Multi-Scale and Multi-Level Feature Assessment Framework for Classification of Parkinson's Disease State from Short-Term Motor Tasks

Xiyang Peng, Yuting Zhao, Ziheng Li, Xulong Wang, Fengtao Nan, Zhong Zhao, Yun Yang*, Po Yang*
Senior Member, IEEE

***Abstract*— *Objective:* Recent quantification research on Parkinson's disease (PD) integrates wearable technology with machine learning methods, indicating a strong potential for practical applications. However, the effectiveness of these techniques is influenced by environmental settings and is hardly applied in real-world situations. This paper aims to propose an effective feature assessment framework to automatically rate the severity of PD motor symptoms from short-term motor tasks, and then classify different PD severity levels in the real world. *Methods:* This paper identified specific PD motor symptoms using a novel feature-assessment framework at both segment-level and sample-level. Features were selected after calculating SHapley Additive exPlanation(SHAP) value, and verified by different machine learning methods with appropriate parameters. This framework has been verified on real-world data from 100 PD patients performing Unified Parkinson's Disease Rating Scale(UPDRS)-recommended short motor tasks, each task lasting 20-50 seconds. *Results:* The sensitivity for recognizing motor fluctuations reached 88% in tremor recognition. Additionally, LightGBM achieved the highest accuracy for early detection(92.59%) and achieved 71.58% in fine-grained severity classification using 31 selected features. *Conclusion:* This paper reports the first effort to assess multi-level and multi-scale features for automatic quantification of motor symptoms and PD severity levels. The proposed framework has been proven effective in assessing key PD information for recognition during short-term tasks. *Significance:* The explanatory analysis of digital features in this study provides more prior knowledge for PD self-assessment in a free-living environment.**

Date of current version June 21, 2024.
This work was supported in part by the National Natural Science Foundation of China under Grant 62061050.
Xiyang Peng is with the University of Sheffield, Sheffield, South Yorkshire, the United Kingdom. (e-mail: xpeng24@sheffield.ac.uk).
Yuting Zhao is with Yunnan University, Kunming, Yunnan, China. (e-mail: zhaoyuting@mail.ynu.edu.cn).
Ziheng Li is with Yunnan University, Kunming, Yunnan, China. (e-mail: liziheng9050@mail.ynu.edu.cn).
Xulong Wang is with the University of Sheffield, Sheffield, South Yorkshire, the United Kingdom. (e-mail: xl.wang@sheffield.ac.uk).
Fengtao Nan is with Yunnan University, Kunming, Yunnan, China. (e-mail: fengtaonan@gmail.com).
Zhong Zhao is with Yunnan First People's Hospital, Kunming, Yunnan, China. (e-mail: wasx-1128new@163.com).
Yun Yang and Po Yang are co-corresponding authors.
Yun Yang is with Yunnan University, Kunming, Yunnan, China. (e-mail: yangyun@ynu.edu.cn).
Po Yang is with the University of Sheffield, Sheffield, South Yorkshire, the United Kingdom. (e-mail: po.yang@sheffield.ac.uk).

***Index Terms*— decomposition, disease recognition, feature fusion, healthcare wearables system, machine learning, multi-dimensional, parkinson's disease**

I. Introduction

PARKINSON'S disease(PD) [1] is the second most common chronic neurodegenerative disease affecting $2 \sim 3\%$ of the population aged 65 or older. The global burden of disease study [2] estimates that the number of PD cases will double from about 7 million in 2015 to about 13 million in 2040 due to the ageing and increasing life span of the global population. PD is clinically diagnosed by questionnaires [3] and motor diaries [4] according to the Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS), which is time-consuming, less compliance, large recall bias and diary fatigue. A research study [5] shows that prodromal PD could be recognized by a combination of nonmotor markers (olfactory loss, depression, anxiety), motor measures (REM sleep behaviour disorder), and biomarker testing (Dopaminergic PET or SPECT abnormalities). Furthermore, it is also introduced that the markers of the clinical PD stage include progressive bradykinesia, rest tremor, and rigidity.

Recently, most studies [6], [7] have claimed that inertial sensors could quantify some motor symptoms in PD by using the data from accelerometers, gyroscopes, and magnetometers. Specifically, inertial wearable devices [6] are used for quantifying tremors and bradykinesia by sitting, palms extended forward, and finger tapping for 10 seconds. Moreover, the freezing of gait (FOG) symptom [8] has been recognized from the 8-hour walking task with a sensitivity of 73.1% and a specificity of 81.6% by a set of inertial sensors. Though the automated methods based on inertial sensors have outperformed neurologists, they are limited to the lab, using professional devices worn on the thumb and index finger. Therefore, a study [7] has used a smartwatch on the wrist aiming to monitor real-world motor fluctuations in PD. It has been demonstrated that the smartwatch can accurately estimate displacement, showing a strong correlation (r = 0.98) with simulated tremors in different positions. The maximum displacement also correlated significantly (Spearman correlation = 0.80) with 5-level MDS-UPDRS tremor severity, while tremor constancy was positively correlated (Spearman correlation = 0.72) with mean daily

tremor estimates. Moreover, in the free-living environment, using only one wrist sensor [9] has proven the ability to continuously capture resting tremor and hand bradykinesia. There are also studies [10] collecting inertial data of PD from personal smartphones in the wild and found that the utilisation of 454 unlabelled subjects' data could help 9% increase in F1-score for tremor detection in 45 subjects with ground-truth label. Meanwhile, the abnormal Rapid eye movement (REM) percentage [11] has been successfully detected by a smartwatch-based sensor, yielding results in the control group (1.6±1.3%), PD group with clonazepam (2.0±1.7%), and PD group without clonazepam (5.7±7.1%). However, most of these studies are based on long-term monitoring or simulated motor symptoms. Long-term feature assessment frameworks have limited utility and acceptability due to high time consumption, requiring patients to wear devices for several days. Therefore, to extract valuable information from real-life signals, it is necessary to include multi-scale feature in feature assessment.

Some segment level features [12] including time domain (mean, max, min, range, root mean square, axis correlation, autocorrelation, skew, and kurt), frequency domain (dominant frequency, energy, entropy), and hybrid domain could be extracted from each segment after sliding window segmentation. Typically, segments are derived by utilizing a fixed window segmentation with overlapping rates from the raw signals, which is not suitable for the free-living environment due to the presence of diverse motor patterns in continuous signals. An event-based adaptive sliding window [13] approach has been proposed to solve this problem by expanding or contracting the time window according to the detected activity type. And the results showed that it outperforms traditional methods by 15.3% in static conditions, but demonstrates less improvement in dynamic scenarios, mainly due to limited analysis over the long term. Besides, tremendous feature fusion methods have been proposed to better combine different aspect features for training.

Except for the difficulty in deciding the segmentation approach, there are some other challenges when extracting effective features in the free-living environment. There is a gap between data quality that collected in a free-living environment and data collected in a controlled environment. For example, it is common for participants to wear only one consumer-level device instead of multiple professional devices in the free-living environment, because multiple professional wearable devices may bring economic pressure on patients and hinder the patient's daily activities. In the data collected from a free-living environment, the collection frequency may deviate along with the time, and environmental noise [14] may occur. On the other hand, less guidance may result in many anomalous activity patterns [15] exist in a free-living environment, and it is hard for us to recognize the reason (symptom result or intentional human influence) for these anomalous patterns. It is important to analyse the features and remove some personalized information by using suitable signal decomposition and signal transfer methods.

This study aims to assess features extracted from short-term motor tasks in a free-living environment. The feature assessment framework proposed in this study can extract symptom-related information from real motor task signals. The contribution of the study is below:

- A multi-scale(segments, samples) and multi-level(time, frequency, spectrum, autocorrelation) feature assessment framework is proposed for PD symptoms detection and PD severity classification through short-term motor tasks. Walking Around(WA) task shows high performance over all of the short-term motor tasks in fine-grained PD severity classification.
- A comprehensive experimental analysis is carried out to assess the significance of digital features using LightGBM and SHAP. 31 out of 636 features were selected to improve the accuracy of fine-grained PD classification through WA task.
- A detailed evaluation of various classifiers is given to assess different domain features. The ensemble classifier LightGBM exhibits strong performance in short-term motor tasks. In WA task, sample-features contribute to early detection and spectrum-features contribute to fine-grained severity classification.

The remainder of the paper is structured as follows: Sec. II describes the related works in this paper; Sec. III shows the methodology we use in this work; Sec. IV describes the experiments' results; Sec. V offers some discussion; and the conclusion is given in Sec. VI.

II. Related Work

A. PD Diagnosis Criteria

The MDS-UPDRS [16] is a widely used clinical tool for assessing Parkinson's disease. MDS-UPDRS consists of four sections, Part III is the most significant section as it examines motor abilities, while Part IV assesses motor symptoms. However, the MDS-UPDRS has some limitations, such as being subjective, time-consuming, and requiring a clinical setting. There have been attempts [17] to identify PD patients using statistical methods. However, these methods have limitations as they lack accurate quantification of tremor score, which means they can only determine whether there is a tremor, without a more accurate classification of tremor score.

Recently, various low-cost devices have been used to assess motor impairment in Parkinson's disease. For example, wearable sensors on the index finger, arm, and wrist joint are used to test the upper extremity and lower extremity [18] in the free-living environment. Another work [19] used a pen-and-tablet device to test hand movement and muscle coordination to classify PD group and Healthy control group. They introduced some features which correspond to the variability of the pen tip's velocity: the deviation from the horizontal plane, and the trajectory's entropy. In addition, multimodal assessment [20] of PD including speech, handwriting, and gait have been studied in work based on a deep learning model. Unlike these studies, our experiments explored more clinical motor tasks for assessing PD motor performance and conducted a more detailed classification.

Existing research on motor symptoms has been focused on tremor and bradykinesia. A hierarchical framework [9] has been proposed to monitor tremor and bradykinesia in

daily life. They used tremor constancy, tremor amplitude and hand bradykinesia score in at least 2 hours monitoring data. They used a third-order 3.5-7.5 Hz Butterworth IIR filter to preprocess the data and then calculated the tremor amplitude by root mean square. They introduced four indicators to measure bradykinesia: slowness, hesitancy, poverty and absence. Another work [21] introduced amplitude and periods of movements as bradykinesia features, and these features have been proven to correlate well with the UPDRS scores(r=-0.83, p=0.001). Based on this work, we calculated the number of peaks to assess the bradykinesia symptom, which could represent the periods of movement. We calculated the signal amplitude with other statistical features to quantify the severity of the tremor.

