Measurement 262 (2026) 120052

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Transfer learning for assessing Parkinson's disease: Analysis of wrist-worn sensors data and time-series imaging

Mohammed Hammoud [a], Aleksei Shcherbak [a], Ekaterina Bril [b], Maksim Semenov [b], Oleg Sergiyenko [c,*], Andrey Somov [a]

[a] Research Center for Digital Engineering and Innovation, Moscow, 143026, Russia
[b] Burnazyan Federal Medical and Biophysical Center, Moscow, 123098, Russia
[c] Autonomous University of Baja California, Mexicali, 21100, Mexico

ARTICLE INFO

Keywords:
Deep learning
Parkinson's disease
Time-series imaging
Transfer learning
Wearable sensors

ABSTRACT

Parkinson's disease (PD) is a debilitating condition that causes loss of physical activity, tremors, and coordination issues. Current studies use complex experimental testbeds to acquire data for PD diagnosis, and performance still requires improvement. This research aims to differentiate between healthy individuals and those in the early stages of PD using deep learning and a left-wrist-worn sensor with an embedded tri-axis Inertial Measurement Unit sensor. Data were collected from 40 patients while they performed nine daily exercises. The acceleration and gyroscope signals were preprocessed, segmented, and transformed into images using Time Series Imaging (TSI) algorithms, including Gramian Angular Difference/Summation Fields (GADF/GASF), Recurrence plots, Markov Transition Fields, and Continuous Wavelet Transform. We report on a PD assessment system based on the equal-weighted RGB images from accelerometer/gyroscope axes and transfer learning. GADF and GASF algorithms demonstrate superior performance, particularly with ResNet, achieving an average F1-score of 0.986 for HC vs. PwPD1 classification, with perfect detection (F1=1.0) in functional tasks like arm-holding and glass-filling. For PD staging (PwPD1 vs. PwPD1.5), the system maintained strong accuracy (F1=0.878), reaching F1=1.0 in dynamic exercises. Specifically, the key exercises including arm-holding and fist-clenching provided well-calibrated predictions Expected Calibration Error (ECE) $\leq$ 0.045, Brier Score $\leq$ 0.031) for both the detection and staging, ensuring high confidence in the model's outputs. Multiclass tasks yielded F1-scores in the range 0.90–0.91, confirming TSI's potential for scalable, clinical-grade PD monitoring.

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by the degeneration of dopamine-producing neurons, leading to motor symptoms like tremors, stiffness, and impaired coordination, as well as non-motor symptoms such as cognitive decline and depression [1]. It affects 1% of individuals over 60 [2], with prevalence projected to rise from 6.9 million in 2015 to 14.2 million by 2040 [3]. Although treatments exist, PD remains incurable[1] and requires continuous monitoring and medication adjustments. Hospitalization costs account for 69% of total healthcare expenses for PD [4]. Current assessments, such as the MDS-UPDRS, involve clinician-supervised motor activities but rely on subjective evaluations [5], with accuracies ranging from 73% to 84% [6]. PD symptoms persist for years, complicating diagnosis, and medication effectiveness diminishes over time, necessitating continuous adjustments.[2] These factors highlight the need for objective PD diagnosis [7] and/or PD stage

Existing PD detection methods face trade-offs between accuracy, invasiveness, and practicality. A multiscale hybrid attention network was developed to automatically distinguish between Patients with Parkinson's Disease (PwPD) and healthy controls (HC) by capturing complex features in brain images [8]. Two strategies were proposed to enhance performance: Parallel Network Classification (PNC) and Multislice Fusion Classification (MSFC). Both strategies improved performance, with PNC achieving the best results, reaching the highest F1-score of 94.17%. Neuroimaging techniques (e.g., Magnetic resonance imaging (MRI), positron emission tomography (PET)) and cerebrospinal fluid biomarkers offer high accuracy but are costly and invasive.

* Corresponding author.
E-mail address: srgnk@uabc.edu.mx (O. Sergiyenko).

[1] https://www.nhs.uk/conditions/parkinsons-disease/.
[2] http://pdcenter.neurology.ucsf.edu/patients-guide/parkinson%E2%80%99s-disease-medications.

https://doi.org/10.1016/j.measurement.2025.120052
Received 12 May 2025; Received in revised form 25 November 2025; Accepted 10 December 2025
Available online 12 December 2025

Voice-based approaches, and smartphone sensors [9] offer non-invasive alternatives but are sensitive to speaker characteristics and voice disorders. Additionally, muscle weakness affecting the vocal cords at advanced stages of Parkinson's disease (PD) further restricts the analysis of vocal anomalies. Similarly, home nocturnal breathing signals from a breathing belt were also analyzed, although other breath-related conditions may affect the accuracy of this method. Electroencephalogram (EEG) recordings have limitations due to their low sampling rate and challenges in determining the signal source.

Smartphones are widely used for PD detection due to their applicability in real-life situations [9,10]. Studies have utilized accelerometers [10] and gyroscopes [4]. One study analyzed gait data with a smartphone's sensors, highlighting the importance of lifestyle and gyroscope sway features in distinguishing classifications of shaking and tremor. While the study did not use magnetometers, researchers also employed smartwatches with accelerometers [11] to differentiate PwPD from other neurological diseases and HC. Innovations, such as an inertial spoon prototype, were introduced to classify PwPD and assess PD levels [12]. The study indicated that using both gyroscopes and accelerometers improves classification accuracy, with the gyroscope being more effective in detecting PD symptoms due to its ability to detect more noticeable angular deflections. Additionally, a study on Archimedes' spiral drawing using a smart ink pen found that PwPD showed reduced smoothness and less consistent force on the writing surface.

Video recording is a non-contact PD diagnostic technology. A camera-based approach is cost-effective and addresses the typical limitations of the Inertial Measurement Unit (IMU). Researchers used near-infrared recording (NIR) eye video recording and Deep learning (DL)-based pupil segmentation to detect PD and Progressive Supranuclear Palsy (PSP) during five exercises in a clinical setting. Eye movement diagnosis typically requires a complex setup found in clinics [13]. A recent study [14] employed a Convolutional Neural Network (CNN) to classify PD severity using video data from seven tasks, including finger, hand, postural, and leg movements. Other studies have also employed various exercises, including hand tasks, walking, and standing. Although video-based detection is convenient, it is complex and algorithm-dependent.

