PDWearML: Leveraging Daily Activities for Fast Parkinson's Disease Severity Assessment with Wearable Machine Learning

Xulong Wang, Xiyang Peng, Zheyuan Xu, Mingchang Xu, Yun Yang, Menghui Zhou, Zhong Zhao, Peng Yue and Po Yang *Senior Member, IEEE*

***Abstract*— *Objective:* Achieving effective and robust free-living PD severity assessment with wearable intelligence technologies requires a deep understanding of clinically relevant features, representative activities, and machine learning algorithms. *Methods:* We designed a unified analytic framework (PDWearML) to optimise wearable ML approaches with simple daily activities for fast assessment of PD severity. It comprises annotation criteria, feature importance analysis, representative activity combination, and PD severity assessment. We conducted a 12-month study, developing a supervised PD wearable dataset containing 100 PD patients and 35 age-matched healthy controls using Huawei smartwatches and Shimmer. PD severity, assessed by trained physicians using the Hoehn and Yahr (H&Y) scale. *Results:* The results reveal that through optimising multi-level feature extraction and combining three representative daily activities (WALK, ARISING-FROM-CHAIR, and DRINK), our smartwatch-based machine learning approach can assess PD severity in supervised settings within 2 minutes with an accuracy of up to 84.7%. *Significance:* This work holds significant clinical value, offering a potential auxiliary tool for faster, more tailored interventions in PD healthcare. Code is available at code ocean platform and https://github.com/wang-xulong/PDWearML.**

***Index Terms*—Parkinson's disease, fast assessment, subject adherence, wearable intelligence, activities of daily living**

I. Introduction

WITH notable advancement of machine learning (ML) techniques, wearable intelligence (WI) has made significant strides in developing intelligent early-warning and self-management solutions for patients both in hospital and at home [1]–[3]. In particular, WI technology is transforming Parkinson's disease (PD) management by shifting from traditional, hub-based healthcare systems to more personalised, free-living healthcare environments. This approach allows for continuous monitoring and accurate self-assessment, giving clinicians a more comprehensive view of patient conditions to support individualised care. For instance, Apple® engineers have developed a smartwatch-based ambulatory monitoring system that remotely tracks fluctuations in resting tremor and dyskinesia [4]. Additionally, the Personal Kinetigraph(PKG®), an FDA-cleared medical device, is designed to provide a continuous, objective assessment of movement disorder symptoms, such as slowness of movement, stiffness, tremor, and dyskinesias in free-living environments for PD patients [5], [6]. Although these WI technologies are intended to provide fast and accurate assessment of PD patients in real-life settings, achieving high precision for detecting subtle motor fluctuations across diverse PD activities remains a significant challenge. This requires advanced, reliable ML algorithms, cost-sensitive hardware, and clinically relevant data, including kinematic analysis.

Theoretically, assessing PD in free-living conditions using WI technologies can be framed as a human activity recognition and fine-grained classification problem. This approach first requires extracting and learning versatile, representative features from wearable data, followed by the application of suitable ML algorithms to tasks such as activity recognition and PD severity classification. These tasks facilitate personalised assessments, symptom monitoring, and anomaly detection for effective PD self-management in real-world settings. Over recent decades, ML techniques have shown considerable success in identifying PD symptom characteristics when applied to large-scale, multi-variable wearable datasets [7]. WI solutions powered by ML have the potential to surpass standard clinical scales by capturing subtle changes in real-time, enabling more precise and sensitive tracking of motor function [8]. This advancement has prompted the Food and Drug Administration (FDA) to develop new protocols for evaluating the safety and efficacy of ML-based healthcare technologies [9]. These protocols emphasise (1) supporting research on "patient-centred approaches with transparency to users" and (2) developing methodologies for evaluating and improving ML methods. Consequently, creating analytical frameworks to evaluate the effectiveness and robustness of ML algorithms in free-living environments

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership under Grant EP/T517835/1 and China Scholarship Council 202107030007. (Corresponding authors: Po Yang and Yun Yang)

Xulong Wang, Xiyang Peng, Menghui Zhou, Peng Yue and Po Yang are with the School of Computer Science, University of Sheffield, S10 2TN, Sheffield, UK (e-mail: {xl.wang; xpeng24; menghui.zhou; pyue1; po.yang}@sheffield.ac.uk).

Zheyuan Xu, Mingchang Xu, Yun Yang are with the National Pilot School of Software, Yunnan University, Kunming 650106, China (e-mail: xuzheyuan@stu.ynu.edu.cn; mingchangxu@mail.ynu.edu.cn; yangyun@ynu.edu.cn).

Zhong Zhao is with Yunnan First People's Hospital, Yunnan 650021, China (e-mail: wasx-1128new@163.com)

will be crucial to advancing patient-centred PD healthcare.

To analyse the effectiveness of ML algorithms in PD healthcare, researchers have evaluated classic ML methods using extensive real-world data from daily living, combined with clinical scores and patient profiles, both cross-sectionally and longitudinally [10], [11]. Although valuable, these findings are constrained by the limited availability and quality of free-living wearable datasets. This progress has likely been slowed by challenges related to feature reliability, activity representation, and model estimation: i) Feature Reliability. The reliability of key representative features of PD symptoms in free-living environments remains largely untested. While emerging digital features derived from wearable sensors offer the potential for continuous and remote monitoring of PD symptoms, their lack of specificity in free-living conditions has limited their practical utility [4], [5], [12]–[15]. ii) Activity Representation. Free-living environments encompass diverse daily activities, yet it remains unclear which better assesses PD severity. Although recent studies have identified certain activities relevant for evaluating specific PD motor symptoms, analysing PD symptoms in isolation has proven insufficient for accurately assessing overall disease severity [16]–[21]. iii) ML Model Estimation. Identifying the appropriate ML model for accurate PD severity classification remains an open question. Specifically, given the challenges of collecting and labelling high-quality, balanced datasets in free-living conditions for PD, many ML classifiers have focused on distinguishing PD patients from healthy individuals rather than on classifying PD severity grades, such as mild, moderate, or severe. Developing a unified analytical framework to assess wearable intelligence for severity classification will be crucial for enhancing PD patient self-management.

This study addresses these gaps by introducing PDWearML, a unified analytic framework to optimise wearable ML for fast PD severity assessment using minimal data and simplified interactions. It encompasses standardised annotation criteria linking UPDRS-inspired daily activities to H&Y severity grading [22], [23]; multi-scale feature importance analysis to pinpoint clinically relevant signals; representative activity selection for efficient proxies; and a comprehensive evaluation across 12 SOTA ML models to ensure robustness in supervised daily activities settings.

To validate PDWearML, we conducted a 12-month recruitment study, yielding a supervised wearable dataset of daily activities from 100 PD patients and 35 age-matched healthy controls. These participants are primarily older adults from minoritised groups and rural areas of Yunnan Province, China. We used Huawei smartwatches and Shimmer sensors. The PDWearML dataset and code are publicly available at Code Ocean and github. This availability empowers further research on health inequities via accessible digital tools.

The remainder of this paper is organised as follows. Section II details the PDWearML methodology, including data collection and annotation criteria, multi-scale feature extraction and selection, representative activity identification, and severity assessment procedures. Details of the experimental results are reported and analysed in Section III. In Section IV, we discuss our findings and conclude in Section VI.

II. Methods

A. Study design

As shown in Fig. 1, the PDWearML framework, from activity collection to PD severity assessment, is divided into four steps: Step 1: Data collection and annotation. Four task types—clinical, functional, gross motor, and fine motor—comprising 16 activities were collected using Shimmer and Huawei GT3 devices. At the same time, physicians scored participants using the H&Y scale from 0 to 4. Step 2: Feature extraction and selection. For each activity, features were extracted from the accelerometer X, Y, and Z signals using the sliding window method. A subset of key features was then selected based on their importance for subsequent activity correlation analysis and modelling. Step 3: Representative activities selection. First, we identified representative single activities. Then, we used a combination of strategies to achieve a more comprehensive evaluation. Step 4: Severity assessment. We employed five categories of machine learning models to evaluate the effectiveness of using representative activities for PD severity assessment.

Our PDWearML framework adopts a feature engineering approach combined with traditional machine learning models. This strategy prioritises clinical practicality, data efficiency, and model interpretability over end-to-end deep learning paradigms. The PDWearML dataset includes 135 participants. This scale represents a typical small-sample medical scenario. Deep learning often struggles here due to its need for extensive annotated data. Feature engineering counters this limitation. It pre-extracts multi-domain signals such as time-domain means, frequency-domain spectral energy, and autocorrelation peaks. It also incorporates domain knowledge from UPDRS and H&Y scales. Traditional models like random forests or LightGBM thus converge effectively on limited data. The method follows hybrid intelligence principles. It merges human expertise in clinical activity selection with algorithmic automation. This reduces noise and boosts robustness. The approach sacrifices some automatic representation learning for strong interpretability and easy deployment. These traits support fast detection in free-living settings. They also enable clinicians to verify decisions quickly via tools like SHAP analysis.