B. Wearable Signal Features

There are many time-domain and frequency-domain baseline features have been applied in the field of computer-aided diagnosis. In addition to time-domain and frequency-domain features, which are mostly generated from an individual axis of a sensor, there are hybrid-domain [22] features extracted from signals combined with multiple sensors(accelerometer, gyroscope, magnetometer) or multiple axes of a sensor. For example, the signal magnitude area (SMA) is commonly used to assess physical activity levels, especially to distinguish static activities(sitting) from dynamic activities(walking). Tilt, rotation, and yaw angle are calculated by combining the values from both the accelerometer and gyroscope to represent the postural orientation of the subjects. What's more, some contextual [23] features have been proposed to extract the correlation information between contexts in time sequence. However, these features erratically behave in PD diagnosis because of the diversity of environmental settings and participants in a free-living environment. It is hard to decide which features are effective. However, our aim is to present relevant studies about hybrid features to present temporal correlation and spatial correlation between accelerometer and gyroscope.

The effective features of different tasks may differ because of the diversity of motor tasks in the real-life environment. Nowadays, many feature selection algorithms [24] combined with machine learning have been applied to the selection of key features and quantification of disease level, including Principal Component Analysis(PCA), filter, wrapper subset evaluation, and embedded. Factor Analysis(FA) is also commonly used to fuse the signal properties derived from individual sensors and has been employed across various domains, including sport training [25], and computer-aided medication [26]. This technique generates new synthetic features after the process of dimensionality reduction, potentially reducing the computational burden of the model. Simultaneously, it could enhance model accuracy by eliminating redundant (highly correlated) and irrelevant (low variance) features. However, the effectiveness of these feature selection methods is unknown due to the diversity of activities in the free-living environment. Therefore, this study rigorously compares diverse feature selection methods in the quantification of Parkinson's disease (PD) severity and proposes a robust feature selection strategy after comprehensively analyzing the features across distinct motor tasks.

C. Machine Learning for PD Diagnosis

Several decision-level approaches such as voting, boosting, bagging, and stacking have critically combined various intermediate classification results to get an improved result. For example, Sarfaraz Masood [15] explores the correlation between voice features calculated from PD subjects and proposes a two-level ensemble-based feature extraction framework.

For motor symptoms recognition, San-Segundo [22] extracted tremor spectrum features based on non-negative factorization and Convolutional Neural Networks (CNNs) features to detect tremor segments in daily activities and several motor tasks from UPDRS. But it can only detect whether the tremor symptom exists or not, and can not recognize the severity and the type of tremor(resting, postural, and kinetic). Another deep learning-based work [18] found that weights learned from unsupervised pre-training models with unlabelled data would improve the performance of classification. However, the operation of this model requires a large amount of unlabeled data for pre-training.

In the field of PD severity classification, researchers aim to link the features with specific motor symptoms to infer a larger health marker that can reflect people's health conditions. Rehman RZ [27] found the optimal combination of clinically relevant gait characteristics(mean step velocity, mean step length, step length variability, mean step width and step width variability) by SVM for early classification of PD. But their work can only distinguish PD and HC and there is no generalization in their feature selection method.

III. Methodology

This section describes the proposed framework of PD self-diagnosis. As Fig. 2 shows, tremendous wearable technologies(signal decomposition, signal segmentation, signal transformation, feature sorting) combined with machine learning algorithms such as XGBoost or LightGBM) can construct new approaches for effectively discriminating different PD states. In the preprocessing step, apart from normalization, filtering, and segmentation, the raw signals were transferred to more domains to represent the motor fluctuation characters. In the feature extraction step, different feature groups are extracted in various dimensions from both sample-level and segment-level signals. Lastly, features are resorted by reweighted SHAP value and machine learning methods with leave-one-subject-out validation method are used to verify the effectiveness of this framework.

A. Experiment Protocol

In this study, all data collection was conducted under the collaboration of Yunnan University and Yunnan First People's Hospital. The data collection started in September 2022 and ended in March 2024, and was supported in part by the National Natural Science Foundation of China under Grant 62061050. All participants provided written informed consent

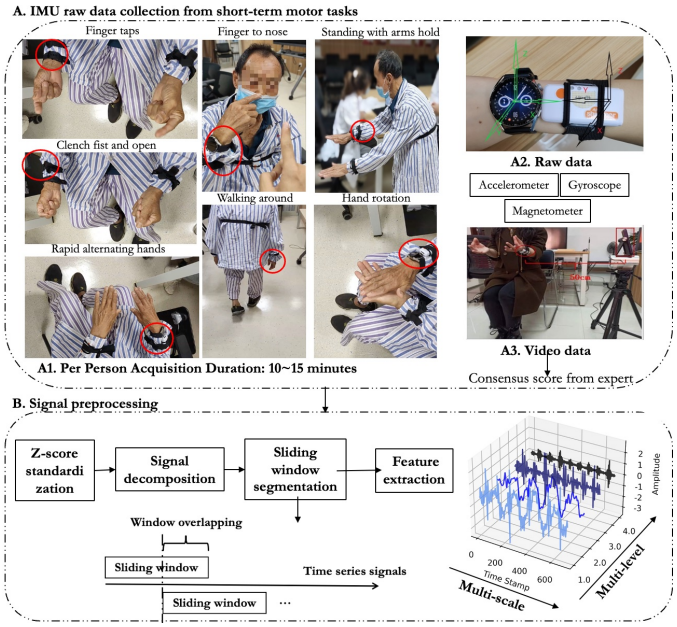

Fig. 1. Data collection and signal preprocessing

prior to their involvement in the study, and all experiments adhered to the guidelines outlined in the Code of Ethics of the World Medical Association (Declaration of Helsinki). This experiment utilized external IMU sensors and involved only non-invasive monitoring of physiological parameters. The experimental protocol and ethical approval were obtained from Yunnan University and Yunnan First People's Hospital. Subsequently, AntData Company and the University of Sheffield focused on data analysis and algorithm design. We performed only one measurement point per patient because of the difficulties in tracking a large number of patients at the same time in real clinical settings. There have been numerous research on long-term monitoring [28], [29] and short-term clinical classification [30], [31], and our work focuses on the latter one. It is challenging to conduct long-term follow-up on patients' conditions or obtain their medical histories. Nevertheless, our experiment simulates the process of motion diagnosis in the patient's clinic, which can reflect the severity of the patient's condition at a certain time. This contributes to quantifying and selecting disease-related features using non-invasive sensors.

During the experiment, a professional wearable device(Shimmer:200Hz sampling rate) was worn on the right side of the wrist, and a mobile phone was fixed 50 centimetres away from the participants throughout the session for recording. The rating scores are based on the assessments of three experts after reaching a consensus according to the UPDRS form. All of the gathered data were used for analysis without any objective inclusion or exclusion criteria. As shown in Tab. I, 100 PD patients were recruited from a hospital for this experiment, and 25 young healthy people and 35 old healthy people were invited as a control group. In Tab. II, participants are instructed to perform 14 activities(Fig. 1) and execute each activity as quickly as they can in around 20 to 60 seconds, which is significantly different from existing PD diagnosis research based on long-time monitoring. The activities in this research are chosen from UPDRS [32] (Finger Tapping, Clench Fist and Open, Hands Rotation, Hand Alternating-right/left, Finger to nose-right/left, Standing with arms held) and daily life(Walk Back and Forth, Arising from a Chair, Drinking Water, Pick Things, Sitting, Standing).

TABLE I
DEMOGRAPHIC DATA OF STUDY POPULATION

Severity	Age	Weight	Height	M/F	P	S
Old Healthy	63.3(10.7)	62.6(11.6)	159.8(8.0)	15:20	35	17687
Young Healthy	22.8(1.0)	64.4(13.3)	171.0(8.5)	18:07	25	9263
PD(mild)	67.5(10.5)	63.3(8.3)	162.4(8.2)	14:12	26	10677
PD(moderate)	66.1(9.8)	57.5(9.2)	160.6(7.6)	22:15	37	15238
PD(severe)	67.8(8.9)	59.1(8.4)	159.7(7.0)	19:18	37	15234

M/F: Male/Female, P: Number of participants; S: Number of Samples

TABLE II
SHORT-TERM MOTOR TASKS DESCRIPTIONS

Activity Abbreviations	Motor Tasks	Duration Time mean (std)
FT	Finger tapping	23.37 (4.44)
COA	Clench and open alternately	23.04 (3.41)
ALTER	Rapid alternating hands	23.09 (4.04)
HR-R	Hand rotation right	23.40 (4.22)
HR-L	Hand rotation left	23.34 (4.36)
FN-L	Finger to nose left	23.32 (3.91)
FN-R	Finger to nose right	22.80 (3.24)
STANDH	Standing with arms hold	22.63 (3.62)
WA	Walking around	50.30 (20.28)
AC	Arising from chair	28.47 (11.83)
DRINK	Drinking water	24.74 (3.78)
PICK	Pick things	25.83 (6.60)
SIT	Sitting	24.57 (6.70)
STAND	Standing	23.48 (3.89)

B. Annotation Criteria for PD

The labels used in this experiment are annotated by signal experts and checked by neurologists from Yunnan First People's Hospital based on data collection videos, which could only reflect the activity score of PD patients during the outpatient period.