The state-of-the-art research targets multimodal-based detection [15] to obtain the highest possible performance. It has been found that integrating various domains and modalities has the potential to improve accuracy and enhance robustness [16–18]. Researchers used smartphone acceleration data and touchscreen typing to classify PwPD/HC and Tremor/PwPD/fine motor impairment (FMI) [9]. Similarly, voice sensors and smartphone accelerometers were attached to both hands to achieve two and five PD severity classifications [19]. The use of wearable sensors [20,21] presents a trade-off between accuracy and feasibility in detecting and classifying PD. The more accurate a multimodal approach is, the more complex the required setup becomes. Video recording cannot provide 24-hour monitoring. Moreover, 24-hour video recording in private houses is not preferred due to safety issues and social acceptance. Most currently available solutions do not track the progress of PD.

This study introduces an intelligent framework for PD detection and severity assessment using a single wrist-worn sensor and Time series imaging (TSI) algorithms. Unlike prior work, our approach transforms raw accelerometer and gyroscope data into images, e.g. Gramian Angular Fields, Recurrence Plots, Continuous Wavelet Transform (CWT), and leverages transfer learning with pre-trained convolutional neural networks (CNNs) to achieve the state-of-the-art performance.

Novelty of this research is: (1) Simplicity and Accessibility: A single-sensor setup reduces a patient burden while maintaining diagnostic accuracy, outperforming multi-sensor systems [17]; (2) TSI-Enhanced Feature Extraction: Converting time-series data into images helps to capture the spatiotemporal patterns of PD symptoms, e.g., tremor and bradykinesia, more effectively than traditional signal processing methods [22]; (3) Comprehensive Exercise Analysis: We identify the optimal daily activities, e.g. glass-filling tasks, for PD detection, achieving perfect classification (F1-score is 1.0) in the early-stage of PD. (4) We systematically investigate how segmentation parameters (window length and overlap) must be adapted to the PD symptom characteristics. This parameter optimization validated across nine clinically-relevant exercises demonstrates that task-specific segmentation is critical for maximizing the performance while maintaining computational efficiency in continuous monitoring scenarios.

The contribution of this research can be summarized as follows:

- Building a multi-input DL model for PD detection and its severity assessment, using time-series imaging (TSI) algorithms and transfer learning.
- Investigation of the impact of different TSI and model architectures on the diagnostic performance.
- Exploration of the relationship between the performance of PD diagnosis and various activities of daily living.
- Investigation of the critical relationship between the segmentation parameters and PD symptom characteristics.

The article is structured as follows: Section 2 reviews relevant research works. Section 3 describes the proposed methodology, implementation details, and evaluation metrics. Section 4 presents the results of the experiments and their discussion. Finally, concluding remarks are provided in Section 5.

2. Related work

2.1. Sensor data analysis

A gyroscope in a smartphone was used to investigate the impact of Sampling Frequency (SF) [23]. The results show that to accurately detect tremors and bradykinesia (BR)(slowness of movement), a minimum frequency of 30 Hz is required using data from either the accelerometer or the gyroscope. The study also found that detecting tremors requires more complex features, such as features related to entropy, than detecting BR. Interestingly, the type of device used did not significantly impact tremor detection, indicating that using just one accelerometer is sufficient for detecting PD. However, using both the accelerometer and the gyroscope improves the accuracy of BR detection. Another study has proposed a Machine learning (ML) [24] based system that uses six IMU sensors to rate PwPD [25]. These sensors were placed on the trunk, wrists, feet, and lumbar spine to capture data during two exercises: walking for two minutes and standing still with closed eyes for 30 s, to measure postural sway.

The effectiveness of using multiple sensors attached to different body parts was analyzed, specifically the wrist and left tibia, to determine the impact of using an accelerometer and gyroscope from the sensor [26]. The results showed that using the wrist sensor, the support vector machine (SVM) produced an accuracy of 88% and 85.71% in complex and minimal configurations, respectively. On the other hand, the IMU sensor from the left tibia achieved accuracies of 84.11% and 85.13% for SVM and K-Nearest Neighbor (KNN), respectively. The study also compared the performance of acceleration attached to the lateral left tibia with that of other sensors, such as EEG and wrist-worn IMU. It was concluded that the impact of using different sensors was limited.

Using multiple sensors is considered complex and requires too much effort from the patient. Additionally, it may be inconvenient for patients. So, the most effective data for detecting tremors in patients with PD was investigated [27]. The authors discovered that using a single wearable sensor placed on the hand dorsum was sufficient for both

BR and tremor detection in the upper extremities. They concluded that having sensors on both sides did not significantly improve the detection performance.

Researchers conducted a study on the classification of HC/PwPD by collecting data using 11 exercises from an IMU sensor attached to the wrists [28]. They employed short-time Fourier-transform (STFT)-based 1D-CNN and investigated STFT configurations, such as window lengths, number of output points, and frequency ranges. They concluded that the STFT window must be at least 2 s long, and the frequency range must include frequencies below 3 Hz. Their proposed approach outperformed the state-of-the-art in 70% of cases. Contrary, another study utilized ML techniques along with IMU sensors attached to the right dorsum to differentiate between HC and PwPD patients in early stages (PwPD1 and PwPD2) [29]. The study explored the significance of various features and the impact of exercises on the diagnosis. The researchers identified the three most effective exercises for efficient PD diagnosis, achieving an Area under the Receiver-Operating Characteristics (ROC) of 0.9 for each classification task.

Similar work proposed an ML-based model for HC/PwPD1 and HC/PwPD2 classification using four IMU sensors attached to the wrist and dorsum of the hands [30]. They used 11 exercises and extracted time, correlation, and Discrete Fourier transform (DFT) features. Compared to the studies [31,32], they achieved the best f1-micro of 0.78 and 0.88 for HC/PwPD1 and HC/PwPD2, respectively. The performance of wearable wireless sensors, handwriting data, and video recording was analyzed and compared [33] to monitor the stage of PD. The best performance was obtained from sensor data with an F1-score of 0.93.

2.2. Time series classification

A time-series signal is a sequential set of measurements that occur in a natural temporal sequence. Different kinds of time series data comprise biomedical signals like EEG and electrocardiogram (ECG), financial data, industrial signals from sensors, weather data, video, and barometric signals such as voice and gestures.