B. Data collection and annotation

The process of data collection is presented in Fig. S2 a. Participants performed specific tasks categorised into clinical, functional, fine motor, gross motor, and fine motor groups, each lasting 20–60 seconds. The entire process was video-recorded. Fig. S2 b displays the sensor details used to collect the data. Two wearable sensor devices were used in our study. The first device, a professional-grade wearable known as Shimmer, was attached to both wrists and operated at a sampling rate of 200Hz. The second device, the commercially available Huawei GT3 watch, was worn on the wrist with more severe symptoms and operated at a sampling rate of 100Hz. Both devices were equipped with accelerometers and gyroscopes. Further details about the devices can be found in

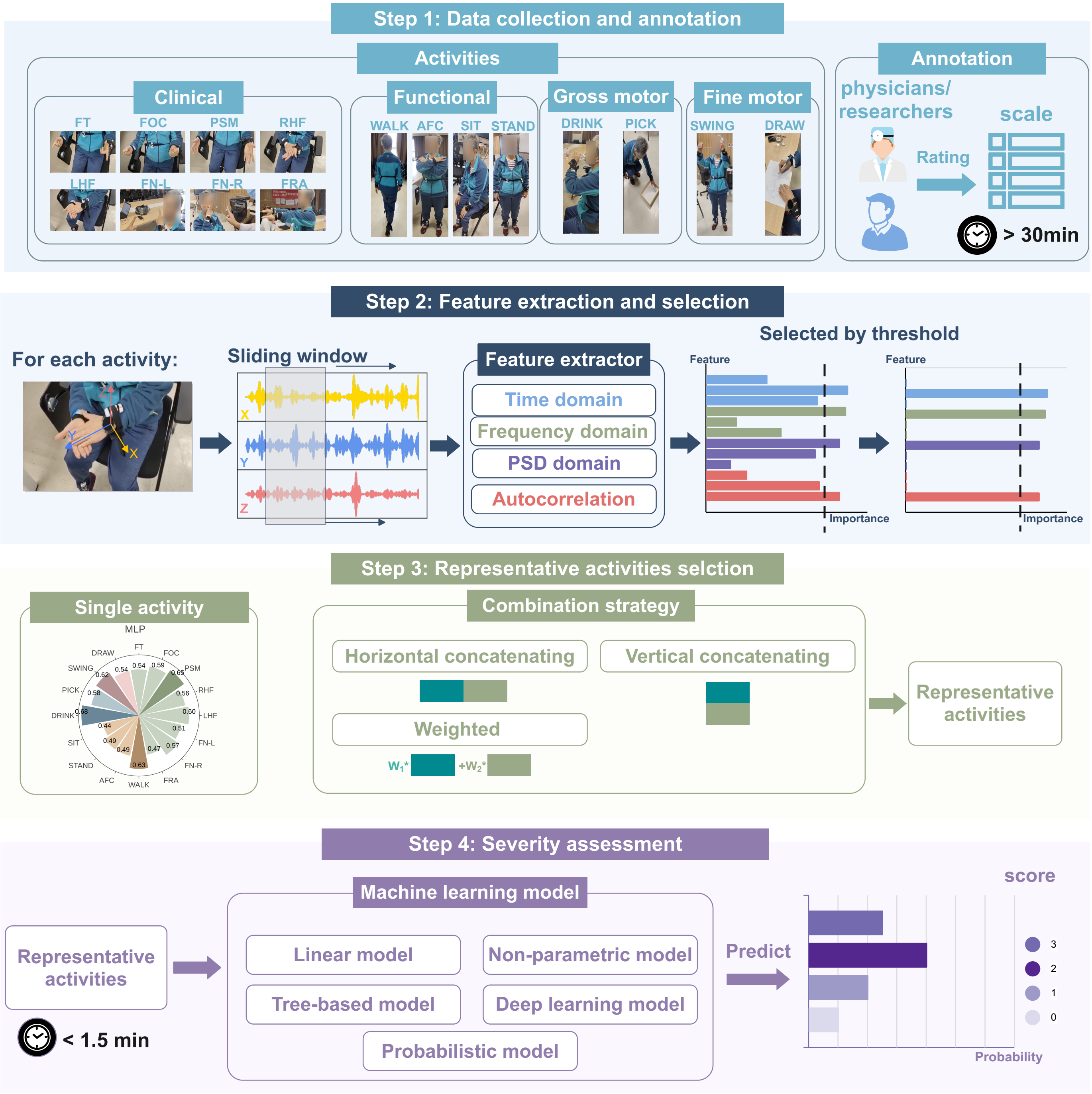

Fig. 1. The PDWearML framework for PD severity assessment. The process is divided into four steps: (1) Data collection and annotation: Four types of activities (clinical, functional, gross motor, and fine motor) are collected using wearable sensors, followed by annotation from physicians and researchers using standardised scales. (2) Feature extraction and selection: For each activity, features are extracted from accelerometer data in multiple domains (time, frequency, PSD, and autocorrelation) using a sliding window approach, with key features selected based on importance thresholds. (3) Representative activities selection: A combination strategy (horizontal, vertical, or weighted concatenation) is used to identify representative activities. (4) Severity assessment: Machine learning models from various classes (linear, non-parametric, tree-based, deep learning, probabilistic) are applied to predict Parkinson's disease severity, with the final score indicating the likelihood of severity grades.

Table S3. Fig. S2 c statistics on the distribution of PD severity among subjects.

Participants. Strict screening criteria were applied to all participants. As shown in Fig. 2, a total of 321 participants (193 with PD and 128 healthy controls, HC) were involved in the study. Based on age-matched and sensor frequency consistency, 83 younger participants and 19 individuals with PD were excluded. Further inspection of the videos and data resulted in the exclusion of an additional 19 PD participants due to lost video recordings. Ultimately, 100 PD patients and 35 HC were selected for the final experimental dataset.

Activities. Referring to the UPDRS guidelines (Table I), we selected 16 easy-to-execute activities categorized into four types: clinical, functional, gross motor, and fine motor. The

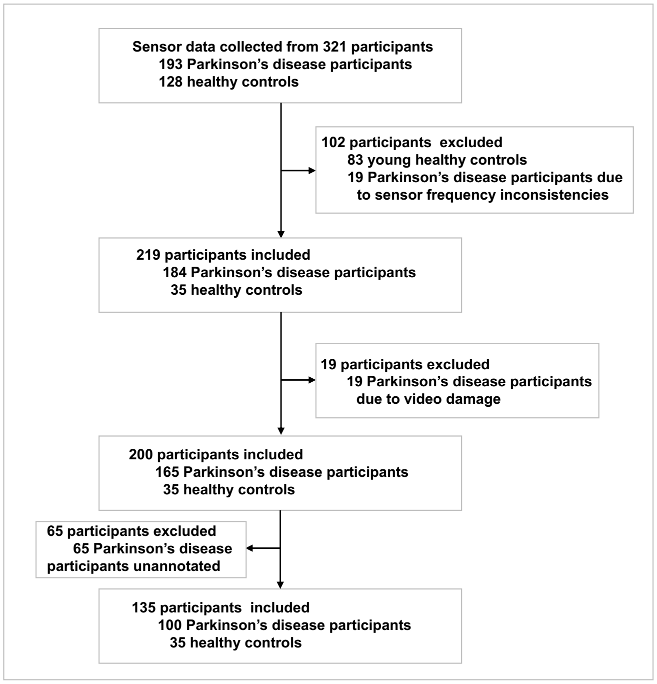
Sensor data collected from 321 participants
193 Parkinson's disease participants
128 healthy controls
102 participants excluded
83 young healthy controls
19 Parkinson's disease participants due to sensor frequency inconsistencies
219 participants included
184 Parkinson's disease participants
35 healthy controls
19 participants excluded
19 Parkinson's disease participants due to video damage
200 participants included
165 Parkinson's disease participants
35 healthy controls
65 participants excluded
65 Parkinson's disease participants unannotated
135 participants included
100 Parkinson's disease participants
35 healthy controls

Fig. 2. Flowchart of participant screening. A total of 321 participants were initially recruited, consisting of 193 Parkinson's disease participants and 128 healthy controls. After several exclusion steps due to factors such as young healthy controls, sensor frequency inconsistencies, video damage, and lack of annotation, 135 participants (100 Parkinson's disease participants and 35 healthy controls) were included in the final analysis.

duration of these activities is provided in Table S4. Most activities were performed in approximately 24 seconds, except for WALK.