*a) **PD severity stages**:* In this experiment, the severity level scores are given according to the description in Part III of the UPDRS form. It is well known that in the Hoehn-Yahr(HY) scale [33], the primary difference between HY-1 and HY-2 is whether functional impairment is present only in one limb. Since all data in this experiment were collected from a single wrist sensor, we chose to use the 'Mild' level to replace both HY-1 and HY-2. Additionally, the sample in HY-4 and HY-5 is relatively small and could result in class imbalance issues when trained by the model directly. Therefore, HY-4 and HY-5 are combined in this study to obtain a relatively balanced dataset. Tab. III presents the criteria for annotating the PD severity. It explains the labelling rules of the score and how they match with mild, moderate, and severe levels of the disease in this experiment. The labelling rules in this table clearly state that the score for a patient is decided according to the number of pauses, the speed, and the rate of decrease of amplitude during short-term motor tasks in most motor

TABLE III
PD SEVERITY ANNOTATION CRITERIA

Severity Level	**Mild**		**Moderate**	**Severe**	
HY Score	severity 1	severity 2	severity 3	severity 4	severity 5
FT, COA, ALTER, HR-R/L, FN-L/R, DRINK, PICK	No problem	1-2 pauses; Slightly slow; Amplitude decreases near the end.	3-5 pauses; Mildly slow; Amplitude decreases at half the tasks.	5 or more pauses; Moderately slow; Amplitude decreases from the beginning	Unable or barely able to perform.
STANDH	25cm extended distances	15∼25cm extended distances	10∼15cm extended distances	5∼10cm extended distances	0∼5cm extended distances
WA	No problem	Slight difficult	One hand does not swing	Neither hand swings	Unable or barely able to perform.
AC	No problem	Slight difficult	Needs one side support	Needs bilateral assistance	Unable or barely able to perform.
Labeled patients (person)	2	24	37	21	16

tasks(FT, COA, ALTER, HR-RL, FN-LR, DRINK, PICK). For example, when it comes to the task of standing with arms held, the straight degree of the hands and the extended forward distance are used to distinguish between different UPDRS grades. When it comes to AC and WA, the degree of hand swing and the need for assistance were used to measure the completion of the movement. It is noticeable that there are two patients with a severity 1 HY score in the Mild severity level. This may be caused by the medical effect during the experimental session.

b) ***PD motor symptoms****:* Amplitude and constancy [9] have been utilized in recognizing PD resting tremors. However, they are only applicable to binary classification for resting tremors without analyzing those tremors occurring during dynamic motor tasks. Another study [7] employed a novel digital biomarker 'displacement' to detect motor fluctuations in daily life. They categorized longitudinal tremors into three levels using boundaries of 0.1cm, 0.6cm, and 2.2cm. Therefore, our research labelled the tremor according to both the abnormal amplitude and constancy of the symptoms. For example, severe tremor means the tremor amplitude observed from the video is $>$2cm and the constancy time is larger than 50%. Amplitude 0.5-2cm tremor amplitude and 26-50% constancy time mean moderate tremor. Amplitude $<$0.5cm and the constancy time less than 25% is mild tremor. Amplitude is the core characteristic of recognizing tremor, followed by constancy. In addition, the number of activities seen in the video is utilized to classify the bradykinesia level. The constancy of interruptions has also been recorded as part of PD motor symptoms.

C. Multi-Scale and Multi-Level Feature Engineering

Tab. IV shows the multi-scale and multi-level features in PD area. These features are extracted after normalization, transformation and segmentation. For example, signal magnitude vector(SMV) and absolute vertical acceleration (AVA) are normally used to fuse the channels from the accelerometer and gyroscope in this experiment.

a) ***Signal preprocessing****:* Data collected from individuals have various initial values at the start of signals, so we normalized the data by using the Z-score algorithm to make the signal vibrate around 0, and use band-pass filtering and sliding window to prepare the data. Then, raw time domain signals could transferred into autocorrelation signals, frequency domain signals, and power spectral density (PSD) signals respectively. In addition, using a single axis to calculate the tremor displacement may be infected by motion drift or gravity components. Therefore, A-axis is used to represent the square root of the sum of squares of three raw axis, and T-axis is the arccos value of the z-axis.

b) ***Multi-Scale Feature Analysing*** *:* Multi-scale features(Fig. 2(c)(d)) include extracting features both at the sample level and the segment level. At the sample level, features are extracted from the whole 20-50s signals and mainly show the trend from the whole sample. The fluctuation of amplitude in the time domain, the concentration area in the spectral domain, and the maximum autocorrelation value in the autocorrelation domain are recommended to be more valuable to show motor symptoms. At the segment level, the best-size activity window needs to contain at least one activity period, so the selection of window size is related to the basic time period of different activities, and then 50 % overlapping is used to avoid cutting a cycle into two windows.

The sample level features mainly include peak numbers, whole sample entropy, difference trends, axis correlation coefficients, and peak values. The peak detection algorithm is important for counting the repetition number of the activities from the whole time of signals. It has been found that the sum of peaks' frequency and amplitude performs better than using peaks' frequency and peaks' amplitude separately. The sum of the peaks' values (peaksXY1-5) in the X (frequency) and Y (amplitude) axis were recorded according to Algorithm 1. In Algorithm 1, the peak is recognized by Minimum Peak Height(MPH), which is formulated as

$$\mathrm{mhp} = Q_{max} + (Q_{max} - Q_{min})/N, \tag{1}$$

where Q_{max} and Q_{min} are respectively the 97_{th} and 5_{th} percentile excluding NAN values, and N is the number of samples. Zero is used for padding if there are less than five peaks in a signal.

The maximum value in the autocorrelation domain is proposed to show the ratio of pauses during the motor tasks in this experiment. Multiple signals after different time delays can be fused by a softmax function for temporal fusion, and similar to this principle, the autocorrelation signal can reflect the similarity between the original signal and the signal after lagging for different time steps. The autocorrelation function

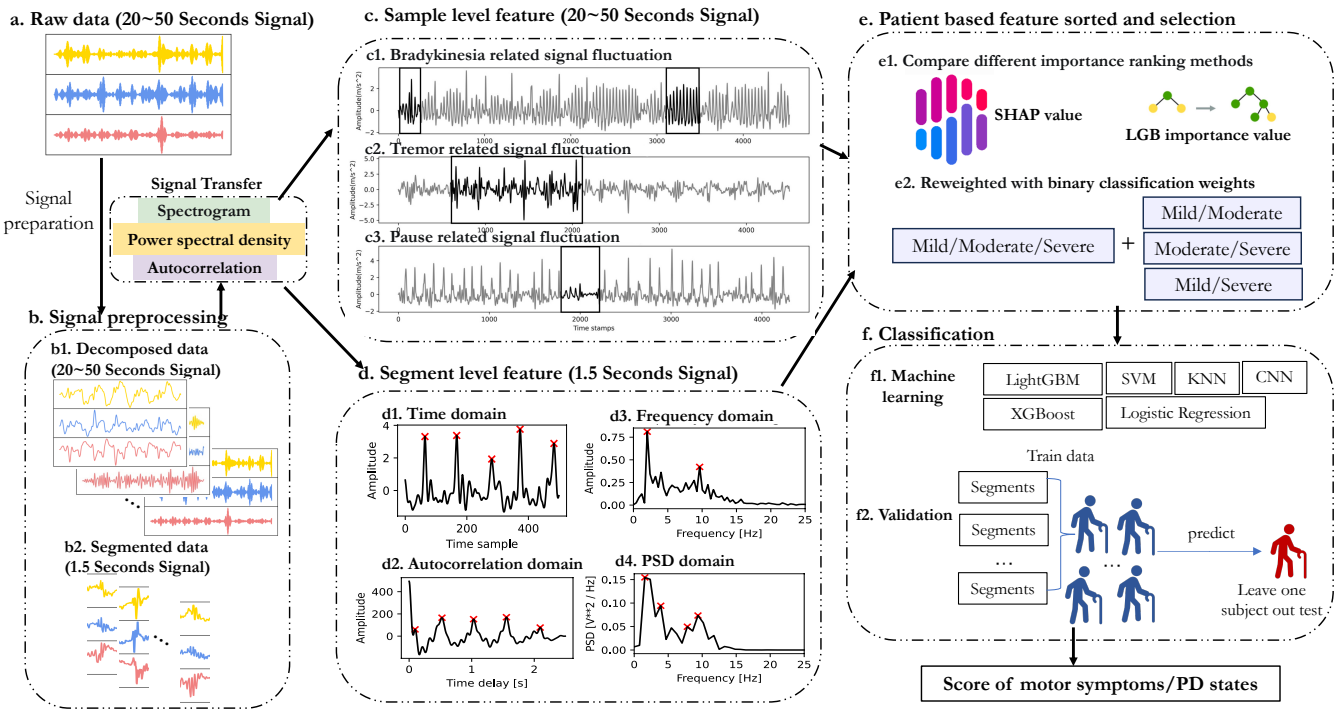

Fig. 2. Proposed feature assessment framework.

TABLE IV
FEATURE EXTRACTION FROM IMU

Sample Level			
	Feature Type	**Feature Abbreviations**	**Dimension**
Time domain	Number of peaks, Number of abnormal peaks, Sample entropy, Information entropy, Windowed mean/variance difference, Axis correlation	peaks_normal/abnormal, fea_sampleX/Y, fea_inforX/Y, meandif, vardif, t_xyCor, t_xzCor, t_xaCor, t_yzCor, t_yaCor, t_zaCor,	14
Frequency domain	Main frequency	f_peakXY1-5, f_DF,	6
Spectrum domain	Main Spectrum	p_peakXY1-5, p_energyXYZ, p_concent,	7
Autocorrelation domain	Autocorrelation coeffiencient	fea_autoy, fea_auto_num, a_peakXY1-5,	7
Segment Level			
	Feature Type	**Feature Abbreviations**	**Dimension**
Time domain	Amptitude envelope area, Mean, Max, Standard, Variance, Information entropy, Log energy, sma, Interquartile range, Skew, Kurt,Root mean square, crest factor, main peak value, sub peak value, peak difference	∼_amp_*, ∼_mean_*, ∼_max_*,∼_std_*, ∼_var_*, ∼_entr_*, ∼_lgEnergy_*, ∼_sma_*, ∼_interq_*, ∼_skew_*, ∼_kurt_*, ∼_rms_*, ∼_cftor_*, ∼_mainX_*, ∼_mainY_*, ∼_subX_*, ∼_subY_*, ∼_difX_*,∼_difY_*,	56
Frequency domain			76
Spectrum domain			76
Autocorrelation domain			76

'∼' means features could be replaced by 't_','f_', 'p_','a_' represent features from time domain, frequency domain, spectrum domain and autocorrelation domain separately. '*' could be replaced by '_x', '_y', '_z' and '_a' means features extracted from the single axis and the fusion axis. sma: signal_magnitude_area, fea_sampleX/Y: sample entropy, fea_inforX/Y: information entropy, meandif: mean of peak difference, vadif: variance of peak difference, t_xyCor: axis correlation coefficients of x and y, DF: domain features, interq: interquatile range, cftor: crest factor

is calculated as

$$R_{xx}(\tau) = \int_{-\infty}^{+\infty} x(t)x(t+\tau)\mathrm{d}t, \tag{2}$$

where τ is a time shift variable monotonically increasing.