Time series classification (TSC) approaches can be divided into two categories: time domain and frequency domain [34]. Time domain methods include auto-correlation, auto-regression, and cross-correlation, while frequency domain methods involve spectral and wavelet analysis. Regarding classification, TSC approaches can be instance-based or feature-based. Instance-based methods assess the similarity between training and test data and assign a label to the most similar class (e.g., KNN with k = 1, Dynamic time wrapping (DTW)). On the other hand, feature-based methods alter the signal to extract more distinct and representative features in a new space.

Long Short-Term Memory (LSTM) is a commonly used DL architecture for TSC. As a type of Recurrent neural network (RNN), LSTM addresses issues such as vanishing gradients, but training requires a long time. In the medical field, LSTM has been applied to classify hand movements, brain decisions, and stress based on EEG signals [35,36]. It has also been used to classify heartbeat sound signals [37] and speech modes [38]. In wireless communication, LSTM has been utilized to classify modulation signals.

A recent approach in TSC is TSI. TSI is a visual depiction of a time-series signal [39]. Essentially, these signals are interpreted as a task of recognizing texture images. This conversion aims to apply DL techniques to image classification algorithms. TSI makes use of various algorithms, such as Gramian Angular Summation Fields (GASF), Gramian Angular Difference Fields (GADF), Markov Transition Fields (MTF), and Recurrence plots (RP). Gramian Angular Fields (GAF) finds the temporal correlation between pairs of time series values. It has two types: summation (GASF [39]) and differential (GADF [39]). MTF represents a field of transition probabilities for a discretized time series, while RP is obtained from pairwise Euclidean distances for each time series value.

The research examined how well RP combined with CNN performed in the context of TSC [34]. This method surpassed other deep architectures and even achieved the state of the art in TSC. Researchers used TSI algorithms, including GASF, GADF, and MTF, to convert extracted features from the eye's pupil into images [40]. The study concluded that the window length selection for these algorithms depends on processing power and timing constraints. In a similar study [41], EEG signals were transformed into images using GASF for epilepsy diagnosis, and a CNN-based approach achieved an F1-score of 0.90. Another study utilized GASF, GADF, and MTF to transform Multivariate series (MTS) signals into images [42]. Researchers reduced image dimensions to lower complexity and combined 2D images into colored images. The impact of different architectures, particularly ConvNet and VGG16, on TSI algorithms was also explored, with both architectures yielding similar results. In another study, multivariate time series sensor data were classified by converting them into images, using GADF, GASF, and MTF [43]. These images were combined into one large image and then input into a CNN. The researchers found that the algorithm and concatenation sequence did not significantly impact the results. Despite the simplicity of the conversion method, it proved to be effective for classification, yielding results comparable to VGG16. Another study suggested a rolling bearing fault diagnosis model using MTF and ResNet, which outperforms state-of-the-art methods in average accuracy [44]. Researchers differentiated between PwPD, HC, and patients with PSP by examining NIR eye video recordings using five different exercises [13]. They initially utilized DL-based segmentation to extract pupil features, including coordinates and minor and major axes. Afterwards, the time-series signals of these features were converted into images and fed into a CNN-based model.

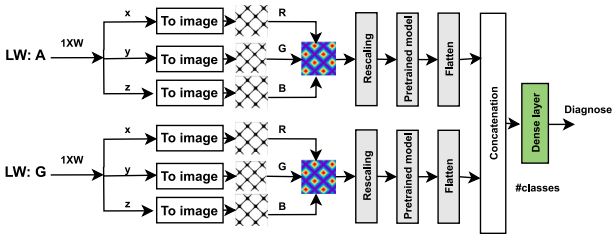
This work offers a non-invasive and simple method for detecting PD and assessing its severity. Compared to other research, our approach is straightforward and can be used in home and clinic settings, allowing for 24-hour monitoring of PD progression. Instead of working with time-series data, we focused on image processing using TSI. These approaches have not been investigated in wearable sensors based on detecting PD and its severity assessment. Images visually represent time-series data, making understanding and interpreting patterns, trends, and anomalies easier. They can capture both local and global patterns in the time series data. Using CNN, we can perform tasks such as classification, regression, or anomaly detection. CNNs can automatically learn and extract meaningful features from the data, improving the accuracy and performance of neurological disease detection models. Images can capture spatial and temporal patterns simultaneously, which is particularly crucial in neurological disease detection, as specific patterns or abnormalities may only become apparent when considering the interactions between different variables over time. Lastly, since medical data is costly and limited, using an image-based approach has opened the opportunity to apply transfer learning via pre-trained models on famous image datasets, such as ImageNet. Transfer learning and TSI have not previously been utilized in wearable sensor-based assessment of PD progression. In our analysis, we examined various TSI algorithms, model architecture, and daily living exercises.

3. Methodology

3.1. Data collection

The data collection process, overseen by neurologists, was conducted under ethical committee approval (No. 1-9-21). Informed consent was obtained from all subjects involved in the study. A total of 40 participants participated in the study, consisting of 18 HC individuals and 22 PwPD. The dataset includes 24 males and 16 females. The disease was classified into five stages using the Hoehn-Yah scale [45], as listed in Table 1, with the following number of patients in each stage: 5 in stage 1, 3 in stage 1.5, and 14 in stage 2.0. Patients with unclear

(a) Experimental setup



(b) DL-based approach for PD detection using left wrist sensors, LW, A, and G stand for left wrist sensor, accelerator, and gyroscope, respectively.

Fig. 1. System overview.

Table 1
Description of the patients' dataset.

Group	Hoehn-Yahr [45]	Total subjects	Female		Male	
			Number	Mean age ± Std	Number	Mean age ± Std
HC		18	5	56.0 ± 2.3	13	64.8 ± 5.9
PD	1	5	3	64.7 ± 3.8	2	62.5 ± 7.8
	1.5	3	2	54 ± 1.4	1	62.0 ± 0
	2	14	6	65.3 ± 8.3	8	68.4 ± 9.5
	–	40	16	–	24	–

diagnoses underwent additional assessments by 3 to 5 neurologists to ensure accurate diagnoses. The Hoehn and Yahr stage 1.5 represents a clinically significant transitional phase in PD progression, characterized by unilateral motor symptoms with emerging axial involvement. This intermediate stage bridges the gap between stage 1 (pure unilateral manifestations) and stage 2 (bilateral symptoms without balance impairment), where patients begin to exhibit subtle midline signs, such as mild axial rigidity, reduced arm swing, or slight postural changes, that do not yet meet the full bilateral criteria. This transitional nature explains the frequent underrepresentation of PD 1.5 cases in research cohorts, as many patients progress rapidly to stage 2 before the study enrollment or receive conflicting classifications during the follow-up assessments.