TABLE I
ACTIVITIES PERFORMED BY PARTICIPANTS FOR THE PD ASSESSMENT

ID	Abbv.	Motor task	Type of task	UPDRS	Time(s)
1	FT	Finger tapping	Clinical	3.4	24.4
2	FOC	Fist open close	Clinical	3.5	24.1
3	PSM	Pro/Sup movements*	Clinical	3.6	24.3
4	RHF	Right hand flip	Clinical	3.6	34.3
5	LHF	Left hand flip	Clinical	3.6	24.7
6	FN-L	Finger to nose left	Clinical	3.16	24.3
7	FN-R	Finger to nose right	Clinical	3.16	24.0
8	FRA	Front raise arms	Clinical	3.17/18	25.1
9	WALK	Walking	Functional	3.10/11	47.6
10	AFC	Arising from chair	Functional	3.9	25.4
11	DRINK	Drinking	Gross motor	2.3	26.4
12	PICK	Pick up something	Gross motor	2.9/11	28.2
13	SIT	Sitting	Functional	3.17/18	24.1
14	STAND	Standing	Functional	3.17/18	24.2
15	SWING	Stand–Swing arms	Fine motor	3.14	24.1
16	DRAW	Drawing on paper	Fine motor	2.7	26.8

* Pro/Sup movements: Pronation supination movements

Annotation. Annotations employed two schemes: physicians used the H&Y scale for clinical assessments, whereas non-clinical researchers applied Table S2, which references the UPDRS-III scoring criteria to capture more granular activity details. The mapping between these annotations is provided in Table S6. The H&Y scale provides expert-driven precision but is difficult to scale in real-world settings owing to scarce clinical annotators. In contrast, the Table S2 scheme supports flexible simulations of home environments akin to self-assessments. Recent studies utilised signal expert annotation to alleviate data scarcity for 'in-the-wild' environments [24], [25]. However, it lacks physicians' expertise, which elevates task complexity. Given these trade-offs, we designate the Table S2 annotations as ground truth for training, validating, and testing our machine learning models. Accordingly, the class distribution of the labels was Normal (N=35), Mild (N=63), Moderate (N=21), and Severe (N=16). To address the potential class imbalance, we employed a random oversampling strategy. This researcher-annotated(RA) standard was demonstrated in our prior foundational work to be effective for identifying features correlated with disease severity [23]. Furthermore, we conducted a separate experiment to analyse the feasibility of whether representations learned using our practical RA standard (Table S2) effectively generalise to predict the scarce, clinical H&Y standard. We trained a machine learning model using RA as ground truth. We then assessed the clinical relevance of this model by evaluating its performance on the clinical H&Y standard.

Ethics. All ethical and experimental procedures adhered to the guidelines outlined in the Code of Ethics of the World Medical Association (Declaration of Helsinki). The experimental protocol and all procedures received full ethical approval from both the Ethics Committee of Yunnan University and the Ethics Committee of Yunnan First People's Hospital. All subjects provided informed consent and signed the consent form.

C. Feature extraction and selection

Fig. S4 presents the workflow of feature extraction and selection.

Data Preprocessing. As the collected data showed inconsistent initial values, we applied Z-score normalisation (transforming a data point by subtracting the mean of its feature and dividing the result by the standard deviation) for short-duration motor tasks to centre the signals around 0 [26]. Next, we applied band-pass filtering(filter between 0.3 Hz and 17 Hz) to isolate the most representative signal components [27]. The data were segmented using a 1.5-second sliding window with a 50% overlap to avoid splitting activity cycles across windows [28].

Feature Extraction. To minimise device and computational demands, feature extraction focused exclusively on accelerometer-derived signals. Multi-scale and multi-level feature selection methods were adopted to ensure the full scope of the assessment [23]. Sample-level features were extracted from the full 20-50 second signals, capturing mean and variance trends along with windowed mean/variance differences (Table S5). However, calculating tremor displacement based on a single axis is often affected by motion drift or gravity. To mitigate this, spatial fusion is applied by combining data from different orientations of the accelerometer. Specifically, A signal magnitude vector (SMV), called as A-axis, is a calculated value representing the overall magnitude or length of a multi-axis signal. A-axis is calculated as the square root of the sum of squares of the three raw axes, providing a more robust measure. In addition, axial correlations between various pairs of axes(XY, YZ, XA, YA, and ZA) were computed for each activity. At the segment level, selecting an optimal activity window size is crucial, as it must capture at least one full activity cycle. Thus, the window size depends on the fundamental time period of each activity. At this stage, raw time-series signals are transformed into autocorrelation signals, frequency-domain signals, and power spectral density (PSD) signals. Additionally, time-domain features are commonly extracted, including statistical metrics such as mean, maximum, minimum, standard deviation (Std), root mean square (Rms), peak-to-peak amplitude (Ptp), zero-crossing rate (Czr), log-energy, percentiles, and interquartile range (Interq). In contrast, frequency-domain features, which represent signal periodicity, are obtained by applying the Fast Fourier Transform (FFT) to the raw data. These features include kurtosis, skewness, dominant frequency (Domifq), spectral energy (SpecEgy), spectral entropy (SpecEnt), and mode. Finally, we extracted 220 features across all activities (Table S5).

Feature Selection. To identify key feature attributes that distinguish disease severity, we employed six complementary feature selection methods: LightGBM Feature Importance Scores, XGBoost Feature Importance Scores, Random Forest Feature Importance Scores, Permutation Importance Scores (with LightGBM as the estimator), RFECV Rankings (using LightGBM), and Boruta Rankings (with Random Forest as the estimator). We achieved stable feature selection by calculating an average rank for each feature across all six methods, an approach that significantly reduced redundant features and improved classification accuracy. Each method offers unique contributions. For instance, LightGBM and XGBoost are fast and effective at identifying high-impact features but may overlook features of medium importance. In contrast, Random Forest excels at handling complex interactions, although it can be more sensitive to noise. Permutation Importance provides unbiased feature relevance estimates but is computationally intensive. RFECV systematically eliminates weaker features by evaluating model performance with feature subsets, though it may struggle with larger datasets. Lastly, Boruta excels at identifying all relevant features, ensuring no critical variables are missed, but it can be time-consuming due to its thoroughness. Finally, by integrating these methods, we leverage their complementary strengths, selecting only the most important features, minimising redundancy, and enhancing model accuracy.

D. Representative activities selection

We employed a process to identify representative activities and their combinations (Fig. S5). For the identification of a single representative activity, 16 activities (Fig. S5 a) were input into machine learning models (Fig. S5 e), and their performance on severity assessments was evaluated (Fig. S5 c). For identifying representative activity combinations, we first applied a branch-and-bound method (Fig. S5 d) to search for optimal combinations. Then, three data combination strategies (Fig. S5 b) were employed for combining activities. The combined data were subsequently fed into the machine learning models and their performance was evaluated (Fig. S5 e, c). The three combination strategies were based on heuristics(supplementary material: combination strategy). The horizontal and vertical strategies expanded the feature space of the dataset and increased the number of samples, respectively, while the weighted combination method assigned weights to activities based on their F1 scores.

E. Severity assessment

In this study, we evaluated five categories of models across twelve different methods. Each method was described in the supporting material: machine learning algorithm. A 5-fold cross-validation process was used to ensure the reliability of the model assessments. The linear models included Logistic Regression with L2 and L1 penalties, as well as Support Vector Machines (SVM) with L2 and L1 penalties. Non-parametric methods were represented by K-nearest Neighbors (KNN), while probabilistic methods included the Naive Bayes classifier. Tree-based methods consisted of Random Forest, XGBoost, and LightGBM. Deep neural networks were represented by 2-layer, 4-layer, and 8-layer neural networks. Each model was assessed using seven key metrics: Accuracy, Precision, Specificity, F1 score, Recall, AUC (Area Under the Curve), and ROC (Receiver Operating Characteristic) curves(the metric formula is detailed in supplementary material: evaluation metric). These metrics provided a comprehensive evaluation of model performance, capturing overall classification ability, precision, handling of imbalanced classes, and the model's sensitivity and specificity in detecting true positive and true negative cases.

