC area under receiver operating characteristics curve. **a** Here we show the average ROC curves of the trained ML models during 16-times repeated leave-on-subject-out cross-validation (CV). The corresponding AUC scores are reported in the right margins as mean ± standard deviation. Overall, integrating body and hand features led to sustained or mildly improved performances in most models. **b** Here we show histograms of the classification accuracies of all models aggregated over all CV iterations. The accuracy scores are reported as mean ± standard deviation at the base of the histogram bars. Similar to the findings when reading the ROC plots, models trained with combined features had comparable accuracies with those trained only on body features, which were more accurate than those trained only on hand features.

counts of features over all leave-one-subject-out CV iterations for each type of classifier model. Specifically, we considered features selected in at least 50% of all iterations as "stable" based on the ensemble paradigm[32], which is a trade-off between retaining all potentially relevant features selected in any CV iteration and retaining only features consistently selected in all iterations to facilitate interpretation.

Performing the analysis on the most performant combined feature set, we identified 9 body and 5 hand features with high stability for interpretation. For all identified features, there exist statistically significant differences ($p \ll 0.05$; independent 2-sample t-test) in group means between the low and high motor symptom severity groups (Fig. 3a). Visualizations of the landmark relationships from which metrics were derived can be seen in Fig. 3c.

The disparities between feature values in the dichotomized groups and their impact on deciding the severity state of patients are compatible with the current clinical understanding of PD and parkinsonism symptoms. Specifically, in patients with higher symptom severity, the model pulled out higher variability in the inter-ankle distance during walking, suggestive of