The data collection process involved nine exercises. Exercise numbers *'2'*, *'3'*, *'5'*, *'8'*, and *'9'* were adopted from the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS). In addition, exercises *'1'* and *'7'* were modified from the MDS-UPDRS. Exercise *'6'* was proposed directly by the neurologists, while exercise *'4'* was included based on relevant research [27]. The nine exercises are shown below in Table 2, along with their ID numbers and their durations.

Patients' data were collected anonymously with a unique identifier. One IMU sensors were attached to the left-hand wrist. We used Sensortile[3] to collect sensor data. It features a built-in 3-axis accelerometer, 3-axis gyroscope, and 3-axis magnetometer. We used a SF of 416 Hz for synchronization purposes, which is determined by the hardware abilities of the IMU and satisfies the Nyquist theorem. To overcome signal attenuation caused by the human body absorbing the 2.4 GHz Bluetooth Low Energy (BLE) signals, we developed a synchronization model and an Android application. The smartphone acts as a wireless sensor node gateway, transferring data to a cloud facility. We used a sensitivity level of ±2g and a range of ±250 degrees per second (dps). Sensor measurements were saved to an SD Card in .csv format, with a timestamp generated from the Real-time clock in the microcontroller, which utilized an external oscillator of 32,768 Hz. The experimental setup is illustrated in Fig. 1(a).

It is important to note that the duration of exercises varied. Additionally, some patients did not have exercises or sensor data, so the number of patients differed based on the exercises and sensors used. We utilized both acceleration and gyroscope data from the sensor. Our work distinguishes healthy individuals from PwPD, including stages 1, 1.5, and 2. The more advanced stages are not included since it is easy to detect them visually by doctors, which is quite complex in the early stages of PD.

3.2. Sensors' data preprocessing

Sensors in this study were sampled at 416 Hz. However, to meet the minimum sampling frequency requirement of 30 Hz [23], we downsampled the signals to 32 Hz to reduce the required computational power. A band-pass filter with the range of *(0.2, 13)* Hz and the order of 6. Frequencies below 0.2 Hz were removed to eliminate the gravity effect, limb orientation [29], and unnecessary frequencies. It is worth noting that filtering was applied before downsampling to prevent aliasing. Research on the PD tremor has identified key frequency ranges: classical rest tremor (3–7 Hz), isolated postural tremor (4–9 Hz), and kinetic tremor during the slow movement (7–12 Hz) [47]. Multiple studies support the use of 32 Hz sampling rate for performing the analysis of the PD movement. Similarly, the tremor frequencies of the upper extremities are lower than 13 Hz [48]. The study [23] demonstrated that the sampling rate of 30 Hz is sufficient for detecting tremors

[3] https://www.st.com/en/evaluation-tools/steval-stlkt01v1.html.

Table 2
Description of nine motor exercises, their clinical alignment with MDS-UPDRS items [46], and sample size per diagnosis group.

Exercise ID	Description	Corresponding MDS-UPDRS items	Primary motor constructs assessed	HC	PwPD 1.0	PwPD 1.5	PwPD 2.0
Ex1	Stand up from a chair, walk there and back, and sit down.	3.9 Arising from Chair, 3.10 Gait	Akinesia (Initiation), Gait, Balance	18	5	3	14
Ex2	Sit down, drop arms, and rest.	3.17 Rest Tremor Amplitude	Rest Tremor	18	5	3	14
Ex3	Bend arms, open and close thumb and index fingers while holding the rest closed to the palm.	3.4 Finger Tapping	Bradykinesia, Rhythm, Amplitude	18	5	3	14
Ex4	Straighten arms and alternately touch nose with index fingers.	3.16 Kinetic tremor of the hands	Bradykinesia, Coordination	18	5	3	12
Ex5	Hold the outstretched arms position.	3.15 Postural Tremor of Hands	Postural Tremor	18	4	3	13
Ex6	Fill a glass of water from the bottle, bring it to the mouth, hold the position for 3 s, and put it back.	3.6 Pronation-Supination 3.15 Postural Tremor of the Hands 3.16 Kinetic Tremor of Hands	Bradykinesia, Coordination	18	5	3	13
Ex7	Clench and unclench fists on the table.	3.5 Hand Movements	Bradykinesia, Fatigue, Rhythm	18	3	3	14
Ex8	Stand, align arms at breast height, palms facing inward.	3.14 Body Bradykinesia, 3.15 Postural Tremor of Hands	Axial Stability, Postural Tremor	18	4	3	13
Ex9	Stand, align arms at breast height, palms facing forward.	3.14 Body Bradykinesia, 3.15 Postural Tremor of Hands	Axial Stability, Postural Tremor	18	5	3	12

and bradykinesia, with higher sampling rates providing no significant improvement. Similarly, another study [49] further confirmed that PD tremor primarily occurs within the 4–6 Hz range, ensuring that 32 Hz adequately captures these key frequencies. In the study [50], it was found that a sampling rate ≥30 Hz was required to detect the tremor using acceleration. The conclusion aligns with the Nyquist theorem, as 30 Hz exceeds twice the highest tremor frequency of interest (13 Hz), preventing aliasing and preserving signal integrity. It comfortably encompasses the full frequency range of various PD tremor and bradykinesia while preventing aliasing artifacts. This confirms that the frequency of 32 Hz is sufficient for the reliable detection and analysis of PD-related motor symptoms [51].