In addition to the autocorrelation coefficient extracting the context information from the signal, there are other context-related features that represent the information between different windows after signal segmentation.

Mean trend μT and windowed mean difference μD [34] are calculated as

$$\mu T = \sum_{i=2}^{N} \left(|\mu_i - \mu_{i-1}|\right), \mu D = \sum_{i=1}^{N} \left(|\mu - \mu_i|\right), \tag{3}$$

where μ_i denotes the mean of segments, and μ denotes the mean of whole sample. In addition, the variance tend σT^2 and windowed variance difference σD^2 are formulated as

$$\sigma T^2 = \sum_{i=2}^{N} \left(\left|\sigma_i^2 - \sigma_{i-1}^2\right|\right), \sigma D^2 = \sum_{i=1}^{N} \left(\left|\sigma^2 - \sigma_i^2\right|\right), \tag{4}$$

where σ_i denotes the variance of segments, and σ denotes the variance of whole sample.

Spectral concentration is the sum of three axis' energy power around the main frequency, which can be formulated

as

$$E_{\mathrm{DF}} = \int_{\mathrm{DF}-0.4}^{\mathrm{DF}+0.4} P_x(f)\mathrm{d}f + \int_{\mathrm{DF}-0.4}^{\mathrm{DF}+0.4} P_y(f)\mathrm{d}f + \int_{\mathrm{DF}-0.4}^{\mathrm{DF}+0.4} P_z(f)\mathrm{d}f, \tag{5}$$

where $\mathrm{SC} = \dfrac{E_{\mathrm{DF}}}{E}$.

*c) **Multi-Level Feature Analysing**:* In the experiment, raw time series data could be transferred into different levels (Fig. 2(d)) including frequency level [35], spectrum level, and autocorrelation level. Some statistical measures such as mean, max, min, standard deviation(Std), root mean square(Rms), peak-to-peak amplitude(Ptp), zero crossing rate(Czr), log-energy, percentiles, interquartile range (Interq) could calculated directly from the transferred data. What's more, some domain-related features including kurtosis, skewness, dominant frequency(Domifq), spectral energy(SpecEgy), and spectral entropy(SpecEnt) could be extracted from the specific domain level.

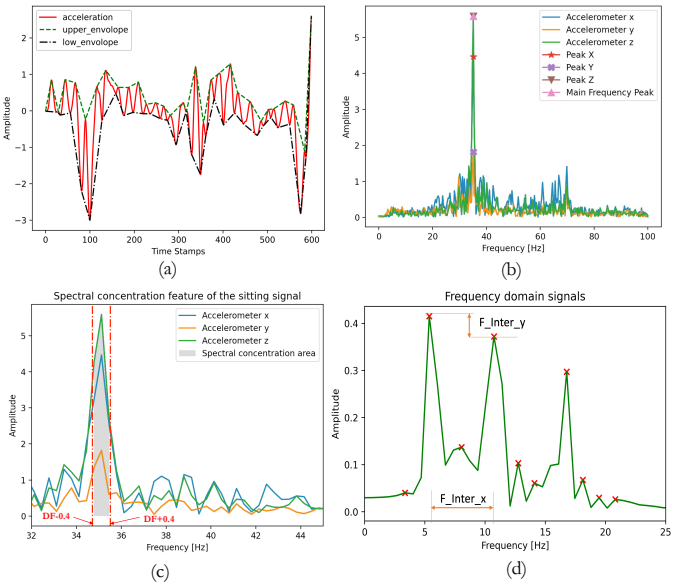

Fig. 3. Key feature visualization. (a) Amplitude envelope, (b) Main frequency domain, (c) Spectral concentration, (d) Peak difference in frequency domain.

For PD patients, tremors can be divided into Rest Tremor 3∼7Hz, Postural Tremor 8∼14Hz, and Kinetic Tremor 1∼2Hz, and the tremor information is mainly contained in the 3∼14Hz frequency band. In this experiment, 0.3∼17Hz bandpass filtering is used to pre-process the signals. In this experiment, fast Fourier transform(FFT) is used to transfer signals from the time domain to the frequency domain, FFT is the most important stochastic signal analysis technique used for analyzing time series signals, which can decompose the raw signals into different periodic components(spectrum).

We demonstrate some effective domain features, i.e., Amplitude envelop, Main frequency domain, Spectral concentration, Low-frequency signals, in Fig. 3.

Algorithm 1 Peak Features Extraction Algorithm

Input: Signals S, Required peak numbers R, Candidate peak points (Peak_x, Peak_y)
Output: FeatureList

while window slides within the sequence **do**
  $\mathrm{mph} \leftarrow \mathrm{mph} = Q_{max} + (Q_{max} - Q_{min})/N$,
  Assigning candidate points as final peak points $peak$ **while** Peak_y $>$ mph.
  count the number of final peak, assigning as M.
  FeatureList.append(sample entropy(Peak_x))
  FeatureList.append(sample entropy(Peak_y))
  if M<R **then**
    feature ← Peak_x +Peak_y **for** peak = 1, ⋯, M
    feature ← 0 **for** peak = M+1, ⋯ , R
  else
    feature ← Peak_x +Peak_y **for** peak = 1, ⋯, R
  end if
  FeatureList.append(feature)
end while
return FeatureList

Amplitude envelop (mAmp) is formulated as

$$\mathrm{mAmp} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathrm{env}_{\mathrm{upper}}(n) - \mathrm{env}_{\mathrm{lower}}(n)\right), \tag{6}$$

where $\mathrm{env}_{\mathrm{upper}}$ means the upper envelope, and $\mathrm{env}_{\mathrm{lower}}$ means the lower envelope. The envelope line is plotted according to the Hilbert function, and mAmp mean value of the amplitude gap between the upper and lower envelope.

In Fig. 3(d), the discrepancy between the value of primary and secondary peaks is recorded, where F_inter_x means the discrepancy between the primary and secondary peaks, and F_inter_y means the discrepancy between the primary and secondary frequencies.

In addition, Crest Factor describes the sharpness of the signal in a segment, which is calculated as

$$\text{Crest Factor (cftor)} = \frac{\text{Peak Value}}{\text{RMS Value}}, \tag{7}$$

where the Peak Value is the maximum peak in the segments, and the (Root Mean Square) RMS Value reflecting the signal's energy magnitude.

*d) **Feature Selection**:* In Fig. 2(e), we present a statistical analysis to select the feature attributes that provided a statistically significant separation between the Mild, Moderate, and Severe levels. We used SHAP value and LightGBM importance to quantify the features for each patient separately and then sorted the features by calculating the sum value for all subjects. Different numbers of reordering features are used to train the model to select the optimal number of key features kinematically. In this study, to further extract the most effective features, we reweight the features based on SHAP value with additional binary classification weights. It could achieve similar accuracy with fewer feature dimensions, but it did not improve accuracy in most of our experiment results and tended to be unstable when motor tasks changed. Even so,

this strategy shows potential for extracting fewer key features through short-term motor tasks.

LightGBM-based permutation importance cannot effectively deal with features with medium influence. For example, some features have a large influence on a small number of samples, but have little influence on the whole, or maintain a moderate influence in all samples. However, SHAP is a model interpretation method which could assess the features more correctly. Its core idea is to calculate the marginal contribution of each feature when it is added to the model, and then take the average of them in the case of all feature sequences.

D. Classification and Evaluation

XGBoost, LightGBM, SVM, KNN, Logistic regression, and Convolutional Neural Networks are investigated in this experiment to recognize PD severity. XGBoost and LightGBM have been well-known as boosting algorithms in recent years, and their base learners are all decision trees with greedy ideas growing. LightGBM provides higher accuracy and shorter training time than XGBoost. For more details about boosting, it is a common ensemble learning algorithm, which trains a series of weak learners and combines the prediction results of all learners as the final prediction result. During the learning process, the later learner pays more attention to the errors in the learning of the previous learner.

We employ the Leave-One-Subject-Out (LOSO) cross-validation as the protocal to evaluate the performance of our method. For a dataset with S patients, there are totally S fold experiments. The test set includes the samples from one particular patient while the training set contains the samples from the remaining patients, in each fold. The final performance collects the results from all test sets and then is calculated based on the collection. Each series of data was cut into N segments according to the fixed window size, and then the probability of these segments being divided into different categories was calculated. The prediction value of a patient is the mode of prediction values for N segments. Throughout the experiment, the test data would not participate in any process of the model training.

IV. Experiments Results

A. Motor symptoms recognition

There have been numerous research [36], [37] focus on discovering the main features related to PD motor symptoms like tremor, bradykinesia, and rigidity. Amplitude, main frequency [38] have been proved the importance when recognizing tremor. Dominant frequency, orientation angle, and peak-to-peak value [39] have been proven to be effective when recognizing bradykinesia. Sample level features including 'peaks_normal', 'fea_sampley', 'fea_samplex', 'fea_infory', 'fea_inforx', 'peaks_abnormal', 'fea_autoy', 'fea_auto_num', 'meandif', 'vardif', have been tested in this experiments, and then the most relevant features are selected after ablation experiments. As Fig. 4 shows, they are highly related to the HY score. In Tab. V, tremor symptoms can be recognized in sedentary or dynamic activities. All motor tasks are used to detect tremors, but in the end, only four motor tasks have

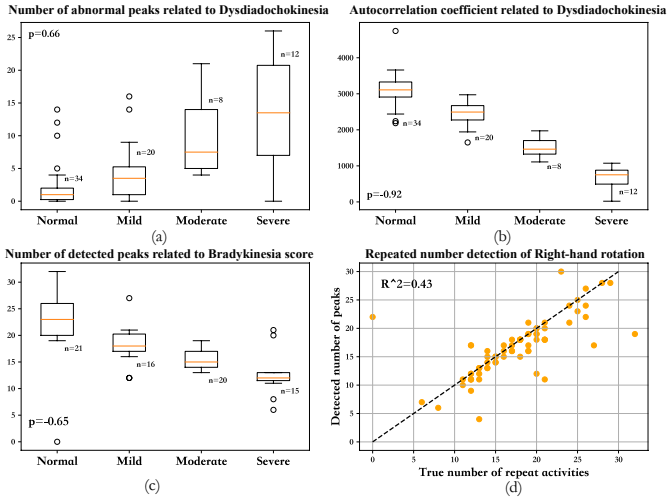

Fig. 4. Effective features from Right-hand rotation related to Dysdiadochokinesia and Bradykinesia

TABLE V
RESULTS OF FINE-GRAINED CLASSIFICATIONS OF MOTOR SYMPTOMS

	Level	precision	recall	f1-score	support
Tremor	**Normal**	0.943	1.000	0.971	33
	Slight	0.778	0.636	0.700	11
	Severe	0.667	0.667	0.667	6
	accuracy	0.880	0.880	0.880	0.88
	macro average	0.796	0.768	0.779	50
	weighted average	0.873	0.88	0.875	50
Bradykinesia	**Normal**	0.792	0.826	0.809	23
	Slight	0.818	0.643	0.720	14
	Severe	0.733	0.846	0.786	13
	accuracy	0.780	0.780	0.780	0.78
	macro average	0.781	0.772	0.771	50
	weighted average	0.784	0.780	0.778	50

a recall value of above 70%, and most of them are static activities. The Sit activity gets the highest recall value at 88%, and the results of the other three activities are all above 70%. To recognize tremor symptoms, features after filtering different bands were fused to better include information from various frequency ranges. Furthermore, the spectrum concentration features and frequency domain features in the base features have made a great contribution to the improvement of tremor recognition results.