III. Results

A. Data quality and distribution

A cross-sectional study recruited 321 participants over 12 months. After excluding subjects failing to match the baseline criteria (Methods II-B participants), 100 patients with varying disease severity (mild, moderate and severe) and 35 age-matched healthy controls were included. Demographics and clinical characteristics are presented in Table S1. Supplementary tables and figures are denoted with an 'S' prefix (e.g., Table S1). No significant demographic differences were found between the two cohorts. The affected side (the hand with the dominant symptoms) was determined by patient self-assessment and was evenly distributed across the cohort. Sensor signal data were collected for 16 activities based on the UPDRS scale, with two sets of criteria used for annotation. To validate our annotation, we conducted an independent experiment to assess the clinical relevance of our model, which was trained exclusively on the Researcher's Annotation (RA) standard. We evaluated this single model against the independent test set, using the physician-provided H&Y scale as the external clinical benchmark. This experiment yielded two critical, contrasting insights. The result was shown in Fig. S1. First, on key functional tasks (e.g., WALK, AFC, DRINK), the model demonstrated exceptionally strong performance. This finding is crucial, as it suggests that our RA standard (Table S2) provides a cleaner and more optimised supervisory signal for these high-noise activities, enabling the model to learn generalisable feature representations that effectively map to the true clinical (H&Y) state. Conversely, on tasks that were more holistic or purely clinical (the source of the physicians' excellent... observation), the model's performance declined. This indicates that the feature representations learned from our functional proxy are task-specific and do not generalise as effectively to these non-functional domains, thus defining the clear boundaries of our current annotation strategy.

B. Feature importance analyses

We analysed the feature importance in terms of the number of features, the ranking of important features and the most important features, respectively. Due to space constraints, we present only two activities (PSW, DRINK) in Fig. 3. These activities also serve as representative examples in the subsequent Fig. 4, facilitating a consistent analysis. First, we observed the effect of the number of features on the performance of the model. The performance of the model reaches a milestone quickly after the number of important features has accumulated to 20. After this point, the F1 score trend stabilised and peaked after adding up to 20 features (Fig.3a), indicating that a small feature set largely drives severity assessment. In both PSM and DRINK activities, second, the feature "autocorrelation coefficient of y-axis"(fea_autoy) consistently ranked as the most important feature, with the highest mean SHAP value in both cases (Fig.3b,c). This suggests that the feature domain of autocorrelation plays a critical role in both types of activity. The difference in feature importance rankings between the two activities implies that certain features are more suited for specific activities. For example, in the PSM activity, axis-based correlation features such as "t_zaCor" and "t_xzCor" are important, while in the DRINK activity, peak-related features(peaks_abnormal, peaks_normal) and energy-related features("p_lgEnergy_y", "p_lgEnergy_x") are more dominant (Fig.3b, c). Further, the panel feature importance offered deeper insight into the variability of each feature's contribution to individual predictions. For "fea_autoy" in PSM, the SHAP values are consistently positive and relatively high for most predictions, indicating that higher values of this feature generally push the model towards higher predictions. In contrast, some other features (e.g., "f_skew_a") show a broader range of SHAP values, indicating a more variable influence on the model's predictions(Fig.3d). Finally, the panel of most important feature provides detailed insights into the behaviour and importance of the "fea_autoy" feature across different severity grades of PD for two representative activities: PSM and DRINK. The "fea_autoy" values were significantly different across all severity grades, showing a noticeable downward trend as severity increases(Fig. 3f, k). This suggests that "fea_autoy" plays an important role in distinguishing different severity grades. There was a strong positive correlation (R = 0.74) between "fea_autoy" values and SHAP values, indicating that higher "fea_autoy" values positively influence the prediction of the normal class(Fig.3g). As the severity increases, the importance of "fea_autoy" became more negative, indicating that higher values of this feature decrease the likelihood of more severe classifications, particularly in the moderate and severe class(Fig.3h, i and j).

Following this analysis, we identified the top 20 features per activity. This selected feature set was then used for the representative activity and comprehensive severity assessments in subsequent sections.

C. Representative activities

A representative activity is defined as one that consistently outperforms other activities of the same category across most models. Pronation supination movements (PSM) was the most representative activity in the clinical type, and it outperformed other clinical-type activities by 4-20% on the logistic regression model(Fig.4a: Logistic regression). PSM also achieved F1 scores of 60%, 65%, 54%, 64%, and 52% on other algorithms (SVM, MLP, LightGBM, and KNN), positioning it as the optimal clinical-type activity (Fig.4a). WALK was the most representative functional activity, achieving optimal F1 scores of 57%, 55%, 63%, 46%, 58% and 55% across six models. For gross motor activities, DRINK outperformed PICK, achieving optimal F1 scores of 61%, 61%, 63%, 57%, 67% and 59% on six models. DRINK had a better F1 score than PSM, implying that the short-term activities of the daily living type have comparable severity assessment performance as the clinic-type activities(Fig.4a). Overall, the representative activities(with ID in Table.I) of clinical, functional, gross and fine are PSM(3), WALK(9), DRINK(11) and SWING(15) respectively. They are dynamic activities and cover the detection of symptoms of kinetic tremor, bradykinesia and gait(Table.I).

Each UPDRS activity typically aims to assess a specific function. Thus, we aimed to identify the minimal set of activities that most effectively enhances PD severity assessment. Combining activities (AFC, DRINK) could provide a better performance, delivering 77% of the F1-score(Fig.4e). Combining activities (WALK, AFC, DRINK) gained 81.4% F1 score as the optimal three-activity combination, significantly outperforming the most representative single-activity(DRINK, 9) by 18%(Fig.4f). Meanwhile, combining activities (FN-R, WALK, AFC, DRINK) presented an 81.8% F1 score as the optimal four-activity combination(Fig.4g). In summary, when combining activities to assess severity, the best activity groups are FN-R, WALK, AFC, and DRINK. These cover clinical, functional, and gross activity tasks and can typically be done in less than 2 minutes.

On the other hand, three heuristics including horizontal combination, vertical combination and weighted combination were utilised to assess the performance of the activity combinations. The horizontal combination was observed to have overall better performance than the vertical and weighted combination(Fig.4b, c and d). This reveals that increasing the dimensions of the feature space helps the model to classify PD severity. Therefore, the horizontal combination of activity data should be prioritised as the preferred strategy.

D. Comprehensive severity assessment

We investigated 12 typical ML models. To thoroughly estimate the reliability of the model, five metrics—accuracy,