After preprocessing, we removed 2 s from both the signal ends and segmented the data into overlapping windows. Window size and overlap selection balances frequency resolution, temporal sensitivity, and real-time requirements. The 4–13 Hz tremor range necessitates windows long enough to capture multiple cycles while minimizing latency for closed-loop applications like adaptive DBS [31,52]. Oversized windows may obscure brief symptoms, whereas undersized windows reduce spectral accuracy [30]. High overlap ratios (75%–90%) prevent missed transient detections [33], with prior studies using: 4s windows/1s steps [31,52] , 3s windows/1.5s overlap [30,31], 3s windows/2s overlap [53], 5s windows/2.5s overlap [33], 4s windows/2s overlap [54], and 2s windows/1s overlap [55].

Our 4s window/1s overlap captures 16–52 tremor cycles (4–13 Hz), providing spectral resolution for resting (3–7 Hz, 12–28 cycles) and postural tremor (4–9 Hz, 16–36 cycles) while aligning with MDS-UPDRS protocols [56]. This yields optimal 128 × 128 pixel time-frequency representations for CNN processing [54,55]. Ablation studies varied these parameters, with reported metrics (ROC/PR/training-validation) reflecting per-exercise optima.

3.3. Time-series imaging algorithms for PD signal analysis

In this section, we examine three time-series imaging (TSI) methods — Recurrence Plots (RP), Gramian Angular Fields (GASF/GADF), and Markov Transition Fields (MTF) — for analyzing the motor symptoms of Parkinson's disease (PD) using wearable sensor data. Each technique captures distinct non-linear and non-stationary movement patterns, enabling robust PD detection and severity assessment.

Recurrence Plots (RP) reconstruct phase-space dynamics by computing pairwise Euclidean distances between the time-series points, generating a recurrence matrix that reveals periodic and irregular movement states. RP detects tremor (4–6 Hz oscillations) via the diagonal lines, bradykinesia through the recurrence density (reflecting movement slowness), and dyskinesia or freezing of gait via the scattered points. Key advantage is its noise robustness achieved through the distance thresholding, while preserving the pathological patterns. RP also tracks symptom progression, such as motor decline, though its performance depends on the parameter selection and computational load. When combined with CNNs, RP enhances PD severity classification by providing the interpretable visual representations of symptom dynamics.

Gramian Angular Fields (GASF and GADF) transform the sensor data into polar coordinates, preserving temporal and amplitude relationships critical for PD assessment. GASF emphasizes the structural correlations, making it well-suited for rigidity and bradykinesia analysis, while GADF enhances sensitivity to subtle the movement deviations, aiding early PD detection. Though GASF may attenuate high-frequency tremor details, it excels in tracking progressive motor decline, such as reduced arm swing. Together, GASF and GADF offer the complementary insights — GASF for amplitude stability and GADF for dynamic movement variations — improving the comprehensive symptom evaluation.

Markov Transition Fields (MTF) discretize time series into the quantile bins, constructing a transition probability matrix that captures abrupt state changes in movement patterns. This method is particularly effective in detecting freezing of gait and quantifying symptom variability over time. MTF's probabilistic framework complements RP's phase-space analysis and GASF/GADF's amplitude-angle representations, providing an additional layer of dynamic assessment.

In summary, RP is ideal for analyzing non-periodic tremor and symptom fluctuations, GASF/GADF better characterize rigidity and progressive decline, and MTF offers a probabilistic perspective on symptom transitions. Together, these TSI algorithms form a multi-faceted approach to PD motor assessment, enhancing the early detection, severity classification, and long-term progression monitoring.

3.4. Deep learning

The segmented frames were transformed into images and then fed to the proposed DL model. Fig. 2 shows the output of these algorithms applied to data from patients diagnosed with PwPD1. The data were divided into training and testing datasets based on subject ID, ensuring