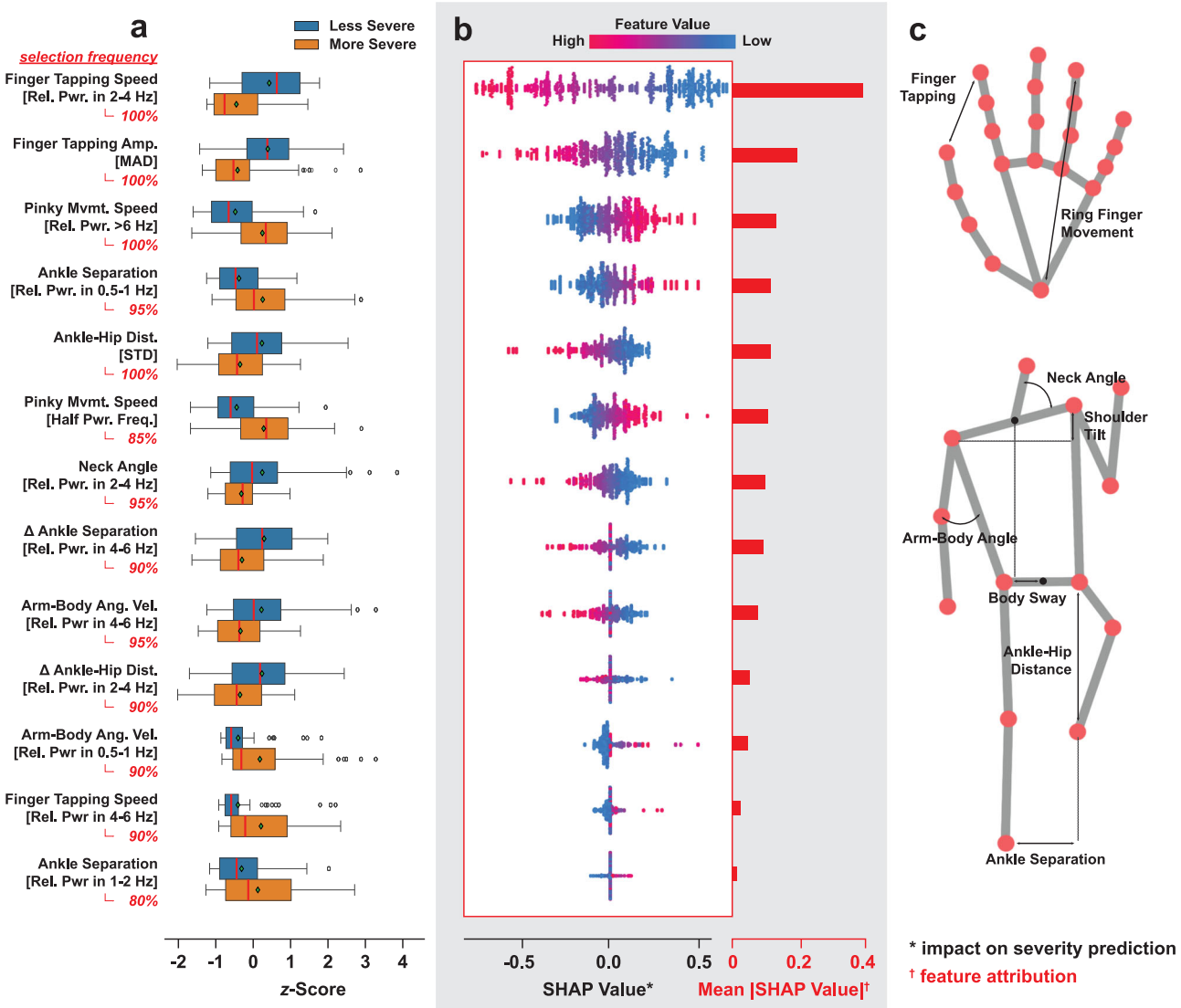

Fig. 3 | Group differences in stable features and their contribution to ML model predictions. STD standard deviation, MAD median absolute deviation. The stable features were chosen according to the criteria that they must be selected in at least 50% of all models trained during cross-validation (CV). **a** Point plot showing the statistically significant differences ($p \ll 0.01$) in group means for the selected features between groups with low and high levels of motor impairment. The feature values are *z*-score normalized. **b** The swarm plots on the left show individual SHAP values of feature values encountered during CV training of a logistic regression (LR) classifier. The histograms on the right show the mean absolute SHAP values of each feature. A higher SHAP value indicates a stronger bias toward predicting the "less severe" class label. A higher mean absolute SHAP value for each feature corresponds to its level of feature attribution in relation to the predicted label. **c** Visualizations of landmark relationships from which kinematic metrics were derived. The body and hand skeletons also show important landmarks and other landmark relationships were based on in our extraction pipeline.

unsteadiness. Additionally, higher movement of the pinkie finger was seen during the finger tapping task which is strictly meant to be restricted to index finger movements only. This may reflect an increase in rigidity or loss of movement discriminability between fingers in more severe clinical Parkinsonian cases. This feature is not classically specifically interrogated in conventional clinical examination of Parkinson's disease. More severe cases also had lower hip-ankle variability during gait, reflective of hip flexion reductions and gait "shuffling". Increased axial rigidity was supported by reduced movement in the neck (relative power in the 2-4 Hz of neck angle) during gait as well as reductions in finger tapping as measured by the finger tapping speed and amplitude.

We also found that the power spectral density for arm-body lateral angular velocity has a distribution shift with the severity of the motor symptoms. Specifically, the relative power in 4–6 Hz of arm-body lateral angular velocity is lower for patients with more severe motor symptoms, while the relative power in 0.5–1 Hz is lower for patients with less severe symptoms (Supplementary Fig. 6). This suggests that with more severe PD motor symptoms, patients have increased lower-frequency changes and reduced higher-frequency changes in arm body lateral angle, suggestive of axial rigidity.

Similarly, there is a distribution shift for the power spectral density of finger tapping speed. Patients with more severe symptoms have higher relative power in 4–6 Hz of their finger tapping speed suggesting jerky/sudden interruptions to regular tapping as this frequency range is higher than the actual tapping frequency. The relative power in 2–4 Hz of the finger tapping speed decreases with severity consistent with incomplete finger tapping for patients with more severe motor symptoms.

The directions of the group differences were consistent with the SHAP values, which measure the impact feature values have on model outputs, estimated with a separate LR classifier using all 14 features (Fig. 3b; see also Supplementary Table 1 for effect sizes of features). In this case, a higher SHAP value biases the model towards assigning a sample the "less severe"