Fig. 4(a)(b) show characteristics related to Dysdiadochokinesia recognition, while Fig. 4(c)(d) show how the number of detected peaks related to Bradykinesia recognition. The data in Fig. 4 are from the Right-hand rotation of 50 PD patients with different severity levels, a band-pass filter(0.2-2hz) was used to preprocess the axis with the highest energy value from the accelerometer in advance.

Fig. 4(a) shows that the number of abnormal peaks is positively correlated with the Dysdiadochokinesia score with a p-value of 0.66. However, only using the number of abnormal peaks for recognition still produces overlap in results, and we found that the autocorrelation coefficient can clearly reflect this property instead. In Fig. 4(b), the Dysdiadochokinesia score can distinguish different levels of disorders according to the maximum autocorrelation coefficient with a p-value of -0.92.

Fig. 4(c) represents that the number of detected peaks can

well distinguish the four levels of Bradykinesia with a p-value of -0.65. In Fig. 4(d), the X-axis represents the number of activity repetitions counted according to the video, and the Y-axis shows the number of peaks detected by peak detection algorithms shown in Algorithm 1. It is clear that the number of activities can be accurately detected by the number of peaks with R2 = 0.43.

In Tab. V, we extended the time domain features and autocorrelation domain features to identify tremor and Bradykinesia respectively and used the XGBoost algorithm to classify them into four categories: normal, mild, moderate, and severe. In the end, the weighted recall for Bradykinesia is 0.78.

B. PD severity classification

The original data from 14 activity types were respectively cut with 0.5s, 1s, 1.5s, and 3s window sizes before feature extraction to explore prior knowledge for proposing an adaptive window segmentation method on PD diagnosis. In our experiment, a 1.5-second window size with 50% overlapping was used because it is suitable for most of the activities.

Tab. VI presents the performance scores of classifiers across various short-term motor tasks for disease severity recognition. In the beginning, logistic regression was used as the baseline, but it failed to achieve high accuracy during multiclass classification. Therefore, LightGBM is employed to achieve higher accuracy. Furthermore, to tackle the imbalanced problem, scores 4 and 5 are merged into the severe-level category for subsequent severity classification. However, we observed that there was no significant increase in accuracy whether or not we combined the minority class samples. In the end, the results show that ALTER, FN-R, STANDH, WA, AC, PICK and STAND classifiers exhibit commendable performance, where FN-R achieves scores ranging from 0.6413 to 0.7667, and WA attains a score of 0.6737 in PD fine-grained classification. Therefore, further experiments will be conducted to optimise the results based on these three activities. Specifically, the effectiveness of features will be verified by comparing different feature selection methods. Additionally, to enhance the accuracy of multi-class classification, we analyzed different binary classifications and applied the feature importance learned from binary classification to the model. The data collection time for a patient is always under 10 minutes. Nearly no participants stop participating due to the length of the collection time. However, some severely ill patients are unable to complete the full set of experiments because of their physical limitations.

Tab. VI illustrates the process of selecting the optimal activities WA using the LightGBM model from multiple activities. In addition, the accuracy of severity fine-grained classification in ALTER, FN-R, STANDH, WA, AC, PICK and STAND are all above 0.6. In table Tab. VII, the accuracy of WA is further improved to 0.7158 by utilizing a 31 dimensions key feature subset. All the features from Tab. VIII are derived from only one activity WA. Even though multiple activities has been analyzed in this experiment, the results show that the best activity could be selected to assess PD through our generalized framework.

Results in Fig. 5 report the results of PD severity classification on the ALTER activity. Initially, the classification accuracy (mild/moderate/severe) for the ALTER activity was 0.6289. After applying the SHAP value feature ranking, the accuracy improved to 0.7526 with the use of 96-dimensional features. Furthermore, an accuracy of 0.5979 could be obtained with the use of 16-dimensional features after combining the feature weights of the three-class classification (mild/moderate/severe) and the binary classifications (mild/moderate, mild/severe). Fig. 5 (a-c) illustrates the accuracy per individual using leave-one-subject-out validation methods. Fig. 5(d) presents the overall results, indicating that the binary classification between mild and severe levels achieved the highest F1-score (0.896), followed by early detection with an F1-score of 0.843, and fine-grained quantification (3 classes) for PD with an F1-score of 0.753. The figure clearly shows that the 3-class classification (mild, moderate, and severe levels) performed less satisfactorily, with 24 out of 100 samples misclassified. Among these, 17 were misclassified into the severe class, with 13 having a true label of moderate and 4 from mild level. Additionally, 7 were misclassified into the moderate label, where 5 from mild-level, and 2 from severe-level. Notably, no samples were misclassified into the mild level. In the following experiments, videos combined with doctors' ratings are used to correct the activity labels to approximately 5% of the total data. Accuracy from WA activity is more stable, achieving an accuracy of 0.7159 with a small number of features (31 dimensions), which helps in model simplification. It is also found that most patients classified through the ALTER activity had only 60%-80% of their segments correctly classified, whereas patients classified through WA activity had almost every segment correctly classified. This further confirms that the ALTER activity relies on segment-level features, while the WA activity relies on sample-level features.

This experiment recorded the time of diagnosis and the time since the last medication intake for PD patients but did not select participants based on these values. In the previous work [40], we focused on the anomalies and label uncertainty in the free-living environment. In this work, our aim is to find a more generalized framework combined with feature analysis and machine learning algorithms to assess Parkinson's disease. Using strict selecting criteria may reduce our sample size and make our model less generalizable.

More importantly, drug deprivation may lead to injuries for patients, so patients were kept under periodic evaluation and levodopa and/or dopamine agonist treatment for ethical and safety reasons. As recorded in the personal information of patients, the mean (variance) of the diagnosis time for the PD group in this study was 71 (78), and the mean (variance) of the time since the last medication intake was 188 (137). The influence of medication on the results of this experiment is limited because the labels were annotated according to real-time video. Our study focuses on quantifying the differences between patients with different disease degrees through feature analysis and machine learning algorithms.

Some medication records and raw Initial Measurement Units(IMU) data are missing due to sensor disconnections or incomplete data caused by patients' conditions. The number of patients with missing specific activity data is shown in

TABLE VI
RESULTS FOR DETECTING PD SEVERITY FROM SINGLE MOTOR TASK

	LOGISTIC			LightGBM		
	PD/OHC	PD/HC	Fine Grained	PD/OHC	PD/HC	Fine Grained
FT	0.7797	0.7917	0.4184	0.8814	0.8333	0.5204
COA	0.7541	0.7059	0.4200	0.8852	0.5490	0.5500
ALTER	0.8525	0.6667	0.3814	0.8361	0.7451	0.6186
HR-R	0.8525	0.8235	0.4848	0.9016	0.6863	0.5657
HR-L	**0.9016**	0.6863	0.4343	**0.9180**	0.5882	0.5657
FN-L	0.8305	0.6875	0.4396	0.8983	0.7500	0.5604
FN-R	0.8833	0.7000	0.4565	0.7667	0.6000	0.6413
STANDH	0.5862	**0.8889**	0.4396	0.7586	0.8000	0.6044
WA	0.7500	0.8776	**0.5263**	0.8500	**0.9184**	**0.6737**
AC	0.8000	0.6500	**0.5167**	0.8667	0.725	0.6167
DRINK	**0.9259**	0.6471	0.4194	**0.9259**	**0.9412**	0.5484
PICK	0.7812	**0.9000**	0.3810	0.7500	0.7667	0.6190
SIT	0.6364	0.5000	0.3400	0.8485	0.6000	0.4800
STAND	0.8043	0.8214	0.4211	0.8043	0.7857	**0.6491**

The best and second results are shown in boldface. PD/OHC means binary classification between PD mild level and Old Healthy Control, PD/HC means binary classification between PD mild level and Young Healthy Control. Fine Grained means three class classification among mild, moderate and severe PD severity levels.

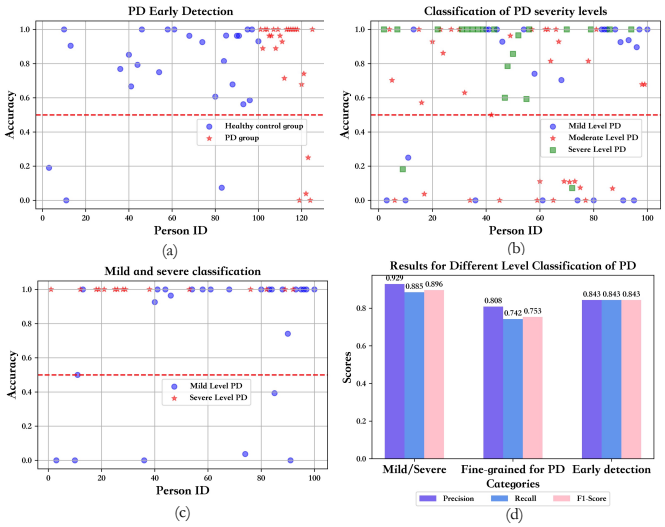

Fig. 5. Patient-based PD recognition results. (a): PD early detection from healthy control group;(b): PD severity classification among mild, moderate, and severe levels. (c): PD binary classification between mild and severe levels. (d): Results for different level severity classifications.

parentheses: HR-R(30), FN-R(2), STANDH/SIT(3), WA(4), AC(16), DRINK/PICK(20), STAND(14). It will lead to serious small sample issues and class imbalance issues if we keep selecting patients according to their medication conditions. Thus all available data were maximally utilized in this experiment.