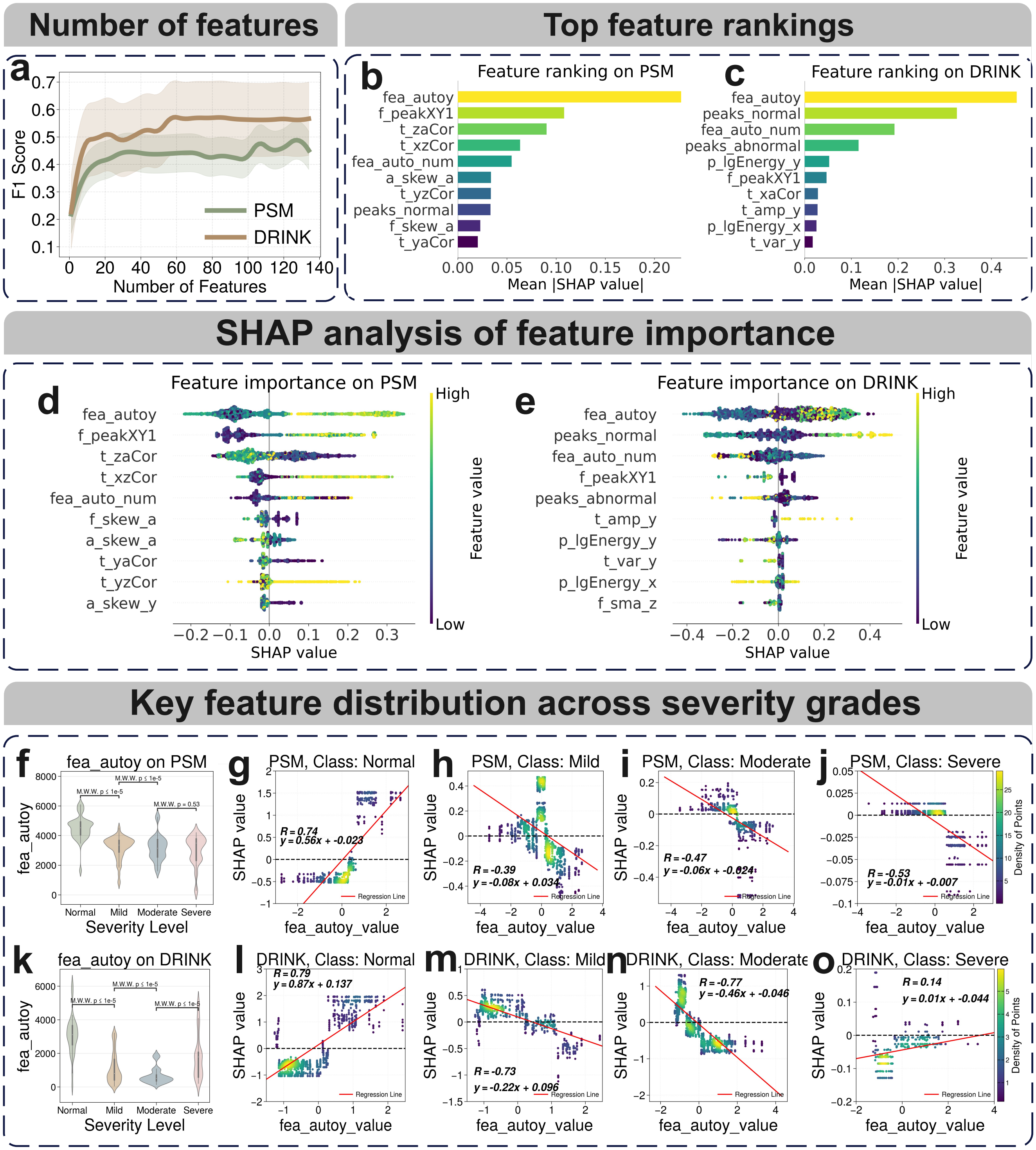

Fig. 3. Feature importance analysis. **a.** Trend of F1 score with the number of features. The increasing number is in descending order of feature importance. The solid lines represent the mean F1 score and the shaded regions indicate the standard deviation. **(b-c).** Top feature rankings in selected activities. **(d-e).** The SHAP importance of features in representative activities. Large SHAP values indicate that the feature is important to the severity assessment. **(f-o).** Violin plots display the distribution of "fea_autoy" values for different PD severity grades (Normal, Mild, Moderate, Severe). Scatter plots show the relationship between "fea_autoy" values and their corresponding SHAP values for different PD severity grades. The Mann-Whitney-Wilcoxon test (M.W.W.) is used to determine if there is a significant difference in the distribution of the two feature sets.

precision, specificity, F1 score, and recall—were used to evaluate its classification performance. Owing to space limitations, results are presented for only the most representative activities and their combinations. The deep learning model was found to yield the best overall performance. Subsequently, a separate, fine-grained analysis of the classification results

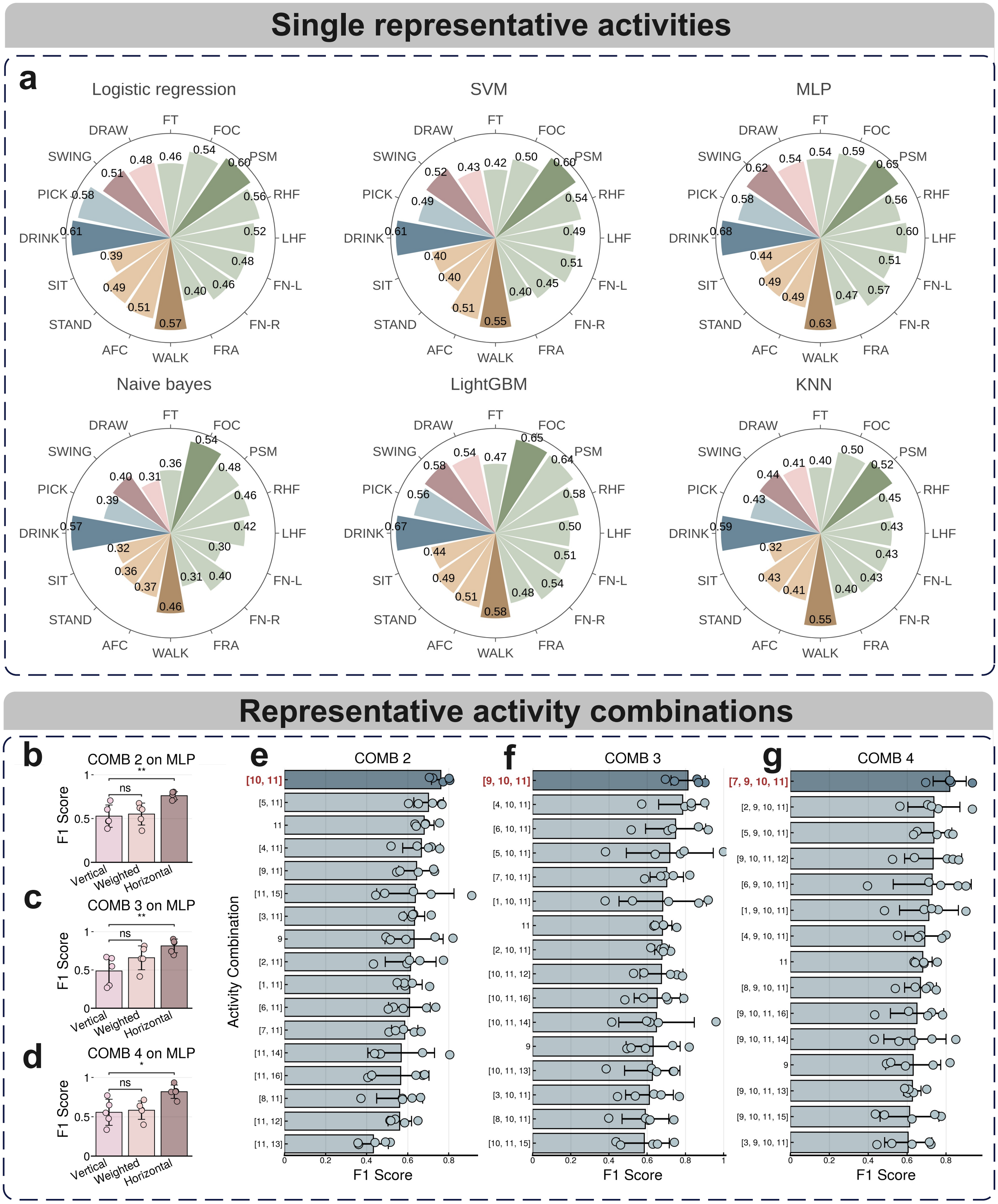

Fig. 4. Representative activities and their combinations. **a.** The severity assessment of all 16 activities was evaluated by their F1-scores, using six classifiers (Logistic Regression, SVM, MLP, Naive Bayes, LightGBM, and KNN). Different colours denote activity types: clinical, functional, gross motor, and fine motor. Darker colours highlight the activity with the highest assessment performance within each task category. **b-d.** Representative activity combination. COMB * indicates the number of combinations of activities; the darker colours represent the optimal activity combinations. **e-g.**Three activity combination strategies were evaluated: horizontal, vertical, and weighted. The horizontal approach proved significantly better than the others. Significance levels are denoted as **ns**(not significant) for $P > 0.05$; *$P \leq 0.05$, **$P \leq 0.01$

for these activities was conducted in the context of identity learning. For the most representative activity, DRINK, the MLP network with 8 hidden layers was identified as the optimal model with 75.3% accuracy, 73.6% precision, 94%