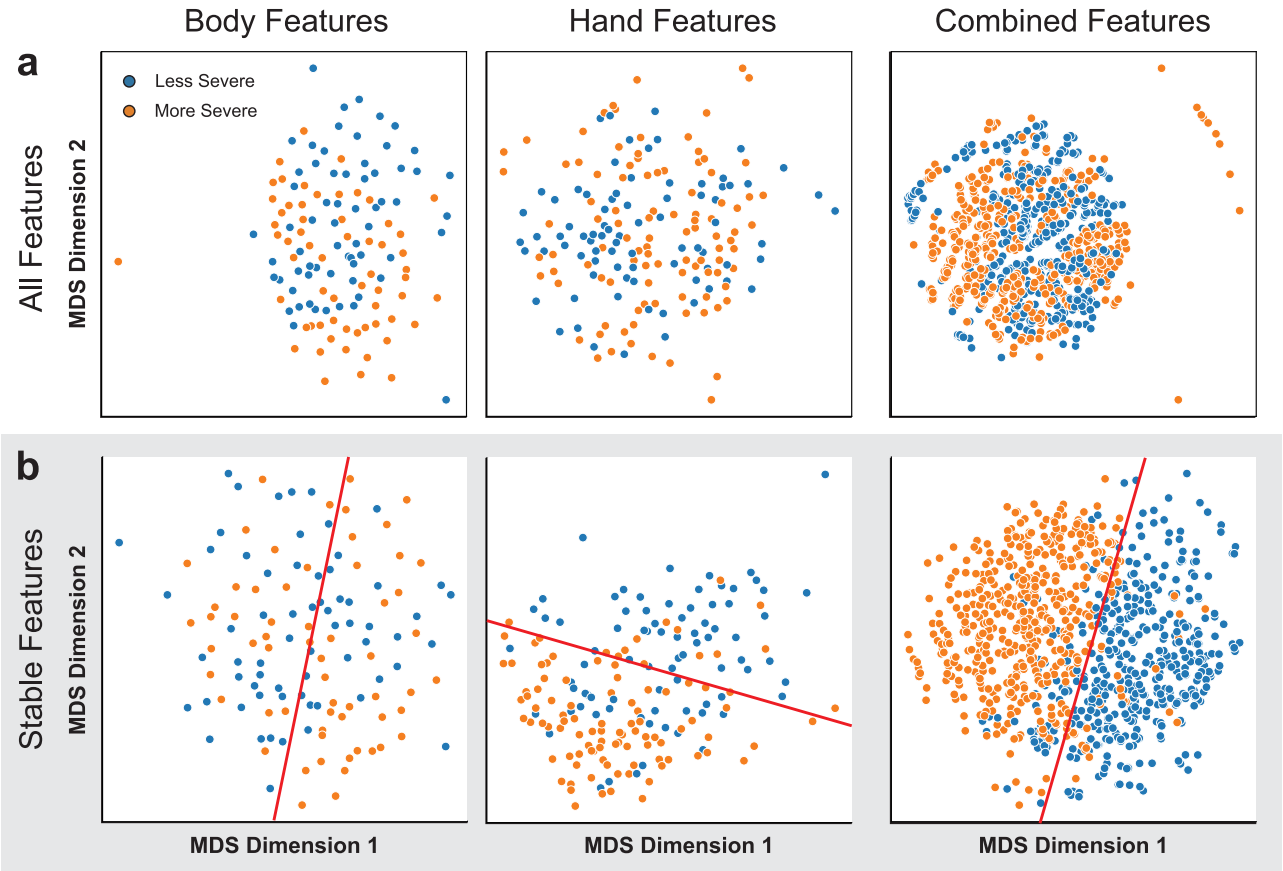

Fig. 4 | Scatter plots comparing 2D projections of datasets with and without stable feature selection. The projections were computed using MDS (multi-dimensional scaling). **a** Poor separability of class clusters in 2D projection space when all features in the datasets were used. **b** Improved separation (especially when both body and hand features were included) of class clusters when only stable features were used.

label. Features whose values have a positive correlation with the estimated SHAP values were consistent with features with higher group means for participants with less severe motor symptoms. Moreover, based on the mean absolute SHAP value of each feature, which reflects overall feature attribution, body features had the most predictive power, higher than hand features associated with finger tapping.

To further validate the optimality of the selected stable features, we compared the 2D projection results of each feature set (body, hand, combined) with and without stable feature selection (Fig. 4). Here, the projection algorithm used was MDS (multidimensional scaling), which tends to preserve the global structures present in the data. When all features generated by our framework were retained prior to projection, the final class clusters were poorly separated in all three cases (Fig. 4a). In contrast, the class clusters saw improved linear separability when only stable features were retained prior to projection. Samples containing body or combined features experienced the most significant increase in separability, consistent with the high classification performance in models trained with them. (Fig. 4b). These results suggest that the stable features identified through ensemble analysis well-characterized the low and high motor impairment groups and are optimal for the classification task at hand. In conclusion, this form of ensemble analysis not only unified the CV iterations and increased model interpretability but also helped to identify and preserve the most important features.

Discussion

Our study demonstrates that an interpretable computer vision - ML pipeline is able to accurately classify Parkinsonian motor severity (high versus low) using very brief video recordings of gait and finger tapping. The classification accuracy was mildly improved using a combined pipeline that incorporated both walking and finger movements, which also condensed the number of stable movement features to evaluate for easy interpretability. The combined model classification was competitive to previous studies that have relied upon complex and extensive hardware solutions including wearable sensors and prolonged, formalized clinical videos. Notably, our study was performed using only very short video clips from a retrospective clinical library; video recordings were not performed in a laboratory under standardized conditions nor optimized for computer vision or machine learning. This supports translation and scalability to real-world applications both in-clinic and potentially extended to outside the clinic.

Clinical movement disorder diagnosis and tracking has shown limited fundamental conceptual advancement over the course of the last 200 years, relying mainly on subjective (expert) recognition and classification of symptoms from visual inspection. This is unsurprisingly significantly limited by inter-rater variability and lack of reproducibility. However, in addition to providing an accurate, objective, evaluation of Parkinsonian motor severity, unlike many ML schemes, our algorithm has the potential to provide, rather than obscure, clinical insight, through bespoke feature stability analysis. Here the algorithm identified a number of classic Parkinsonian features including finger tapping speed and arm swing during gait which serve as useful validation. Moreover, our pipeline identified other features including pinky finger stability during the finger tapping task (likely independent from dyskinesia; see Supplementary Fig. 7) and neck angle features that appear to drive classification and might serve as useful features for clinical classification, that may to date have been underappreciated in medical training. Overall, the stability-driven approach serves to (1) allow clinical validation of identified features (by checking against clinical knowledge), (2) objective quantification of feature importance, and (3) identification of new features with clinical potential. This serves to build

clinical confidence in the pipeline, but this technique could also support improved clinical education with objective validation of clinical phenotyping. A bidirectional relationship between clinical expertise and ML-led feature identification has the potential to improve classification accuracies and clinical knowledge.

Here our algorithm was used to classify high versus low Parkinsonian severity instead of predicting MDS-UPDRS scores directly with regressive models due to limited data availability. In the future, with larger datasets this could be further trained to provide objective, quantitative, diagnostic evaluations that could support classical clinical diagnostic pathways. In addition to single time point evaluations, this approach could in principle be extended to at-home implementation. This has the potential to support chronic symptom tracking in order to monitor naturalistic motor fluctuations over time toward personalized optimization of therapy. Online pose estimation (with immediate deletion of raw video footage) plus automated person recognition have the potential to address potential privacy and security concerns.