C. Feature assessment with Machine Learning methods

Results from Tab. VIII totally depends on WA activity. It provides a comprehensive overview of performance scores for various feature groups (the description of each feature set is shown in Tab. IV) and classifiers in the context of PD severity classification. Performance metrics are evaluated based on different feature dimensions and classifiers(LGB, SVM, KNN, XGB, Logistic regression and Convolutional Neural Network(CNN)), considering early detection (binary classification between OHC and mild PD) and the fine-grained classification (mild/moderate/severe). We transformed the original feature datasets into 3D before CNN training. For WA activity, there are 6254 segments in the training data and 76 segments in the test data, and a window size of 20 with a step size of 20 are used to transfer the data to three dimension data. The small amount of data makes it unsuitable for most deep-learning algorithms in this experiment. From Tab. VIII, it is clear that in early detection, Sample_features performs better than Seg_features, except for Logistic regression. Additionally, Time_features and Autocorr_features perform better than Frequency_features and Spec_features, and Acc_features perform better than Gyro_features. The results show difference in the fine-grained PD classification. Specifically, Time_features perform poorly in the fine-grained classification, whereas Spec_features perform well. Furthermore, Gyro_features perform slightly better than Acc_features.

TABLE VII
FEATURE SELECTION AND SORTING RESULTS IN PD SEVERITY CLASSIFICATION FOR FN-R, WA, AND STAND MOTOR TASKS.

	LGB Permutance			SHAP Value		
Selected Feature Num.	**FN-R**	**WA**	**STAND**	**FN-R**	**WA**	**STAND**
16	0.6413	0.6316	0.5789	0.6196	0.6421	0.5789
31	0.6522	0.5789	0.6140	**0.6304**	**0.7158**	0.5965
46	0.6196	0.6421	0.6140	0.5870	0.6632	0.5263
61	0.6522	0.6421	0.6140	0.5761	0.6947	0.5965
76	0.6413	0.6421	0.6316	0.5978	0.6421	**0.6316**
91	0.6522	0.6211	**0.6842**	0.5978	0.6947	0.6140
106	0.6522	0.6105	0.6316	0.5978	0.6842	0.5789
121	**0.6630**	0.6105	0.6316	0.5870	0.6947	0.5789
136	0.6522	0.6105	0.6140	0.5870	0.6526	0.5789
151	0.6413	**0.6632**	0.6316	0.5978	0.6526	**0.6316**
166	0.6522	0.6211	**0.6842**	0.5870	0.6526	0.5439
181	0.6413	0.6105	0.6140	0.5761	0.6421	0.6140

The best results are shown in boldface.

Tab. IX provides detailed information on the best feature subsets used for PD fine-grained classification. In this experiment, we recorded the importance of features for each patient and calculated their sum and standard deviation. The features are sorted by the sum, as we believe this enhances the generalizability of the feature selection module, while the standard deviation reflects the stability of the features. Spectrum domain features performed well in most tasks, while sample-level features excelled in the WA activity due to the high amplitude of arm swings when walking.

Fig. 6 and Tab. VII intricately elucidate the impact of distinct feature selection dimensions and methodologies, validated using activity FN-R (20s), and WA (50s), STAND(50S) respectively. The experiment starts with an initial dimension of 16, and subsequent increments of 15 dimensions per observation until 175 dimensions. Fig. 6 (a-c) illustrates the fluctuation of accuracy as the number of feature dimensions increases. The results illustrate that employing feature selection technology based on SHAP values effectively reduces the feature dimensions from 636 to 31 for activities FN-R, WA and STAND respectively. This reduction contributes to an enhancement in accuracy from 0.6737 to 0.7158. In Fig. 6 (d-f), the mean and variance of features, ranked by their LightGBM permutation

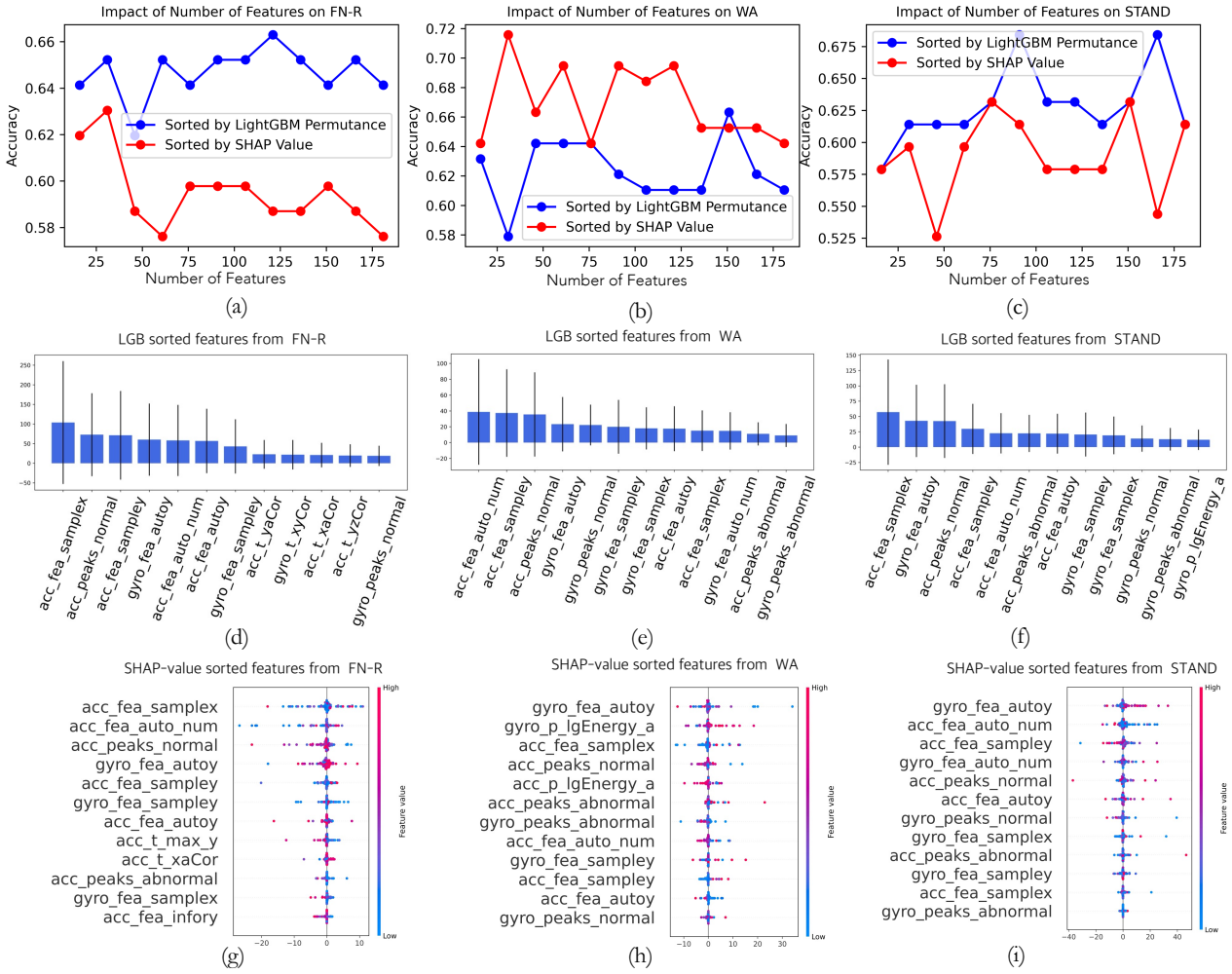

Fig. 6. Feature selection and analysis of FN_R, WA and STAND. (a-c): Accuracy of Mild, Moderate and Severe severity levels classification when selecting different number of features (sorted by SHAP value); (d-f): Features sorted by LightGBM importance; (g-i): Features sorted by SHAP value.

TABLE VIII
PD CLASSIFICATION RESULTS OF DIFFERENT FEATURE GROUPS AFTER FEATURE SELECTION AND FUSION

Classifiers	LGB		SVM		KNN		XGB		Logistic		CNN	
Feature	Early Detection	Fine Grained	Early Detection	Fine Grained	Early Detection	Fine Grained	Early Detection	Fine Grained	Early Detection	Fine Grained	Early Detection	Fine Grained
Sample_Features	**0.8500**	0.6526	0.6667	0.5579	0.6833	0.4737	**0.6500**	0.5895	0.6667	0.6000	0.7500	0.5158
Seg_Features	0.7500	0.6316	**0.7667**	0.5684	0.5000	0.5579	0.6000	0.5579	**0.7667**	0.5579	0.6000	0.5263
Time_Features	0.8333	0.5053	0.7000	0.4737	0.6500	0.5053	0.6000	0.3789	0.6833	0.5158	0.7167	0.4947
Frequency_Features	0.7333	0.6421	0.7333	0.5474	0.4333	0.4000	0.5333	0.5053	0.7167	0.5579	0.5667	0.4842
Autocorr_Features	**0.8500**	0.6632	0.7167	0.5158	0.6500	**0.5789**	0.6333	**0.6211**	0.7000	0.5263	0.6833	0.5368
Spec_Features	0.7500	0.7053	0.6833	**0.5895**	0.5333	0.5368	0.5500	0.5789	0.6667	**0.6105**	0.6500	0.5158
Acc_Features	0.8333	0.5895	0.7500	0.5053	0.6167	0.4842	**0.6500**	0.3895	**0.7667**	0.5158	0.7167	0.5053
Gyro_Features	0.8167	0.5579	0.6333	0.5474	0.5500	0.5474	0.6000	0.5789	0.6500	0.5789	0.7167	0.5368
All_Features	**0.8500**	0.6737	0.7333	0.5263	0.6000	0.5368	**0.6500**	0.5895	0.7333	0.5158	0.6667	**0.6000**
Selected_Features	0.8000	**0.7158**	0.6833	0.5263	**0.7667**	0.4421	**0.6500**	0.6105	0.6833	0.5158	**0.7667**	0.5368

The best results are shown in boldface. 'Early' means binary classification between HC and Mild PD, 'Fine-grained' means multiclass classification among mild, moderate, and severe PD.

values, are presented. Furthermore, Fig. 6 (g-i) displays the 15 features with the highest SHAP mean. Each point on the graph corresponds to the SHAP value of a single patient, with colours indicating the original value of the feature. Notably, the results indicate a substantial prevalence of sample_fea over segment_fea in the key features from WA activity. Additionally, the proportion of axis correlation in permutation value remains consistently high, regardless of changes in activity states.