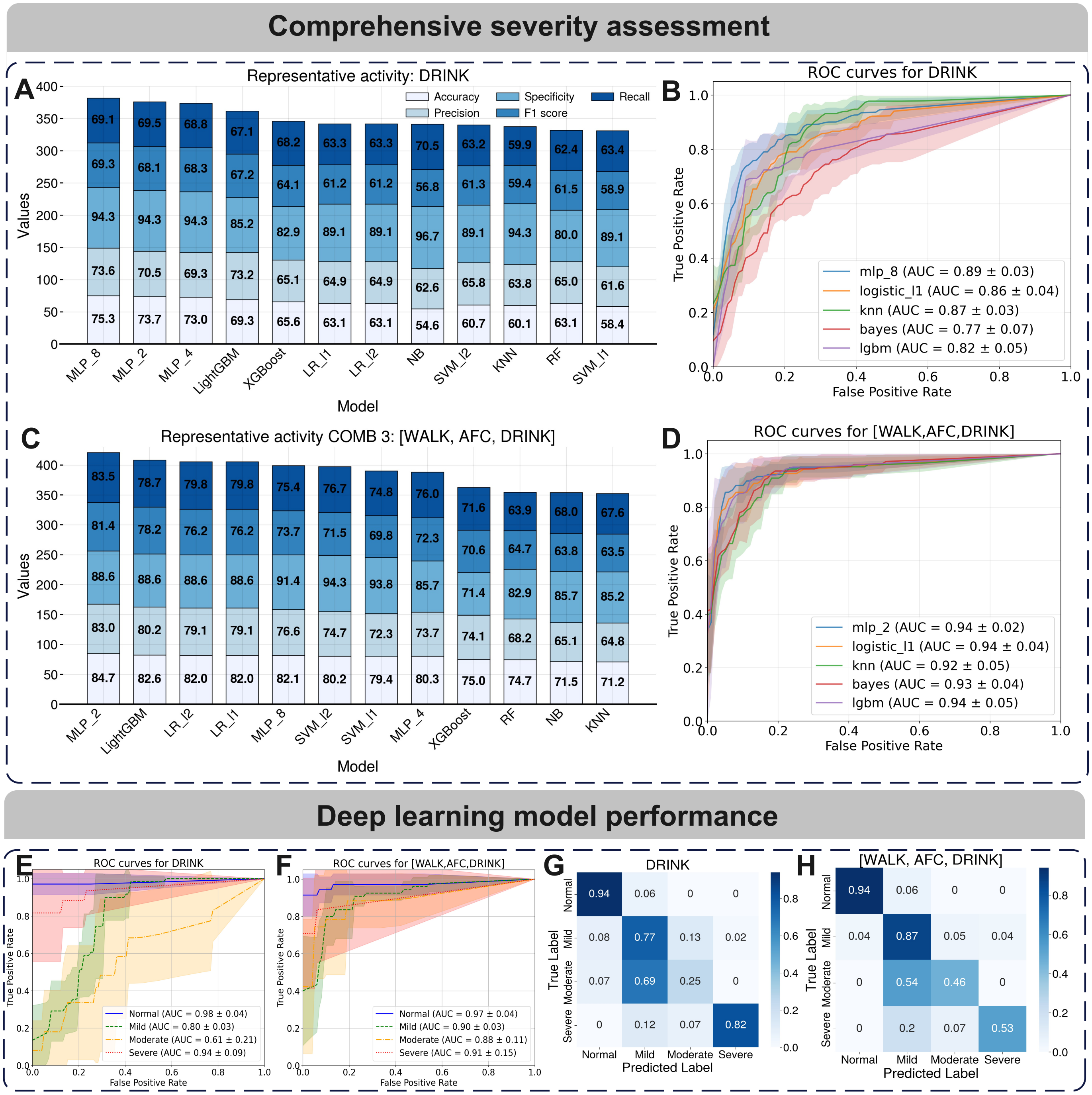

Fig. 5. The performance of various machine learning models in predicting PD severity based on representative activities(DRINK, and COMB 3). **a, c.** The accuracy, precision, specificity, recall, and F1 score for twelve ML models. Each bar's total height reflects overall performance (higher is better), while colored segments indicate contributions from each metric. **b, d.** ROC curves for five types of ML models, with the area under the curve(AUC) showing model performance in different prediction tasks. The "deep learning model performance" section highlights the best-performing models(MLP), with **e, f** showing ROC curves for different severity grades(Normal, Mild, Moderate, Severe), and **g, h** depicting the confusion matrices, indicating how well the model differentiates between these severity grades.

specificity, 69.3% F1 score and 69.1% recall respectively(Fig. 5a). An accuracy of 75.3% reflects the overall correctness of the model's predictions and serves as a baseline for comparison with other metrics. Compared to a precision of 73.6%, the higher specificity (94.3%) indicates the model's better ability to correctly identify non-PD cases, which is crucial in scenarios like disease screening that require a reduction in the false positive rate. This trend was consistent across other models as well. Deep learning models (MLP_2, MLP_4, MLP_8) demonstrated the best overall performance, followed by gradient boosting tree models (LightGBM and XGBoost). The model's superior specificity (94.3%) and accuracy (≥73%) primarily explain why MLP outperformed others. Deep learning models were identified as optimal in the activity combination. The combination of three activities—WALK, Arising from Chair, and DRINK—achieved 84.7% accuracy, 83% precision, 88.6% specificity, 81.4% F1 score, and 83.5% recall with this model (Fig. 5c,d). The deep learning model performed well in predicting both normal and severe PD cases, with AUC values exceeding 0.90 for both activity sets (Fig. 5e, g). Including multiple activities improved classification accuracy for moderate cases, which are typically more challenging to classify and were most frequently confused with the 'Mild' class, highlighting the difficulty in distinguishing these adjacent stages. (Fig. 5f, h). However, further refinement may be needed to improve the model's predictions for moderate cases.

E. Justification on Huawei GT3 Watch

The pipeline for the smartwatch is identical to that of Shimmer. This includes data collection protocols, feature extraction and selection (Top 20 features), representative activities and their combinations, model training, and evaluation procedures, ensuring a direct and fair comparison. In feature importance analysis, "fea_autoy" was identified as the most important feature (Fig.6b). Higher values of "fea_autoy" to FT were associated with more severe grades of PD severity (Fig.6a,c). Consistent with findings from shimmer sensors, autocorrelation features play a key role in severity assessment. Among the representative activities, clinical-type activities outperformed other activity types, suggesting that clinical activities remain a crucial foundation for self-assessment (Fig.6d,e, and f). Notably, unlike PSM, which was the best clinical classification on the Shimmer device, FT performed best on the watch. This difference may be attributed to the varying abilities of the devices to capture subtle disease signals. In severity assessment, deep learning models outperformed other types of machine learning models in overall performance, while the LightGBM model demonstrated strengths in precision, F1 score, and recall (Fig.6g, h). These findings underscore the complexity of consumer-grade sensors, and PDWearML can effectively assist researchers in selecting appropriate models for severity assessment in various scenarios.

IV. Discussion

This work presents PDWearML, a unified framework for assessing PD severity, and evaluates its effectiveness across both Huawei and Shimmer sensors. By identifying key digital features and representative activities through comprehensive machine learning evaluation, PDWearML efficiently classifies motor grades, leading to significant improvements in the speed and accuracy of PD severity assessments. Our key findings are: (1). Autocorrelation is a critical feature domain for recognising severity. (2). PD severity assessment using the daily living activity (DRINK) shows performance comparable to clinical-type activity (PSM). (3). Deep learning models outperform other ML models, with the F1 score being a better evaluation metric. (4). The proposed framework is equally effective on the Huawei GT3 watch for analysing key features, representative activities, and comprehensive severity assessments.

Although numerous features have been proposed for PD recognition, identifying activity-specific keys in free-living settings remains challenging. Recent studies indicate that axis-related and autocorrelation features contribute valuable information for activity recognition and quantifying motor symptoms. [29], [30]. Our study similarly found that features reflecting overall autocorrelation, such as "fea_autoy" and "fea_auto_num", as well as axis-related features like "xzCor" and "yzCor" were effective in assessing severity (Fig.3b,c and Fig.6b,c). These results were consistent even with data collected from a smartwatch, demonstrating the stability and reliability of these features.

We have the following analyses of representative activities. First, the clinical type of activity across all algorithms was significantly higher than that of the others. This suggests that clinical activities are reliable for assessing disease severity. Second, the activity combination provides a more comprehensive symptom view of estimating severity. In two-activity combinations, DRINK combined with AFC or LHF improved classification performance compared to using DRINK alone. In three-activity combinations, DRINK and AFC combined with FT, FN-L, LHF, RHF, or WALK improved classification performance compared to using DRINK alone. Therefore, the introduction of clinical-type activities significantly improved performance, suggesting their ability to classify fine-grained severity grades. However, we observed the slow performance increase of this heuristic combination strategy, which inspired the future proposed use of more effective data fusion strategies: e.g. multi-task learning and multi-modal fusion. Generally, classifying severity is a more complex problem than PD detection. Lonini et al. found that the detection of bradykinesia and tremors from fine motor tasks or WALK has comparable accuracy to the PD detection utilising clinical type [31]. However, in the present study, fine motor tasks are difficult to have the same ability to tackle fine-grained classification problems as clinical activities. We reserve this discussion for 'representative activities' and 'severity assessment' in the Supplementary Information.

It is important to contextualise our study's protocol. The literature defines "free-living" monitoring as the capture of spontaneous, non-structured activities over extended periods in a patient's natural, unconstrained environment, without external prompts or clinical supervision [32], [33]. This stands in contrast to traditional, rigid laboratory assessments, which often involve highly instructed tasks (e.g., specific gait protocols) using specialised multi-sensor equipment. Our protocol was designed as a translational bridge between these two paradigms. While conducted in a supervised setting, our "daily activities" design moves beyond rigid protocols by minimising participant burden. We focused on capturing short, minimally guided activities (e.g., DRINK, ARISE-FROM-CHAIR) using only a single wrist-worn device. This design, validated on consumer-grade smartwatches, facilitates fast assessment (¡2 min) and prioritises the real-world feasibility and patient compliance necessary to pave the way for effective at-home deployment.