The accuracies reported in our study are not directly comparable to existing studies which reported relatively higher classification accuracy differentiating PD from healthy control using movement data (average accuracy of 89.1 ± 8.3%)[46] for three primary reasons: First, our study classifies severity in patients diagnosed with Parkinsonian symptoms (rather than against healthy controls) and attempts to predict high versus low total MDS-UPDRS Part III score. This is a more difficult task compared to the classification of finger tapping or gait scores, which is primarily used in past studies. Second, our approach is designed to use short (~ 5) second-long video segments captured by consumer-grade devices such as smartphones, tablets, simple digital cameras. Out-of-the-box pose estimation by MediaPipe may also, but not always, be suboptimal for subjects under pathological conditions[47–49]. The pipeline may benefit from tracking algorithms trained specifically on precision-labeled, disease-specific data to meet the demands of high accuracy and specificity for clinical use. Previous studies, on the other hand, benefit from the movement data collected using a variety of camera-based, sensor-based, or other miscellaneous recording devices captured over multi-minute-long recordings. Therefore, they lacked the technical simplicity offered by our framework[50]. Third, we prioritize model interpretability over performance to determine highly stable and predictive features of hand and body, providing reliable clinical insights. Tangentially, preliminary analyses showed that the inclusion of non-motor confounding variables, such as cognitive metrics (e.g., MoCA, UPDRS.1) and demographic information (e.g., age, sex), may markedly improve the accuracy of the pipeline as a diagnostic tool. Specifically, the performance-boosting capacity of cognitive metrics illustrates the multi-systemic and cross-domain interactive characteristics of PD. Nevertheless, inclusion of confounding variables could distract from the detection of salient motor features that may inform clinical decisions. Therefore, we chose not to include them in the main analysis. But their potential in aiding automated diagnosis, particularly if interpretability is less of focus, are important and warrant further investigation.

There are a number of limitations to the present study. First, the relatively small cohort size might limit or bias the model performances reported. The distribution of MDS-UPDRS Part III scores of our cohort does not cover the entire score space (0 to 132 points), which could bias the classification performances and feature findings reported. However, UPDRS scoring is a non-linear scale with significant right-sided skew and clinical evidence and published data supporting a median score on the lower side, consistent with our score distribution. Many findings presented herein are also consistent with the current understandings of PD motor characteristics at lower vs. higher PD severity states in general[39,51,52]. These suggest that the pipeline can find clinically valid and meaningful features despite presence of fewer more advanced severity states. Nevertheless, a larger cohort of patients is required to fully validate the framework's reported performance and objectivity, particularly with respect to important features identified. Ideally, the cohort should contain more diverse subjects, with patients at all stages of disease progression and with healthy controls. Having a larger cohort could enable fine-grain analysis such as regression on UPDRS scores to validate the framework's utility in obtaining detailed clinical diagnoses. Moreover, inclusion of healthy controls could allow for direct comparative analysis against competing methods in the context of performance and clinical relevance. Second, the performances reported may be affected by the quality of the video data, which was not collected in a manner specifically designed and optimized for a computer vision task. This has resulted in some suboptimal data unsuitable for ML analysis. Standardization of video acquisition and analysis protocols will be beneficial for the framework's objective assessment of PD symptoms, as well as adaptability and scalability. However, this also demonstrates the robustness of our technique to real-world clinical application. Additional pre-processing should also be explored to increase the pose position estimate validity, including evaluation of a wider suite of available pose estimation computer vision software packages. However, this dataset represents a floor to classification accuracy. Finally, we chose to perform stability analysis based on selection frequency by LASSO during training to promote feature interpretability. However, the method might be overly conservative and ignorant of some important features, especially those with collinearity. Future developments could consider methods such as creating "proto-features" from clusters of correlated features[53,54], which might better preserve important features.

In conclusion, our framework effectively expands upon previous research in PD quantification and addresses many of the shortcomings for a simple yet comprehensive video-based solution. Analyzing a cohort of Parkinsonism patients with the proposed framework, we showed that our approach extracted and identified salient kinematic features that could be used to train accurate ML models for predicting low- and high-severity states for motor impairment with high accuracy. Follow-up studies should focus on further refining the framework, increasing the degree of automation, and validating it in larger, representative cohorts. The framework should also be extended to incorporate additional motor modalities, such as facial expressions and speech, as well as non-motor modalities such as neurological data. Future directions should also include exploring the framework's utility in predicting other clinically relevant outcomes in PD and in application to other neurological movement disorders such as dystonia and essential tremor.

Methods

Participants and assessment of motor symptoms

A cohort of patients presenting with Parkinsonism symptoms, including idiopathic PD and PSP, was recruited at the UCSF Movement Disorder and Neuromodulation Center. Qualitative and Quantitative assessments of motor and non-motor symptoms of participating patients were conducted by a movement disorders neurologist based on MDS-UPDRS, the H&Y scale, and the MoCA scale. The main assessment metric used in this study for measuring motor symptom severity was the MDS-UPDRS Part III score[22,55]. The MDD-UPDRS uses an ordinal scale ranging from 0 to 132, where higher values indicate greater motor impairment. Motor assessments using MDS-UPDRS Part III were conducted for the "off" and "on" dopaminergic medication states. For the "off" state, participants were withdrawn from their medications for a minimum of twelve hours prior to assessment. For the "on" state, participants were given a standard morning dose of Levodopa medication and evaluated one hour later. Not all patients were prescribed dopaminergic medication, and thus some were only assessed in the "off" state. Concurrently with the clinical assessments, video recordings of participants were captured with a consumer-grade iPad tablet recording device at full HD (1920 × 1080) resolution and 30 frames per second, mounted on a tripod in the clinical care area. During recording, participants were instructed to perform finger-tapping and gait tasks as described in MDS-UPDRS sections 3.4 and 3.10. For the finger-tapping task, participants were instructed to tap their index finger on their thumb a minimum of ten times with maximally achievable speed and amplitude. For the gait task, participants were instructed to walk toward and away from the camera for a minimum of 10 meters (30 feet) each way. The resultant videos underwent

no additional editing. All participants provided written consent forms for the use of personal health information in research and release. Privacy and confidentiality protection have been explicitly addressed with UCSF IRB approval. The raw videos were not authorized for release.