V. DISCUSSION

TABLE IX
FEATURES SORTED BY SHAP VALUE IN THREE KEY ACTIVITIES

	FN-R			WA			STAND		
	feature	**sum**	**mean/std**	**feature**	**sum**	**mean(std)**	**feature**	**sum**	**mean(std)**
1	gyro_fea_samplex	10.21	0.22(1.32)	gyro_fea_autoy	170.38	3.55(18.29)	gyro_fea_autoy	69.84	2.41(11.17)
2	acc_t_xaCor	8.10	0.17(1.30)	acc_fea_auto_num	163.47	3.41(17.50)	acc_peaks_abnormal	53.86	1.86(7.67)
3	acc_t_yzCor	5.67	0.12(0.64)	gyro_fea_auto_num	68.58	1.43(7.93)	gyro_p_lgEnergy_a	40.33	1.39(6.74)
4	gyro_peaks_normal	5.06	0.11(0.84)	acc_peaks_abnormal	55.82	1.16(7.51)	gyro_fea_sampley	17.50	0.60(3.54)
5	gyro_fea_sampley	4.96	0.11(2.06)	acc_fea_autoy	38.20	0.80(5.86)	gyro_fea_auto_num	14.36	0.50(2.15)
6	gyro_f_skew_y	4.66	0.10(0.56)	acc_fea_samplex	35.74	0.74(4.26)	acc_fea_auto_num	13.38	0.46(2.71)
7	acc_t_skew_y	4.20	0.09(0.55)	gyro_fea_samplex	32.56	0.68(5.00)	acc_t_xzCor	8.51	0.29(1.25)
8	acc_p_lgEnergy_y	3.18	0.07(0.45)	acc_peaks_normal	23.01	0.48(5.61)	acc_fea_autoy	7.66	0.26(1.91)
9	acc_t_xzCor	2.52	0.05(0.56)	gyro_peaks_abnormal	10.14	0.21(1.27)	acc_t_yzCor	5.04	0.17(0.73)
10	acc_f_skew_a	1.99	0.04(0.33)	gyro_peaks_normal	8.42	0.18(4.66)	gyro_peaks_normal	4.50	0.16(1.27)
11	gyro_p_amp_a	1.62	0.03(0.20)	acc_f_mainX_z	3.33	0.07(0.39)	acc_fea_sampley	4.38	0.15(1.92)
12	gyro_p_lgEnergy_x	1.54	0.03(0.17)	gyro_t_amp_y	2.15	0.04(0.24)	gyro_t_xyCor	2.84	0.10(0.39)
13	acc_t_xyCor	1.43	0.03(0.62)	acc_a_subY_x	1.89	0.04(0.22)	acc_t_yaCor	2.70	0.09(0.58)
14	acc_a_amp_x	1.25	0.03(0.20)	acc_a_peakXY3	1.80	0.04(0.22)	acc_p_interq_a	2.28	0.08(0.31)
15	acc_p_lgEnergy_z	1.20	0.03(0.20)	acc_t_zaCor	1.56	0.03(0.28)	acc_t_zaCor	1.61	0.06(0.27)
16	acc_p_interq_y	1.15	0.02(0.19)	acc_t_xaCor	1.17	0.02(0.13)	acc_p_rms_y	1.19	0.04(0.22)
17	gyro_f_cftor_z	1.13	0.02(0.20)	gyro_a_amp_x	0.95	0.02(0.10)	acc_p_max_y	0.90	0.03(0.21)
18	acc_peaks_abnormal	1.06	0.02(0.96)	acc_f_skew_z	0.86	0.02(0.09)	gyro_p_lgEnergy_x	0.87	0.03(0.17)
19	acc_t_skew_x	0.97	0.02(0.13)	acc_t_max_a	0.54	0.01(0.06)	acc_vardif	0.62	0.02(0.09)
20	acc_f_mainX_z	0.94	0.02(0.13)	gyro_fea_infory	0.49	0.01(0.33)	acc_t_xaCor	0.61	0.02(0.15)

TABLE X
SUMMARY OF RESEARCH ON WEARABLE DEVICES APPLIED TO PD DIAGNOSIS.

Work(Year)	**Subjects**	**Data Collection Paradigm**	**Research Objectives**	**Optimal Results**
Nikhil et al. (2020) [9]	50HC, 31 PD	45min ADL monitoring	Tremor detection (binary classes)	ACC: 83%
Butt et al. (2020) [31]	64 PD, 50 HC	2.5-minute laboratory tasks	Quantify PD severity (regression)	ACC: 81.4%, RMSE: 0.101
Alexandros et al. (2020) [41]	14PD, 8HC	75-seconds ADL task	Distinguish patients and HC (binary classes)	SEN: 86%, SPEC: 93%
Rob et al. (2021) [7]	225PD, 171 HC	6 months of ADL monitoring	Detecting symptom changes (binary classes)	ACC: 94%
Mathias et al. (2021) [42]	66PD	6 days ADL monitoring	The frequency of PKG data changed treatment decisions	31.8% results changed, 88% dialogue improved
Proposed	**100PD, 35HC**	**20-seconds clinical task**	**Distinguish patients and HC, Quantify PD severity (3 classes)**	**Distinguish PD/HC ACC: 92.59%, Quantify PD ACC: 71.58%**

PKG: Parkinson KinetiGraph; PD/HC means binary classification between PD mild group and Old healthy control group in this table.

A. Clinical Application

In real clinical data acquisition, standardizing activity patterns is challenging due to environmental differences and patient diversity. However, a precise evaluation of the disease requires considering statistically significant features across various experimental paradigms. It is hard to establish solid disease-related features for PD severity classification in real-life environments. In this study, we propose an automated feature assessment framework for the quantification of the severity of motor symptoms, which could actively address changes in different environments with various tasks.

As depicted in the studies summarized in Tab. X, some quantification experiments [7], [9], [28], [29], [42] are based on long-term monitoring and are trained with longitudinal data. These studies provide less prior knowledge about key digital biomarkers at the segment level. Other investigations [24], [30], [31], [41] only require 1-2 minutes of activity data. However, these studies are confined to detecting the presence of disease and do not undertake a more comprehensive analysis of symptoms and the severity of the disease. Our model employs shorter clinical medical activities, yielding better results in distinguishing PD from HC and making notable progress in the fine classification of PD patients. We introduce an feature assessment framework for PD severity, utilizing a substantial amount of data (100 PD, 35 HC) evaluated in a real clinical scenario.

B. Feature analysis with classifier performance

In clinical settings, the acquired data exhibits variability due to the challenge of ensuring that individuals consistently perform motor tasks in a uniform posture. There are multiple features have been proposed to recognize PD, but few of them are suitable for all situations. Therefore, we performed a systematic literature review(Tab. IV) about existing PD motor features in various domains, and proposed a feature assessment framework to learn the key features in free-living environments. As illustrated in Fig. 2, our framework considers features at both multi-scale and multi-domain levels, and extracted features undergo analysis using various feature selection algorithms and machine learning methods. In

comparison to existing methods, our approach demonstrates adaptability to different experimental settings involving various motor tasks. Furthermore, the enhanced feature selection algorithm incorporated in the framework proves effective in feature reduction in particular activities. According to the experimental results, the key features keep changing under different tasks. For example, sample-level features, time-domain features, and autocorrelation-domain features are suitable for the early detection of PD, while spectrum-domain features are more suitable for the fine-grained PD classification. In conclusion, we present a novel multi-scale and multi-level feature assessment framework and verified its effectiveness from different domains and levels.

Most recent works did not highlight the choice of their architecture and training hyperparameters. This study trained six machine learning models for the automatic quantitative scoring of PD severity and found the ensemble learning algorithms(LightGBM) perform better than others. Furthermore, we chose a strict validation strategy(leaving one subject out) for unbiased evaluation of the feature performance. Each patient is selected as the test set in a loop and does not participate in the training when selected as test set. Finally, we conclude the performance of the classifiers is influenced by different feature subsets. The DRINK achieved 0.9259 accuracy in PD early detection. The automatically selected features with LightGBM were observed to reach 0.7158 accuracy for the fine-grained PD severity classification.

C. Limitations and future work

Existing studies have demonstrated that PD symptoms predominantly manifest as hand motor fluctuations. Consequently, our study is based on unilateral wrist sensors, aiming to use a limited number of sensors and achieve high accuracy. This approach benefits self-testing for PD patients in their daily lives. However, it may overlook symptoms in other body parts and their interconnections within individuals.

None of the participants were required to discontinue the drug to cooperate with the experiment. Drug deprivation may lead to injuries for patients, so patients were kept under periodic evaluation and levodopa and/or dopamine agonist treatment for ethical and safety reasons. We understand that without using exclusion or inclusion criteria, excluding the influence of medication on the experimental results can become quite challenging. However, the influence of medication on the results of this experiment is limited because the patient labels used were annotated according to real-time video, capturing the severity of movement symptoms.

Moreover, the experiment collects discrete data from each clinic visit of the patients, lacking longitudinal data comparisons for the same patient across different periods. Although we achieved 72% accuracy for mild/moderate/severe PD classification using a leave-one-subject-out validation method with around 100 patients, the accuracy declined to 62% when using an 8:2 split between training and validation with leave-one-out validation for the test set. This decline is primarily caused by the small number of samples and the issue of unbalanced data. Dividing the validation set reduces the number of training samples. Therefore, we plan to increase the collection of patient samples in the next stage and seek related data enhancement algorithms. Despite these limitations, our definition of motor symptoms through short-term motor tasks accurately represents the current mobility of the patient, making it applicable in clinical practice and potentially aiding in PD self-diagnosis at home.

VI. Conclusion

This work represents the initial attempt to assess highly relevant features from multi-scale(temporal, frequency, spectrum, autocorrelation) and multi-level(sample, segment) levels for the automatic classification of PD severity from short-term motor tasks.