Our PDWearML framework employs feature engineering combined with traditional machine learning, rather than end-

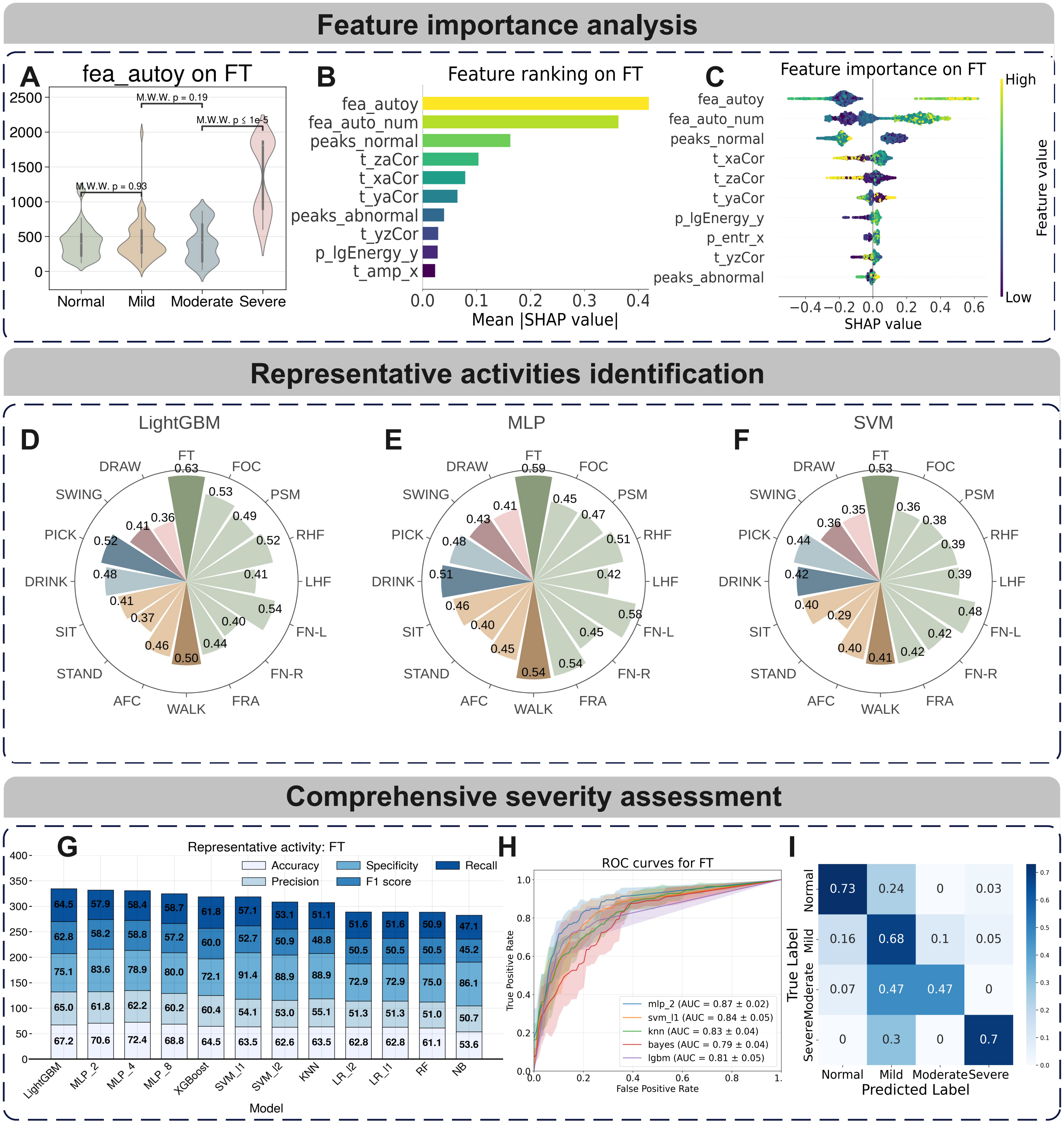

Fig. 6. Feature, activity and assessment analysis of PDWearML on Huawei GT3. **a-c** Important features analysis in FT activities. **d-f** Representative activities on three types of ML models. **g** presented the performance of FT for all ML models using five metrics, while the confusion matrix of the optimal ML model on FT is shown in **h**. **i** ROC curves and the corresponding AUCs for five categories of machine learning models.

to-end deep learning, to balance interpretability, data efficiency, and deployment feasibility in resource-constrained PD wearables. This choice stems from key considerations in PD wearable studies, including small-sample challenges and clinical needs for transparency [26]. Interpretability stands out as a core advantage. Feature engineering enables SHAP analysis to highlight clinically relevant signals. For instance, it reveals tremor peaks linked to H&Y stages. This aids physician validation. In contrast, end-to-end DL produces opaque decisions. These limit traceability and erode trust [34]. Data efficiency further justifies the method. Embedding UPDRS domain knowledge fits our n=135 cohort well. It curbs overfitting effectively. DL, however, demands massive annotated data. This leads to poor generalisation in the wild. Robustness completes the rationale. Hybrid intelligence blends expert activity curation with automation. It dampens noise reliably, as our consistent F1 scores demonstrate. DL automates features but struggles with sparse datasets [35]. Overall, these trade-offs promote transparent PD interventions. This study does not discourage deep learning; on the contrary, it encourages it. Systematic experiments reveal that the MLP model delivers the strongest overall performance. This underscores the nonlinear mapping between signal features and disease classification. Thus, future work should prioritise establishing the importance of key wearable signals for disease grading in free-living settings. Hybrid feature-DL models hold promise for these advancements.

A key challenge in extending our observed clinic-based assessments to in-the-wild home environments lies in bridging controlled protocols with unstructured daily life. Our results on the Huawei Watch demonstrate the feasibility of consumer-grade devices for classifying PD severity during short-term activities. Hence, the core obstacle in home monitoring is not sensor precision, which our Huawei experiments partially affirm, but data context deficiency: distinguishing meaningful segments, such as walking or drinking, from 24/7 noise streams [36]. Our study simplifies this by identifying WALK, AFC, and DRINK as efficient proxy activities that capture PD severity in under 2 minutes [37]. This enables a two-stage deployment on devices like the Huawei Watch. Stage 1 deploys a lightweight, always-on activity recogniser to capture these proxies from continuous data. Stage 2 then activates a variety of ML models for fast, precise severity scoring on isolated segments [38]. Thus, it leverages scalable consumer hardware while defining precise measurement targets [24].

Our study underscores a pivotal paradigm shift from sporadic clinical snapshots to remote, high-frequency clinical spot-checks. Unlike continuous passive monitoring, which often struggles with context identification, annotation and noise, We validate the utility of short-term, specific daily activities (e.g., drinking, walking) as standardised tasks. This approach shows the feasibility of effectively migrating the rigour of standardised clinical tasks into the home environment [39]. Critically, this strategy of activity combination addresses the inherent limitations of widely used scales like the UPDRS, specifically inter-rater variability and subjective bias [40]. By integrating complementary motor tasks (COMB 3: WALK, AFC, DRINK), our models achieved clinical-grade precision in severity grading, with MLP demonstrating robust performance (AUC $\approx$ 0.94). This confirms that combining brief, targeted motor protocols significantly enhances the objective quantification of disease severity compared to single-task assessments or subjective ratings. The most significant translational implication lies in the feasibility of fast severity assessment. Existing research highlights the difficulty of capturing motor fluctuations and ”On-Off” phenomena using infrequent clinical visits [41]. We demonstrate that precise pathological features can be decoded within short time windows, validating the capability for high-frequency evaluations. This allows for the granular mapping of intraday symptom curves, providing a vital tool for detecting subtle motor fluctuations that traditional scales often miss. Consequently, this paradigm offers a scalable digital health solution for optimising personalised medication regimes.

V. Limitations

While PDWearML demonstrates promising potential, several limitations warrant objective discussion. First, regarding data annotation, the reliance on video-based labelling by raters, as opposed to face-to-face interactions, presents a constraint. This approach risks overlooking subtle clinical signs (e.g., rigidity tone) and non-verbal cues essential for precise severity evaluation, potentially introducing labelling noise. Second, we observed inconsistent feature performance across hardware platforms. As shown in our comparative analysis, the ”representative activities” identified differed between devices: research-grade sensors (Shimmer) prioritised gross motor tasks (e.g., walking, drinking), whereas the consumer smartwatch (Huawei GT3) showed higher sensitivity to distal movements (e.g., finger tapping). This discrepancy indicates that feature efficacy is currently device-dependent, and the framework's cross-device robustness requires further optimisation. Third, the management of noise in free-living environments remains a challenge. Extraneous factors stemming from varying sensor placement, diverse activity contexts, and participant non-compliance can introduce significant signal artifacts. The current absence of robust, automated noise detection and correction mechanisms may undermine assessment reliability in completely uncontrolled settings. Finally, the study population limited the model's generalizability regarding PD subtypes. While we covered various severity stages, the dataset may not fully represent the distinct kinematic signatures of tremor-dominant versus akinetic-rigid subtypes. Future work will focus on multi-centre data collection to enhance population diversity and develop unified feature frameworks that generalise across heterogeneous devices and environments.