Pose estimation and signal processing

For each video, kinematic (movement) time series at select body and hand landmarks (Supplementary Fig. 8) were extracted using the Pose and Hands tracking solutions from MediaPipe[14]. MediaPipe is an open-sourced framework for building multimodal machine learning pipelines. It is cross-platform (server, iOS, Android) and uses a graph-based pipeline to perform processing and inference functions on multimodal input streams, such as vision and audio. Through this framework, we employ pre-trained hand and pose models, which were trained on ~30 k and 85 K (25 k of which performing fitness exercises), respectively. All of these images were annotated by humans[56]. The hand model infers 21 3D landmarks of a hand from a single frame, while the pose model predicts the location of 33 pose landmarks. We found empirically that the depth (z) components of the extracted 3D landmarks are noisier and more inaccurate, thereby not informative to the analysis. Therefore, the depth components are discarded upon extraction, resulting in 2D x,y-coordinates to be normalized downstream. The two-stage pipeline used by MediaPipe for pose and hand tracking consists of (1) an autoencoder detector, similar to feature pyramid networks, for finding bounding box for pose or hand[57], and (2) an convolutional encoder tracker for landmark localization informed by the bounding box. While current state-of-the-art approaches rely primarily on powerful desktop environments for inference, MediaPipe pose and hand models achieve real-time performance on mobile phones. MediaPipe, as well as alternatives such as DeepLabCut and OpenPose, have been adopted for accurate extraction of kinematic data from video recordings for clinical analysis of PD[24,58–60].

Moderate Gaussian smoothing was applied to the extracted signals to reduce random fluctuations due to tracking inaccuracies, implemented as a weighted sum over a 5-point rolling Gaussian window ($\sigma = 0.5$) with the python data manipulation library pandas (version 1.4.0)[61]. Time points at which major tracking errors in any time series, *e.g.* missing data, invalid numerics, severe flickering, occurred were identified and marked for removal. Specifically, severe flickering was detected by checking for rapid zero-crossing in distance between shoulders. Time points at which specific posing requirements were not met were also marked for removal. For body recordings, the requirements were that the subject being filmed must be fully standing and roughly facing forward or backward relative to the camera. We determine if a subject is standing by verifying that the apparent lower leg length (ankle-knee distance) is at most twice the apparent upper leg length (hip-knee distance). We approximately determine if a subject is facing forward or backward by verifying that the horizontal distances of left-wrist-to-left-hip and right-wrist-to-right-hip are of opposite signs, *i.e.* the hands are on opposite sides of the body. For hand recordings, the requirement was that one and only one hand must be raised (and assumed to be engaged in active finger-tapping). We determine this by verifying that the vertical distances of left and right index-finger-to-wrist are of opposite signs, *i.e.* one hand is pointed up and the other down. For each invalid time point identified, a time window within 0.333 s (10 time points) from the invalid point was removed from all extracted time series.

As a byproduct, the filtering operation effectively segmented the movement data, resulting in sets of short time series at key landmarks grouped by video segments. To equalize the video segments with respect to duration (and by extension the amount of contained information), segments with a duration less than 3 s were discarded and segments with a duration greater than 8 s were further sub-segmented. The minimum and maximum duration values were empirically chosen such that each segment contained sufficient kinematic information, *i.e.* multiple stride or finger-tapping cycles, for non-trivial extraction of motor features while maximizing the number of samples available for subsequent model training and classification. Although the segmentation process is partially driven by tracking error, there are no statistically significant differences in segment durations between low and high severity groups (Supplementary Fig. 3c).

Computing relative movement measurements

Similar to strategies seen in existing literature[60], we generated relational time series from the filtered and segmented kinematic time series based on predefined interactions between two or more landmarks. In most studies, a small subset of distance or angular relationships between select landmarks may be chosen to be evaluated with select few (sometimes complex) metrics based on experimental or clinical expectations. Here, we chose to include many first-order relationships between available landmarks as long as they have physical and physiological meaning (e.g. limb or body angle, distance, speed) and are not occluded in the videos, followed by blanket application of a relatively large set of kinematic and spectral metrics (e.g., mean, standard deviation, spectral entropy). These relationships between landmarks, which are time series, collectively capture aspects of healthy and pathological movement characteristics without explicit tailoring of feature metrics to each relationship. This process enables automation, increases information to include, and encourages novel feature discoveries. The list of time series selected is provided here:

1. Neck angle (angle between nose, mid-shoulder, and left shoulder)
2. Arm-body lateral angle (angle between elbow, shoulder, and hip, lateral to body)
3. Left/right wrist-shoulder distance
4. Left/right ankle-hip distance
5. Ankle separation (horizontal displacement between ankles)
6. Knee separation (horizontal displacement between knees)
7. Body sway (horizontal displacement between mid-shoulder and mid-hip)
8. Leg raise (vertical displacement between ankles)
9. Shoulder tilt (vertical displacement between shoulders)
10. Hip tilt (vertical displacement between hips)
11. Finger tapping (distance between thumb and index finger of active hand)
12. Middle finger movement (distance between middle finger and wrist of active hand)
13. Ring finger movement (distance between ring finger and wrist of active hand)
14. Pinky finger movement (distance between pinky finger and wrist of active hand)

The pose features listed above reflect distance, displacement, and angle measures between landmarks. To produce information on speed, velocity, and angular velocity for analysis, we also compute the first-order derivatives of the time series, resulting in equivalent numbers of additional time series for further processing. Note that the filtering step ensured that only one hand will be actively performing finger-tapping for any given time segment; therefore, distinguishing between left and right-sided hand movements was unnecessary as the resting hand will necessarily have little to no movement and could have its time series discarded.