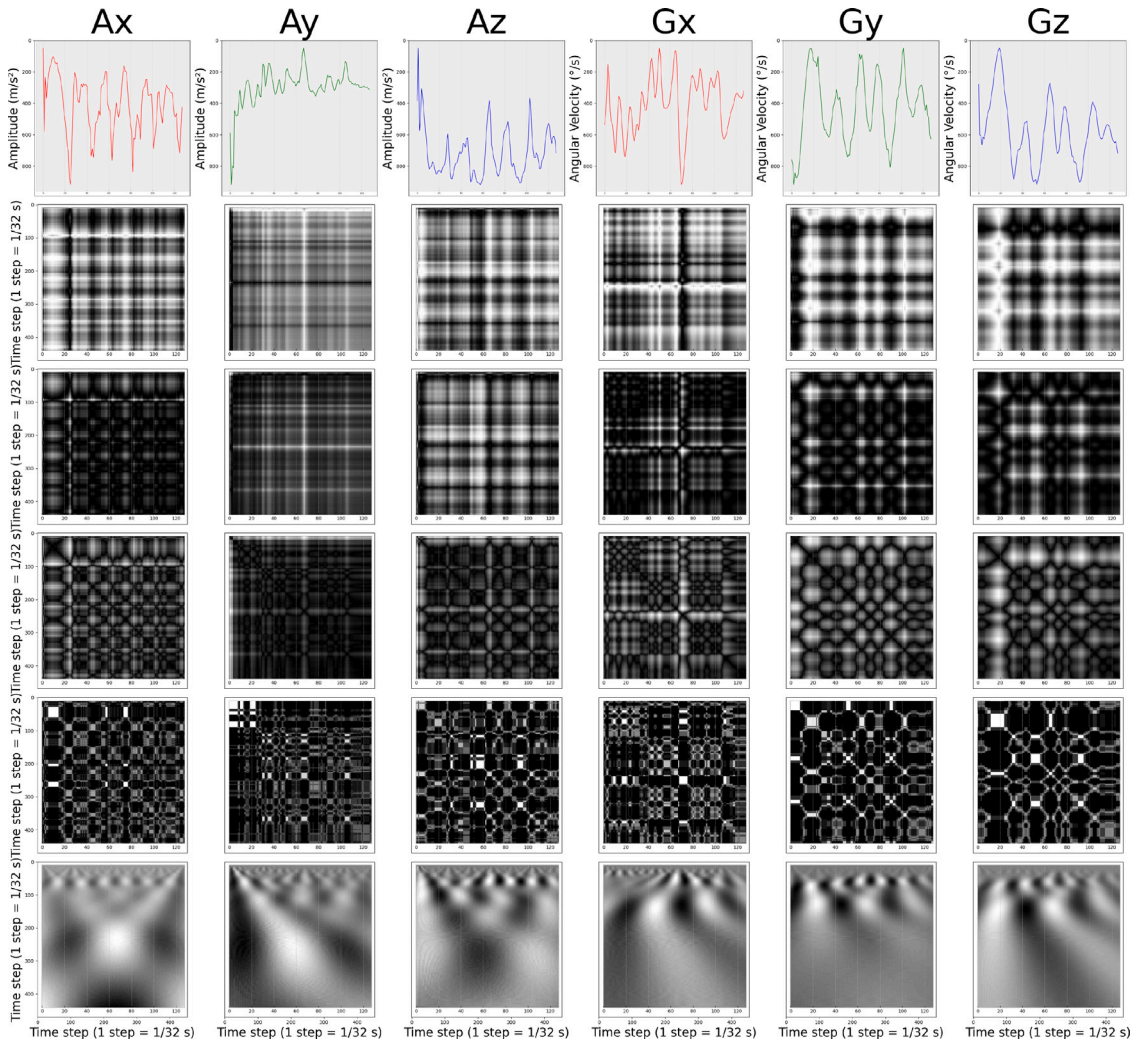


Fig. 2. Different TSI algorithms output (GADF, GASF, MTF, RP, and CWT) for acceleration and gyroscope data from patient # 42 diagnosed with PD1.

no overlap. There are nine different models, each specifically designed for a particular task. Our architecture comprises a multi-input and single-output design, illustrated in Fig. 1(b). The inputs comprise data from the accelerometer and gyroscope. The time series data from each accelerometer and gyroscope were first scaled to the range of *(−1,1)* and then transformed into images using TSI algorithms, such as GADF, GASF, MTF, RP, and Continuous Wavelet Transform (CWT) [57]. The resulting images from each sensor's axes (x, y, and z) were combined into one RGB image and fed into our model. For a patient's L-second exercise duration, the segmentation yields 'n' frames, each containing '6' time series signals corresponding to the acceleration axes (x, y, and z) and gyroscope axes (x, y, and z). As discussed earlier, applying TSI algorithms on 'n' frames results in 'n' × (6/3) × (W × SF) × (W × SF) × 3. If we have five frames and a SF of 32 Hz, we get 5 × 128 × 128 × 3.

We fused accelerometer and gyroscope data into RGB images, assigning equal weight to all three axes (x, y, z) for each sensor modality. The *x*-axis values from the accelerometer are mapped to the red channel, the *y*-axis to green, and the *z*-axis to blue, with the gyroscope data following the same assignment in separate images. This equal weighting was chosen for several reasons: PD symptoms like tremors and bradykinesia often manifest similarly across all three axes, and it avoids bias that could overlook clinically significant symptoms that may appear predominantly in specific axes, such as *z*-axis tremors during pronation/supination tasks.

Time-series signals were scaled to *(−1,1)*, transformed into images, and input to pre-trained models with various architectures such as VGG16 [58], Inception [59], MobileNet [60], and ResNet [61]. The output from each pathway is flattened, combined, and directed to a dense layer to carry out the classification task.

3.5. Experimental protocol

LOSO cross-validation was used to validate the model's generality. For instance, in the LOSO method, only one subject is retained for testing while the rest are used for training. After training, the model is tested on the testing patient using all of his own data frames, resulting in multiple outputs based on the number of frames. We obtained the final classification results using a majority-based voting technique. The voting technique is used to obtain results based on subjects. Another reason is that frame-based detection is not reliable, as the severity of PD symptoms may fluctuate over time. For the subject-level classification, window-level probabilities are thresholded at 0.5, the standard choice for balanced binary problems; each window is assigned to the positive

class if threshold ≥0.5, and the final subject label is obtained via the majority voting across all the windows belonging to that subject.

All reported results are based on testing frames for patients not used during training. This process is repeated *'n'* times, where *'n'* is the number of patients. Although LOSO is time-consuming, it is suitable for small datasets.

To demonstrate the model's robustness, we present ROC curves for two binary classification tasks: HC vs. PwPD1 and PwPD1 vs. PwPD1.5. Frame-based ROC curves, derived from all testing subject frames (using LOSO cross-validation), offer a comprehensive evaluation of the model's performance. This approach highlights the model's ability to generalize across varying granularities. It ensures validation not only at the subject level (via the voting approach) but also at the frame level (before applying the voting approach), thereby capturing potential fluctuations in symptom severity over time.

In all experiments, the leave-one-subject-out (LOSO) protocol is implemented at the subject ID level. For each fold, all frames/windows belonging to the held-out subject (across all exercises) are excluded from training, excluded from validation and hyper-parameter selection, and used only once for final testing in that fold. This ensures no information from the test subject influenced the preprocessing or model design. No data augmentation was applied.

During training, we employed class weights to mitigate the high imbalance in the dataset. Moreover, the models were evaluated using the F1-micro and specificity. All metrics are micro-averages. We conducted various experiments using different exercises, different TSI, and pre-trained models from different architectures. The deep learning models were implemented using various architectures (VGG16, InceptionV3, MobileNetV2, ResNet50V2) with their top layers disabled, trained on multiple TSI representations, including GADF, GASF, MTF, RP, and CWT. Models were optimized using Rectified Adam with the learning rate 0.001 and trained for 50 epochs with the batch size 64. The input data consisted of 4 s frames with the 1 s overlap, using the classification threshold 0.5. Hyperparameter tuning for the architecture was not applied in this work. The network was implemented using Python 3.10.0 as a programming language, TensorFlow, and Keras framework. PyTs[4] library was utilized to apply TSI.

Our primary results are achieved using 4 s window with 1 s overlap. However, we conducted a comprehensive ablation study examining the multiple factors including TSI methods, model architectures, exercises, sampling rates, sensor axis weighting schemes, and segmentation parameters.

The sensor data were segmented using sliding windows with various lengths and overlap configurations. We systematically evaluated the following window/overlap parameter combinations for each exercise: (5s, 3s), (5s, 2s), (5s, 1s), (4s, 3s), (4s, 2s), (4s, 1s), (3s, 2s), and (3s, 1s).