VI. Conclusion

In this work, we introduce PDWearML, an accessible machine learning framework designed to support real-world evidence studies for fast and accurate assessment of PD severity in supervised conditions. To minimise participant burden, we designed a brief and intuitive data collection protocol based on simple daily activities and simplified device interactions, reducing fatigue for older adults with limited digital literacy and enhancing long-term adherence. We validated PDWearML by comprising 100 PD patients and 35 age-matched controls, demonstrating that optimised multi-scale feature extraction (notably autocorrelation) and combining representative activities (WALK, Arising-from-Chair, and DRINK) enable fast severity assessment in under two minutes with accuracy up to 84.7%. By publicly releasing the dataset and code, we ensure reproducibility, provide baseline results for future comparisons, and encourage further development of wearable-based PD assessment tools. Future work should address noise detection and processing in free-living data, explore richer data fusion strategies, and refine annotation processes to capture subtle clinical cues.

Acknowledgment

This work was supported by the China ScholarshipCouncil (No.202107030007
and the Engineering and PhysicalSciences Research Council (EPSRC) Doctoral Training Part-nership (EP/T517835/1). We also acknowledged our key col-laborators, including Yunnan University, Yunnan First People'sHospital, AntData Company, and the University of Sheffield.

References

[1] E. Dorsey *et al.*, "The emerging evidence of the parkinson pandemic," *Journal of Parkinson's disease*, vol. 8, no. s1, pp. S3–S8, 2018.

[2] E. Leroy *et al.*, "Deletions in the parkin gene and genetic heterogeneity in a greek family with early onset parkinson's disease," *Human genetics*, vol. 103, pp. 424–427, 1998.

[3] X.-a. Bi *et al.*, "A novel cernne approach for predicting parkinson's disease-associated genes and brain regions based on multimodal imaging genetics data," *Medical Image Analysis*, vol. 67, p. 101830, 2021.

[4] R. Powers *et al.*, "Smartwatch inertial sensors continuously monitor real-world motor fluctuations in parkinson's disease," *Science translational medicine*, vol. 13, no. 579, p. eabd7865, 2021.

[5] R. I. Griffiths *et al.*, "Automated assessment of bradykinesia and dyskinesia in parkinson's disease," *Journal of Parkinson's disease*, vol. 2, no. 1, pp. 47–55, 2012.

[6] C. Moreau *et al.*, "Overview on wearable sensors for the management of parkinson's disease," *npj Parkinson's Disease*, vol. 9, no. 1, p. 153, 2023.

[7] A. S. Chandrabhatla *et al.*, "Co-evolution of machine learning and digital technologies to improve monitoring of parkinson's disease motor symptoms," *NPJ digital medicine*, vol. 5, no. 1, p. 32, 2022.

[8] C. Sotirakis *et al.*, "Identification of motor progression in parkinson's disease using wearable sensors and machine learning," *npj Parkinson's Disease*, vol. 9, no. 1, p. 142, 2023.

[9] A. Intelligence *et al.*, "based software as a medical device (samd) action plan," *Food and Drug Administration*, pp. 2021–06, 2021.

[10] A. S. Gala *et al.*, "The digital signature of emergent tremor in parkinson's disease," *npj Parkinson's Disease*, vol. 10, no. 1, p. 147, 2024.

[11] A. Papadopoulos *et al.*, "Detecting parkinsonian tremor from imu data collected in-the-wild using deep multiple-instance learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2559–2569, 2019.

[12] N. Mahadevan *et al.*, "Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device," *NPJ digital medicine*, vol. 3, no. 1, p. 5, 2020.

[13] M. D. Czech *et al.*, "Improved measurement of disease progression in people living with early parkinson's disease using digital health technologies," *Communications Medicine*, vol. 4, no. 1, p. 49, 2024.

[14] M. K. Erb *et al.*, "mhealth and wearable technology should replace motor diaries to track motor fluctuations in parkinson's disease," *NPJ digital medicine*, vol. 3, no. 1, p. 6, 2020.

[15] O. Y. Chén *et al.*, "Building a machine-learning framework to remotely assess parkinson's disease using smartphones," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 12, pp. 3491–3500, 2020.

[16] S. Patel *et al.*, "Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors," *IEEE transactions on information technology in biomedicine*, vol. 13, no. 6, pp. 864–873, 2009.

[17] B. M. Eskofier *et al.*, "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for parkinson's disease assessment," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 655–658.

[18] G. Rigas *et al.*, "Assessment of tremor activity in the parkinson's disease using a set of wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 478–487, 2012.

[19] M. Ullrich *et al.*, "Detection of unsupervised standardized gait tests from real-world inertial sensor data in parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2103–2111, 2021.

[20] J.-W. Kim *et al.*, "Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with parkinson's disease," *Medical & biological engineering & computing*, vol. 49, pp. 365–371, 2011.

[21] M. Giuberti *et al.*, "Automatic updrs evaluation in the sit-to-stand task of parkinsonians: Kinematic analysis and comparative outlook on the leg agility task," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 803–814, 2015.

[22] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[23] X. Peng *et al.*, "Multi-scale and multi-level feature assessment framework for classification of parkinson's disease state from short-term motor tasks," *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 4, pp. 1211–1224, 2025.

[24] H. Li *et al.*, "Evaluating the utility of wearable sensors for the early diagnosis of parkinson disease: Systematic review," *Journal of Medical Internet Research*, vol. 27, p. e69422, 2025.

[25] A. Papadopoulos *et al.*, "Leveraging unlabelled data in multiple-instance learning problems for improved detection of parkinsonian tremor in free-living conditions," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3569–3578, 2023.

[26] L. Sigcha *et al.*, "Deep learning and wearable sensors for the diagnosis and monitoring of parkinson's disease: A systematic review," *Expert Systems with Applications*, vol. 229, p. 120541, 2023.

[27] J. L. Adams *et al.*, "A real-world study of wearable sensors in parkinson's disease," *npj Parkinson's Disease*, vol. 7, no. 1, p. 106, 2021.

[28] P. Yue *et al.*, "Wearable-sensor-based weakly supervised parkinson's disease assessment with data augmentation," *Sensors*, vol. 24, no. 4, p. 1196, 2024.

[29] P. Yang *et al.*, "Activity graph based convolutional neural network for human activity recognition using acceleration and gyroscope data," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6619–6630, 2022.

[30] J. R. Williamson *et al.*, "Detecting parkinson's disease from wrist-worn accelerometry in the u.k. biobank," *Sensors*, vol. 21, no. 6, p. 2047, 2021.

[31] L. Lonini *et al.*, "Wearable sensors for parkinson's disease: which data are worth collecting for training symptom detection models," *NPJ digital medicine*, vol. 1, no. 1, p. 64, 2018.

[32] H. Zhang *et al.*, "mhealth technologies towards parkinson's disease detection and monitoring in daily life: a comprehensive review," *IEEE reviews in biomedical engineering*, vol. 14, pp. 71–81, 2020.

[33] S. Del Din *et al.*, "Free-living monitoring of parkinson's disease: Lessons from the field," *Movement Disorders*, vol. 31, no. 9, pp. 1293–1313, 2016.

[34] N. M. Nayan *et al.*, "An interpretable and balanced machine learning framework for parkinson's disease prediction using feature engineering and explainable ai," *PLoS One*, vol. 20, no. 10, p. e0333418, 2025.

[35] C. Quan *et al.*, "End-to-end deep learning approach for parkinson's disease detection from speech signals," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 2, pp. 556–574, 2022.

[36] R. San-Segundo *et al.*, "Parkinson's disease tremor detection in the wild using wearable accelerometers," *Sensors*, vol. 20, no. 20, p. 5817, 2020.

[37] A. Mirelman *et al.*, "Digital mobility measures: a window into real-world severity and progression of parkinson's disease," *Movement Disorders*, vol. 39, no. 2, pp. 328–338, 2024.

[38] J. L. Adams *et al.*, "Using a smartwatch and smartphone to assess early parkinson's disease in the watch-pd study," *npj Parkinson's Disease*, vol. 9, no. 1, p. 64, 2023.

[39] A. J. Espay *et al.*, "Technology in parkinson's disease: challenges and opportunities," *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, 2016.

[40] D. A. Gallagher *et al.*, "Validation of the mds-updrs part i for nonmotor symptoms in parkinson's disease," *Movement Disorders*, vol. 27, no. 1, pp. 79–83, 2012.

[41] F. Caillava-Santos *et al.*, "Wearing-off in parkinson's disease: neuropsychological differences between on and off periods," *Neuropsychiatric Disease and Treatment*, pp. 1175–1180, 2015.