Time series assessment with kinematic metrics

Once we computed the relational time series, as the apparent body and hand sizes change depending on the distance between the subject and the camera, we normalized the new time series at each time point to equalize the scale and allow direct comparison. For body signals, normalization was achieved by dividing the time series by the body length (distance between mid-shoulder and mid-hip) and centered mid-hip. For hand signals, normalization was achieved by dividing the time series by the palm length (distance between wrist and midpoint between bases of index and pinky fingers) and centering at the wrist. The normalization was followed by kinematic metric evaluation of each time series, during which various temporal (distribution-based) and spectral (spectral density-based) metrics were used to provide a statistical summary of the body and hand movement characteristics (Supplementary Table 2). This process was repeated for the derivatives of the

relational time series, calculated as the rate of change between adjacent time points.

The computed metric values formed the feature vectors associated with each time segment and were used in subsequent ML training and classification. Since only one hand is performing active finger-tapping at any time in the videos, features from the non-active hand are discarded. Considering the common unilateral development of PD symptoms, we needed to retain both left and right-sided body features. This posed a challenge to compare features between subjects whose manifestation of unilateral features might be on opposite sides of the body. To address this, we recategorized left and right-sided features of the same type as "minimum" and "maximum" features by ranking numerically, thereby eliminating the sidedness of features. In addition, to unify the body and hand kinematic features and increase classification performance, we generated combined feature vectors from all valid combinations of body and hand feature vectors, so long as they were associated with the same patient record, *i.e.* same patient under the same medication status.

Feature selection and classification

The scikit-learn package (version 1.1.2) in Python was used for feature selection and classification on low and high motor symptom severity for all three feature sets generated by our framework. To generate training and validation data in a robust manner, we conducted the aforementioned procedure with a custom leave-one-subject-out CV, repeated 16 times to account for model variability due to hyperparameter tuning. The leave-one-subject-out protocol selects a different patient at each CV iteration and isolates all samples related to that patient as the testing set. This form of CV is necessary to prevent dependency on hidden confounding variables and thus data leakage. Once partitioned, the training and validation sets were centered and normalized feature-wise based on means and standard deviations from the training set. Then, to reduce feature dimensionality, address multicollinearity, and improve interpretability, we applied to all feature sets as a pre-training step LASSO feature selection, which is an L1-regularized, sparsity-inducing algorithm. Finally, classification performances were evaluated using seven different ML algorithms: (1) linear discriminant analysis (LDA)[62]; (2) logistic regression (LR)[63]; (3) support vector machine (SVM)[64]; (4) random forest (RF)[65]; (5) adaptive-boosted trees (AdaBoost; AB)[66]; (6) K-nearest neighbors (KNN)[67]; and, (7) Gaussian naive Bayes (GNB)[68]. The regularization parameter for LASSO feature sparsity, as well as relevant hyperparameters for each classification model, were chosen via hyperparameter optimization with nested leave-one-subject-out CV during each CV iteration. Classification accuracies and AUC scores were recorded for each classifier, where accuracy measures the model's ability to classify on the defined class labels and AUC measures the model's sensitivity and generalizability by looking at all probabilities for assigning labels. Accuracy and AUC scores during CV were aggregated and reported as mean ± standard deviation.

Feature stability analysis

To address the challenge of interpreting membership variance in salient feature subsets selected via LASSO during CV, we performed feature stability analysis based on frequencies of selection. Here, we defined a feature as being stable if it was selected by LASSO in at least 50% of all CV iterations. Once the stable features have been identified, their contributions to model predictions were evaluated with SHAP analysis on the most performant model using the python package SHAP (version 0.40.0). Additional validation was performed by comparing the 2D projections of the datasets with and without stable feature selection. The algorithm used for projection was multidimensional scaling (MDS), an unsupervised dimensionality reduction algorithm that preserves global structures in the data.

Statistical analysis

Continuous variables were presented as mean ± standard deviation and compared between low and high motor scoring groups with an independent 2-sample *t*-test if normally distributed or with the Mann–Whitney *U*-test if otherwise. Categorical variables were presented as counts and compared between scoring groups with Fisher's exact test, which is analogous to the Chi-squared test but suitable for small-sized samples. A *p*-value less than 0.05 was considered statistically significant.

Data availability

We have publicly released the kinematic data from 31 participants with Parkinsonism. The data is freely available at https://github.com/abbasilab/Video-Tracking-PD.

Code availability

The software package is freely available at https://github.com/abbasilab/Video-Tracking-PD.