These parameters were determined through an exhaustive offline grid search conducted before the main LOSO experiments. The optimal configuration for each exercise was selected based on the preliminary performance analysis and then frozen for all the subsequent experiments. Crucially, this approach ensured that parameter selection was completely decoupled from the LOSO evaluation, preventing data leakage from the test subjects from influencing the segmentation strategy and ensuring no test data contamination during the parameter selection.

Model calibration evaluates the agreement between the predicted probabilities and observed outcomes, where a perfectly calibrated model has predicted the probabilities $X\%$ that correspond to $X\%$ empirical event rates [62]. We assessed the calibration using three metrics. The Expected Calibration Error (ECE) quantifies the weighted average of absolute differences between the accuracy and confidence across the probability bins. The Brier Score (BS) measures the mean squared probability error, with the lower values indicating better forecasts. To address ECE's sensitivity to binning, we used 5 bins for the subject-level predictions and 10 bins for the frame-level predictions, following recommendations to mitigate the binning bias [63].

[4] https://pyts.readthedocs.io/en/latest/index.html.

4. Results

In our study, we conducted a series of experiments to thoroughly examine the impact of different daily living exercises on overall performance. We implemented cutting-edge transfer learning and TSI algorithms, along with various architectural approaches to ensure comprehensive analysis. The severity assessment of PD was successfully achieved using binary and multi-class classifiers. It is worth noting that the results reported in this section are based on the segmentation parameters: a window length 4 s and 1 s overlap. All results reported in the Appendix are based on the best-found segmentation parameters from the ablation study, in Section 4.3.4.

4.1. Models performance

As mentioned, LOSO cross-validation was used to obtain multiple outputs for the tested patients. These outputs represent predictions for individual frames. However, what is important is the final prediction for the subject rather than the predictions for individual frames. To obtain the final prediction, we used a majority-based voting technique. All results in the results section are subject-based, which results from applying the majority-based voting technique.

The average testing results (F1-micro, and specificity micro) for all subjects during LOSO validation are listed in Tables A.10 and A.11 for four tasks, including binary and multi-classification. Table 3 summarizes the best-performing model architectures for each TSI algorithm, corresponding to each exercise for the targeted classification tasks.

Table 3 lists the best results for each task across TSI algorithms.. Both frame-based and subject-based results were included. Table 5 lists the metrics of the best-performing models, including the optimal model architecture, exercises, and TSI algorithm. It provides a comprehensive overview of the top configurations for both tasks: HC vs. PwPD1 and PwPD1 vs. PwPD1.5.

4.2. Discussion

Our initial findings, based on 4s-window size and 1s overlap, consistently demonstrate the superior performance of ResNet over other models in a range of exercises and TSI in most of our experiments. Furthermore, the GADF and GASF algorithms emerged as top performers, leveraging an angular perspective to analyze the trigonometric addition and subtraction between each data point, thereby facilitating the identification of temporal correlations across different time intervals.

The average F1-micro scores, a key metric in our research, are as follows: 1.0 for HC/PwPD1, 0.98 for HC/PwPD1.5, 0.9 for PwPD1/PwPD1.5, and 0.91 for multi-class classification. These scores indicate that our approach effectively distinguishes healthy individuals from early-stage PD. Notably, the performance in differentiating HC from PwPD1.5 is slightly lower than in HC/PwPD1, a finding that warrants further investigation.

PwPD1.5 represents an intermediate stage between stage 1 and 2. It was challenging for doctors to classify it as either stage 1 or stage 2. Our approach performed well because both PwPD1 and PwPD1.5 are in the early stages. The similarity in symptoms between PD stages 1 and 1.5 is reflected in the F1-micro score (0.9) for their classification. PD's Severity assessment can be achieved by applying a single classifier or multiple binary classifiers. However, using multiple binary classifiers requires training numerous models, which can be time-consuming.

Table 3
The best-performing model architectures for each TSI algorithm, corresponding to each exercise for the targeted classification.

	TSI/Exercise ID	1	2	3	4	5	6	7	8	9	Avg
Subject-based											
HC/PwPD1	GADF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	GASF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	MTF	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.86	0.96	0.97
	RP	1.00	0.78	1.00	0.91	0.91	1.00	0.95	0.91	1.00	0.94
	CWT	1.00	0.96	1.00	1.00	0.91	1.00	0.95	1.00	1.00	0.98
HC/PwPD1.5	GADF	1.00	1.00	1.00	0.95	1.00	1.00	0.85	1.00	1.00	0.98
	GASF	1.00	0.95	0.95	1.00	0.95	1.00	0.85	0.95	1.00	0.96
	MTF	1.00	0.86	1.00	1.00	0.95	1.00	0.90	0.90	0.95	0.95
	RP	1.00	0.90	0.95	0.90	0.86	1.00	0.95	0.90	0.90	0.93
	CWT	1.00	0.95	1.00	0.86	0.95	1.00	0.90	0.95	1.00	0.96
PwPD1/PwPD1.5	GADF	1.00	1.00	0.71	1.00	1.00	0.86	1.00	0.50	1.00	0.90
	GASF	1.00	0.71	0.71	1.00	0.83	0.86	0.83	0.67	0.86	0.83
	MTF	1.00	0.57	0.86	1.00	1.00	0.86	1.00	0.67	0.71	0.85
	RP	1.00	0.71	0.86	1.00	0.67	0.57	0.83	0.67	0.71	0.78
	CWT	1.00	0.71	0.71	0.86	0.83	0.86	0.67	0.67	0.71	0.78
Multi	GADF	0.98	0.85	0.90	1.00	0.72	0.98	0.84	0.95	0.92	0.90
	GASF	0.98	0.90	0.93	1.00	0.79	0.98	0.84	0.89	0.84	0.91
	MTF	0.98	0.88	0.93	0.90	0.82	0.98	0.89	0.89	0.82	0.90
	RP	0.98	0.85	0.81	0.70	0.72	0.80	0.84	0.73	0.89	0.81
	CWT	0.98	0.90	0.88	0.98	0.82	0.93	0.86	0.76	0.82	0.88

4.3. Ablation and sensitivity analysis

The ablation study is a systematic process of removing or modifying specific components within a model or methodology to evaluate their individual contributions to overall performance. In the context of this work, conducting an ablation study enables us to identify the significance of various elements, such as the selection of TSI algorithms, deep learning architectures, sensor data modalities, e.g., accelerometer vs. gyroscope, specific exercises, and preprocessing techniques, e.g., window size, segmentation parameters, and sampling frequency. By isolating and analyzing these components, we can better understand their impact on the model's effectiveness and optimize the system for accurate PD assessment.