Sample-features, especially 'amplitude area', 'normal/abnormal peak numbers', and 'max autocorrelation value', demonstrate higher accuracy in detecting motor fluctuations through Right-hand rotation and sit tasks. For fine-grained severity classification, 'finger-to-nose (right)', 'walking around', and 'standing' are identified as the most effective tasks, while 'drinking', 'hands rotation' perform better in early detection. Through WA task, SHAP performs better than LightGBM, and finally, 31 explainable features are extracted to improve the accuracy of fine-grained classification.

Detail feature assessment experiments are conducted through WA task. Two points of corresponding analysis are concluded. Firstly, three types of features, i.e., Sample_features, Time_features, and Autocorr_features, have outstanding performance in early PD detection, simultaneously, Acc_features outperform Gyro_features. The results are different from that in fine-grained PD classification, where Spec_features outperform Time_features, and Gyro_features outperform slightly than Acc_features.

Acknowledgment

Thanks to all the co-authors who contributed to this project. The data collection was led by Yun Yang and Zhong Zhao, facilitated by the collaboration between Yunnan University and Yunnan First People's Hospital. Thanks to Yuting Zhao, Ziheng Li, and Fengtao Nan for their efforts in data collection and data cleaning. The data analysis process was conducted under the leadership of Po Yang from the University of Sheffield. Xiyang Peng was in charge of framework design, feature analysis, and paper drafting. Xulong Wang suggested the algorithms and assisted with proofreading.

References

[1] J. Jankovic and E. K. Tan, "Parkinson's disease: etiopathogenesis and treatment," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 91, pp. 795–808, Aug. 2020.

[2] V. L. Feigin *et al.*, "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet Neurology*, vol. 18, pp. 459–480, May 2019.

[3] M. Stacy, A. Bowron, M. Guttman, R. Hauser, K. Hughes, J. P. Larsen, P. LeWitt, W. Oertel, N. Quinn, K. Sethi, *et al.*, "Identification of motor and nonmotor wearing-off in parkinson's disease: comparison of a patient questionnaire versus a clinician assessment," *Movement disorders: official journal of the Movement Disorder Society*, vol. 20, no. 6, pp. 726–733, 2005.

[4] R. A. Hauser, J. Friedlander, T. A. Zesiewicz, C. H. Adler, L. C. Seeberger, C. F. O'Brien, E. S. Molho, and S. A. Factor, "A home diary to assess functional status in patients with parkinson's disease with motor fluctuations and dyskinesia," *Clinical neuropharmacology*, vol. 23, no. 2, pp. 75–81, 2000.

[5] R. B. Postuma and D. Berg, "Advances in markers of prodromal parkinson disease," *Nature Reviews Neurology*, vol. 12, no. 11, pp. 622–634, 2016.

[6] H. Dai, G. Cai, Z. Lin, Z. Wang, and Q. Ye, "Validation of inertial sensing-based wearable device for tremor and bradykinesia quantification," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 997–1005, 2020.

[7] R. Powers, M. Etezadi-Amoli, E. M. Arnold, S. Kianian, I. Mance, M. Gibiansky, D. Trietsch, A. S. Alvarado, J. D. Kretlow, T. M. Herrington, *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.

[8] M. Bachlin, M. Plotnik, D. Roggen, I. Maidan, J. M. Hausdorff, N. Giladi, and G. Troster, "Wearable assistant for parkinson's disease patients with the freezing of gait symptom," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 436–446, 2009.

[9] N. Mahadevan, C. Demanuele, H. Zhang, D. Volfson, B. Ho, M. K. Erb, and S. Patel, "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.

[10] A. Papadopoulos and A. Delopoulos, "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, p. 3569–3578, July 2023.

[11] Y.-F. Ko, P.-H. Kuo, C.-F. Wang, Y.-J. Chen, P.-C. Chuang, S.-Z. Li, B.-W. Chen, F.-C. Yang, Y.-C. Lo, Y. Yang, *et al.*, "Quantification analysis of sleep based on smartwatch sensors for parkinson's disease," *Biosensors*, vol. 12, no. 2, p. 74, 2022.

[12] X. Peng *et al.*, "Experimental Analysis of Artificial Neural Networks Performance for Accessing Physical Activity Recognition in Daily Life," in *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, (Exeter, United Kingdom), pp. 1348–1353, IEEE, Dec. 2020.

[13] C. Ma *et al.*, "Adaptive sliding window based activity recognition for assisted livings," *Information Fusion*, vol. 53, pp. 55–65, Jan. 2020.

[14] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889, 2020.

[15] S. Laskaridis, D. Spathis, and M. Almeida, "Federated mobile sensing for activity recognition," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, (New Orleans Louisiana), pp. 858–859, ACM, Oct. 2021.

[16] 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: MDS-UPDRS: Clinimetric Assessment," *Movement Disorders*, vol. 23, pp. 2129–2170, Nov. 2008.

[17] K. Harish *et al.*, "Tremor quantification and its measurements on parkinsonian patients," in *2009 International Conference on Biomedical and Pharmaceutical Engineering*, (Singapore, Singapore), pp. 1–3, IEEE, Dec. 2009.

[18] M. Singh, P. Prakash, R. Kaur, R. Sowers, J. R. Brašić, and M. E. Hernandez, "A deep learning approach for automatic and objective grading of the motor impairment severity in parkinson's disease for use in tele-assessments," *Sensors*, vol. 23, no. 21, p. 9004, 2023.

[19] C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in parkinson's disease," *Biomedical Signal Processing and Control*, vol. 31, pp. 174–180, 2017.

[20] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: a deep learning approach," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1618–1630, 2018.

[21] Z. Lin, Y. Xiong, G. Cai, H. Dai, X. Xia, Y. Tan, and T. C. Lueth, "Quantification of parkinsonian bradykinesia based on axis-angle representation and svm multiclass classification method," *IEEE Access*, vol. 6, pp. 26895–26903, 2018.

[22] R. San-Segundo *et al.*, "Parkinson's Disease Tremor Detection in the Wild Using Wearable Accelerometers," *Sensors (Basel, Switzerland)*, vol. 20, p. 5817, Oct. 2020.

[23] P. Gupta and T. Dallas, "Feature Selection and Activity Recognition System Using a Single Triaxial Accelerometer," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 1780–1786, June 2014.

[24] C. Ma, P. Zhang, J. Wang, J. Zhang, L. Pan, X. Li, C. Yin, A. Li, R. Zong, and Z. Zhang, "Objective quantification of the severity of postural tremor based on kinematic parameters: A multi-sensory fusion study," *Computer Methods and Programs in Biomedicine*, vol. 219, p. 106741, 2022.

[25] W. Qi *et al.*, "A Smartphone-Based Adaptive Recognition and Real-Time Monitoring System for Human Activities," *IEEE Transactions on Human-Machine Systems*, vol. 50, pp. 414–423, Oct. 2020.

[26] J. Qi *et al.*, "Examining sensor-based physical activity recognition and monitoring for healthcare using Internet of Things: A systematic review," *Journal of Biomedical Informatics*, vol. 87, pp. 138–153, Nov. 2018.

[27] R. Z. U. Rehman *et al.*, "Selecting Clinically Relevant Gait Characteristics for Classification of Early Parkinson's Disease: A Comprehensive Machine Learning Approach," *Scientific Reports*, vol. 9, p. 17269, Nov. 2019.

[28] C. L. Pulliam, D. A. Heldman, E. B. Brokaw, T. O. Mera, Z. K. Mari, and M. A. Burack, "Continuous assessment of levodopa response in parkinson's disease using wearable motion sensors," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 159–164, 2017.

[29] H. Khodakarami, P. Farzanehfar, and M. Horne, "The use of data from the parkinson's kinetigraph to identify potential candidates for device assisted therapies," *Sensors*, vol. 19, no. 10, p. 2241, 2019.

[30] E. Rovini, C. Maremmani, A. Moschetti, D. Esposito, and F. Cavallo, "Comparative motor pre-clinical assessment in parkinson's disease using supervised machine learning approaches," *Annals of biomedical engineering*, vol. 46, pp. 2057–2068, 2018.

[31] A. H. Butt, E. Rovini, H. Fujita, C. Maremmani, and F. Cavallo, "Data-driven models for objective grading improvement of parkinson's disease," *Annals of Biomedical Engineering*, vol. 48, pp. 2976–2987, 2020.

[32] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, *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.

[33] M. M. Hoehn and M. D. Yahr, "Parkinsonism: onset, progression, and mortality," *Neurology*, vol. 17, no. 5, pp. 427–427, 1967.

[34] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1780–1786, 2014.

[35] J. Qi *et al.*, "Advanced internet of things for personalised healthcare systems: A survey," *Pervasive and Mobile Computing*, vol. 41, pp. 132–149, Oct. 2017.

[36] N. Bahador *et al.*, "Multimodal spatio-temporal-spectral fusion for deep learning applications in physiological time series processing: A case study in monitoring the depth of anesthesia," *Information Fusion*, vol. 73, pp. 125–143, Sept. 2021.

[37] D. Vos *et al.*, "Discriminating progressive supranuclear palsy from Parkinson's disease using wearable technology and machine learning," *Gait & Posture*, vol. 77, pp. 257–263, Mar. 2020.

[38] A. Abrami *et al.*, "Using an unbiased symbolic movement representation to characterize Parkinson's disease states," *Scientific Reports*, vol. 10, p. 7377, Apr. 2020.

[39] S. Bhat *et al.*, "Parkinson's disease: Cause factors, measurable indicators, and early diagnosis," *Computers in Biology and Medicine*, vol. 102, pp. 234–241, Nov. 2018.

[40] L. Tao, X. Wang, F. Nan, J. Qi, Y. Yang, and P. Yang, "Effective severity assessment of parkinson's disease with wearable intelligence using free-living environment data," in *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*, pp. 1–10, IEEE, 2023.

[41] A. Papadopoulos, D. Iakovakis, L. Klingelhoefer, S. Bostantjopoulou, K. R. Chaudhuri, K. Kyritsis, S. Hadjidimitriou, V. Charisis, L. J. Hadjileontiadis, and A. Delopoulos, "Unobtrusive detection of parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques," *Scientific Reports*, vol. 10, no. 1, p. 21370, 2020.

[42] M. Sundgren, M. Andréasson, P. Svenningsson, R.-M. Noori, and A. Johansson, "Does information from the parkinson kinetigraph™(pkg) influence the neurologist's treatment decisions?—an observational study in routine clinical care of people with parkinson's disease," *Journal of Personalized Medicine*, vol. 11, no. 6, p. 519, 2021.