Received: 9 November 2023; Accepted: 14 June 2024;
Published online: 25 June 2024

References

1. Fasano, A. et al. Characterizing advanced Parkinson's disease: OBSERVE-PD observational study results of 2615 patients. *BMC Neurol.* **19**, 50 (2019).
2. Sanchez-Luengos, I. et al. Predictors of health-related quality of life in Parkinson's disease: the impact of overlap between health-related quality of life and clinical measures. *Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil.* **31**, 3241–3252 (2022).
3. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet Lond. Engl.* **397**, 2284–2303 (2021).
4. Armstrong, M. J. & Okun, M. S. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA* **323**, 548–560 (2020).
5. 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. J. Mov. Disord. Soc.* **23**, 2129–2170 (2008).
6. Löhle, M. et al. Validation of the PD home diary for assessment of motor fluctuations in advanced Parkinson's disease. *Npj Park. Dis.* **8**, 69 (2022).
7. Zolfaghari, S. et al. Self-Report versus Clinician Examination in Early Parkinson's Disease. *Mov. Disord. J. Mov. Disord. Soc.* **37**, 585–597 (2022).
8. Huckvale, K., Venkatesh, S. & Christensen, H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *Npj Digit. Med.* **2**, 88 (2019).
9. Matias, R., Paixão, V., Bouça, R. & Ferreira, J. J. A Perspective on Wearable Sensor Measurements and Data Science for Parkinson's Disease. *Front. Neurol.* **8**, 677 (2017).
10. Navani, D., Block, V., Cree, B. & Abbasi-Asl, R. Diurnal Step Count Patterns in Progressive Multiple Sclerosis (P3-3.011). *Neurology* **100**, 3320 (2023).
11. Habets, J. G. V. et al. Rapid Dynamic Naturalistic Monitoring of Bradykinesia in Parkinson's Disease Using a Wrist-Worn Accelerometer. *Sensors* **21**, 7876 (2021).
12. Pang, Y. et al. Automatic detection and quantification of hand movements toward development of an objective assessment of tremor and bradykinesia in Parkinson's disease. *J. Neurosci. Methods* **333**, 108576 (2020).
13. Khodakarami, H. et al. Prediction of the Levodopa Challenge Test in Parkinson's Disease Using Data from a Wrist-Worn Sensor. *Sensors* **19**, 5153 (2019).
14. Lugaresi, C. et al. MediaPipe: A Framework for Building Perception Pipelines. https://doi.org/10.48550/ARXIV.1906.08172. (2019)
15. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
16. Graving, J. M. et al. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* **8**, e47994 (2019).

17. Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. https://doi.org/10.48550/ARXIV.1611.08050. (2016).
18. Park, K. W. et al. Machine Learning–Based Automatic Rating for Cardinal Symptoms of Parkinson Disease. *Neurology* **96**, e1761–e1769 (2021).
19. Espay, A. J. et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. *Mov. Disord. J. Mov. Disord. Soc.* **34**, 657–663 (2019).
20. Sibley, K. G., Girges, C., Hoque, E. & Foltynie, T. Video-Based Analyses of Parkinson's Disease Severity: A Brief Review. *J. Park. Dis.* **11**, S83–S93 (2021).
21. Van Kersbergen, J. et al. Camera-based objective measures of Parkinson's disease gait features. *BMC Res. Notes* **14**, 329 (2021).
22. Morinan, G. et al. Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population. *Npj Park. Dis.* **9**, 10 (2023).
23. Baltrusaitis, T., Ahuja, C. & Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
24. Lim, W. S. et al. An integrated biometric voice and facial features for early detection of Parkinson's disease. *Npj Park. Dis.* **8**, 145 (2022).
25. Makarious, M. B. et al. Multi-modality machine learning predicting Parkinson's disease. *Npj Park. Dis.* **8**, 35 (2022).
26. Rupprechter, S. et al. A Clinically Interpretable Computer-Vision Based Method for Quantifying Gait in Parkinson's Disease. *Sensors* **21**, 5437 (2021).
27. Sarapata, G. et al. Video-Based Activity Recognition for Automated Motor Assessment of Parkinson's Disease. *IEEE J. Biomed. Health Inform.* **27**, 5032–5041 (2023).
28. Lonini, L. et al. Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *Npj Digit. Med.* **1**, 64 (2018).
29. Sica, M. et al. Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review. *PLoS One* **16**, e0246528 (2021).
30. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl Acad. Sci. USA* **116**, 22071–22080 (2019).
31. Abbasi-Asl, R. & Yu, B. Structural Compression of Convolutional Neural Networks with Applications in Interpretability. *Front Big Data* **4**, 704182 (2021).
32. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. *Inf. Fusion* **35**, 132–147 (2017).
33. Tolosi, L. & Lengauer, T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* **27**, 1986–1994 (2011).
34. Nogueira, S., Sechidis, K. & Brown, G. On the stability of feature selection algorithms. *J. Mach. Learn Res.* **18**, 6345–6398 (2017).
35. Lundberg, S. M. et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
36. Gunduz, H. Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. *IEEE Access* **7**, 115540–115551 (2019).
37. Jha, A. et al. The CloudUPDRS smartphone software in Parkinson's study: cross-validation against blinded human raters. *NPJ. Park. Dis.* **6**, 36 (2020).
38. Zitser, J. et al. Parkinsonism of uncertain clinical significance (PUCS): A proposed new diagnostic entity. *J. Neurol. Sci.* **451**, 120696 (2023).
39. Martínez-Martín, P. et al. Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. *Park. Relat. Disord.* **21**, 50–54 (2015).
40. Egerton, T., Williams, D. R. & Iansek, R. Comparison of gait in progressive supranuclear palsy, Parkinson's disease and healthy older adults. *BMC Neurol.* **12**, 116 (2012).
41. Cordato, N. J., Halliday, G. M., Caine, D. & Morris, J. G. L. Comparison of motor, cognitive, and behavioral features in progressive supranuclear palsy and Parkinson's disease. *Mov. Disord.* **21**, 632–638 (2006).
42. Ling, H., Massey, L. A., Lees, A. J., Brown, P. & Day, B. L. Hypokinesia without decrement distinguishes progressive supranuclear palsy from Parkinson's disease. *Brain* **135**, 1141–1153 (2012).
43. Wang, X., Garg, S., Tran, S. N., Bai, Q. & Alty, J. Hand tremor detection in videos with cluttered background using neural network based approaches. *Health Inf. Sci. Syst.* **9**, 30 (2021).
44. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996).
45. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **4**, 627–635 (2013).
46. 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).
47. Seethapathi, N., Wang, S., Saluja, R., Blohm, G. & Kording, K. P. Movement science needs different pose tracking algorithms. https://doi.org/10.48550/ARXIV.1907.10226. (2019)
48. Friedrich, M. et al. Visual perceptive deep learning for smartphone video-based tremor analysis: VIPER-Tremor. Preprint at https://doi.org/10.21203/rs.3.rs-3692906/v1 (2023).
49. Friedrich, M. U. et al. Smartphone video nystagmography using convolutional neural networks: ConVNG. *J. Neurol.* **270**, 2518–2530 (2023).
50. Kour, N., Sunanda & Arora, S. Computer-Vision Based Diagnosis of Parkinson's Disease via Gait: A Survey. *IEEE Access* **7**, 156620–156645 (2019).
51. Evers, L. J. W., Krijthe, J. H., Meinders, M. J., Bloem, B. R. & Heskes, T. M. Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS. *Mov. Disord.* **34**, 1480–1487 (2019).
52. Lang, A. E. et al. Deep brain stimulation: Preoperative issues. *Mov. Disord.* **21**, S171–S196 (2006).
53. Faletto, G. & Bien, J. Cluster Stability Selection. https://doi.org/10.48550/ARXIV.2201.00494 (2022).
54. Ghosal, G. & Abbasi-Asl, R. Multi-modal prototype learning for interpretable multivariable time series classification. *ArXiv Prepr. ArXiv210609636* (2021).
55. Grover, S., Bhartia, S., Akshama, Yadav, A. & Seeja, K. R. Predicting Severity Of Parkinson's Disease Using Deep Learning. *Procedia Comput Sci.* **132**, 1788–1794 (2018).
56. Bazarevsky, V. et al. BlazePose: On-device Real-time Body Pose tracking. https://doi.org/10.48550/ARXIV.2006.10204 (2020).
57. Lin, T.-Y. et al. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2117–2125 (IEEE, Honolulu, HI, 2017).
58. Li, M. H., Mestre, T. A., Fox, S. H. & Taati, B. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *J. Neuroeng. Rehabil.* **15**, 97 (2018).
59. Güney, G. et al. Video-Based Hand Movement Analysis of Parkinson Patients before and after Medication Using High-Frame-Rate Videos and MediaPipe. *Sensors* **22**, 7992 (2022).
60. Dentamaro, V., Impedovo, D. & Pirlo, G. Gait Analysis for Early Neurodegenerative Diseases Classification Through the Kinematic Theory of Rapid Human Movements. *IEEE Access* **8**, 193966–193980 (2020).
61. McKinney, W. Data Structures for Statistical Computing in Python. In *Proc. of the 9th Python in Science Conference* (SciPy, Austin, Texas, 2010).

62. Tharwat, A., Gaber, T., Ibrahim, A. & Hassanien, A. E. Linear discriminant analysis: A detailed tutorial. *AI Commun.* **30**, 169–190 (2017).
63. Schober, P. & Vetter, T. R. Logistic Regression in Medical Research. *Anesth. Analg.* **132**, 365–366 (2021).
64. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **13**, 18–28 (1998).
65. Breiman, L. Random Forests. *Mach. Learn* **45**, 5–32 (2001).
66. Freund, Y., Schapire, R. E. & AT&T Labs. A Short Introduction to Boosting. *J. Japanese Soc. Artificial Intell.* **14**, 771–780 (1999).
67. Taunk, K., De, S., Verma, S. & Swetapadma, A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* 1255–1260 (2019).
68. Hand, D. J. & Yu, K. Idiot's Bayes? Not so stupid after all? *Int Stat. Rev.* **69**, 385–398 (2001).

Acknowledgements

The authors would like to thank Chirag Sharma, Russell Ro, Ian Bledsoe, Ethan Brown, Howie Rosen, Carlie Tanner, Michael Geschwind, Kevin Lieu, and Bruce Miller for their comments and contributions to the recruitment and evaluation of the patients. R.A., S.L., and J.L.O. would like to acknowledge support from the UCSF Innovation Ventures.

Author contributions

Study conception and design: R.A., S.L., J.L.O. Data collection: V.N., D.D.C., J.S., S.L., J.L.O. Analysis and modeling: D.D., R.A. with inputs from A.P. Interpretation of results: D.D., R.A., S.L., J.L.O. Manuscript preparation: D.D., R.A., S.L., J.L.O. with inputs from A.P., V.N., D.D.C., J.S. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

MediaPipe is a free and open-source software package, and we implemented a publicly available version of this with all data analysis performed within UCSF. A.P. provided advice and technical advice only. The primary clinical study through which the data was collected was funded by an anonymous Philanthropy gift to the UCSF Movement Disorders and Neuromodulation Center. The UCSF analysis pipeline covering machine learning analysis of kinematic data (independent of specific pose estimation software) for PD severity classification has been filed as a provisional patent (no. 63/530,566). All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41531-024-00742-x.

Correspondence and requests for materials should be addressed to Simon Little or Reza Abbasi-Asl.