4.3.1. The impact of TSI algorithms, deep learning architectures, and exercises

In this work, we applied only one TSI algorithm at a time to transform time-series data into images, which allowed us to determine which algorithm is more efficient and effective for the task. However, we did not explore the potential of combining multiple TSI algorithms simultaneously. This approach is planned for our future work, where we aim to harness the unique strengths and distinguishable patterns of each TSI algorithm. By integrating the complementary TSI techniques, we anticipate capturing a richer representation of the time-series data, ultimately enhancing the model's performance and robustness in detecting PD.

The performance results of different TSI algorithms for both tasks HC vs. PwPD1 and PwPD1 vs. PwPD1.5 are listed in Tables A.11 and A.10. As noted, the GADF performed the best across various tasks and exercises.

We also defined the exercises that are the most informative and determine whether some exercises can be excluded without significantly degrading the performance, as listed in Tables A.11 and A.10. In the current work, the model is built using only one exercise at a time as the input, simplifying the diagnosis procedure. No combinations of exercises were used simultaneously. This approach will be explored in future work, where the model will incorporate multiple exercises together to potentially enhance the performance. However, this may reduce the convenience for both the patients and doctors, as it would require more complex data collection, analysis processes, and more complex models. Here is the **detailed analysis** of the performance for two tasks, HC vs. PwPD1 and PwPD1 vs. PD1.5, using various exercises, various TSI, with various deep learning architectures:

HC vs. PwPD: *Ex.6* stands out, consistently achieving F1-score of 1 across various architectures like InceptionV3, MobileNetV2, and ResNet50V2. It excels by capturing the fine motor control, stability, and coordination—critical for distinguishing HC from PwPD1. Its structured nature provides clear signals for analysis. Similarly, *Ex.9* also performs well, testing postural stability and upper limb coordination, both impaired in PD. These exercises are valuable for capturing subtle differences between the HC and PwPD1. However, some exercises, like *Ex.8* and *Ex.2*, show inconsistent performance. *Ex.8*, for instance, performs poorly with VGG16 (0.36 F1-micro), likely due to the insufficient dynamic movements. The performance of *Ex.2* also varies, achieving lower performance (0.61), as its static nature limits the discriminative features.

ResNet50V2 and MobileNetV2 are the most reliable architectures. ResNet50V2 achieves perfect performance across the multiple exercises, including *Ex.1*, *Ex.3*, *Ex.5*, *Ex.6*, and *Ex.9*, due to its ability to capture intricate patterns. MobileNetV2, while lightweight, also excels, particularly for *Ex.6* and *Ex.9*, making it efficient for the resource-constrained applications.

As a result, focusing on *Ex.6* and strong architectures (ResNet50V2 or MobileNetV2) ensures accurate classification between the HC and PwPD1. In contrast, *Ex.8* and *Ex.2* should be used cautiously due to its inconsistency.

PwPD1 vs. PwPD1.5: Exercises revealed several trends in distinguishing PwPD1 from PwPD1.5. Among the exercises, *Ex.1* and *Ex.4* stand out as the top performers across multiple architectures, such as GADF InceptionV3, ResNet50V2, and RP InceptionV3. These exercises likely perform well because they involve dynamic movements and coordination, which are sensitive to the progression of PD. For instance, *Ex.1* tests gait, balance, and transitional movements, while *Ex.4* evaluates the fine motor control and coordination, both of which are affected as the disease progresses.

In terms of model architectures, ResNet50V2 and InceptionV3 consistently deliver strong results. These architectures' deep learning frameworks enable them to extract complex patterns from the data, making them well-suited for this classification task. However, some exercises show inconsistent or lower performance. For example, *Ex.8* and

Table 4
F1-micro by sampling frequency, and sensor modalities across exercises.

EX ID	1	2	3	4	5	6	7	8	9
HC vs. PwPD1									
Sensors[a]									
AG	1.000	1.000	1.000	0.960	1.000	1.000	0.900	0.950	1.000
A	1.000	1.000	1.000	0.957	0.955	1.000	0.952	1.000	1.000
G	1.000	0.957	1.000	1.000	1.000	1.000	0.857	1.000	1.000
PwPD1 vs. PwPD1.5.									
Sensors[a]									
AG	1.000	0.860	0.710	1.000	1.000	0.860	0.830	0.500	0.860
A	1.000	0.571	0.857	1.000	0.833	0.714	0.833	0.500	0.714
G	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SF[b]									
32	0.909	1.000	0.957	1.000	0.864	0.957	1.000	1.000	1.000
48	0.773	1.000	1.000	1.000	0.864	0.957	1.000	0.909	0.913
64	0.773	1.000	1.000	1.000	0.864	0.957	1.000	0.909	0.913

A - Accelerometer; G - Gyroscope.
[a] Experiment details: ResNetV2 + GADF, Batch size = 64, and sampling frequency of 32 Hz. segmentation (4s, 1s).
[b] Experiment details: ResNetV2 + GADF + (A + G), Batch size = 4, segmentation (4s, 1s).

Ex.2 yield mixed results, with some architectures like MobileNetV2 and VGG16. These exercises may lack sufficient dynamic or discriminative features to reliably differentiate between the PwPD1 and PwPD1.5. Additionally, *Ex.5* shows variability. The static nature of these exercises may limit their ability to capture the disease progression effectively. **According to the results**, it is recommended to use *Ex.1* and *Ex.4*, leveraging robust architectures like ResNet50V2 or InceptionV3. Similarly, avoid relying solely on *Ex.2*, *Ex.5*, and *Ex.8*.

4.3.2. The impact of sensor data modalities

The ablation study results provide valuable insights into the contribution of the accelerometer (A), gyroscope (G), and their combination (A+G) for distinguishing between the HC and PwPD1, as well as between the PwPD1 and PwPD1.5, across various exercises.

As listed in Table 4, for the HC vs. PwPD1 task, the combination of accelerometer and gyroscope (A+G) achieves perfect classification for most of the exercises, indicating that this fusion provides the most discriminative features. The accelerometer alone also performs exceptionally well, achieving perfect performance in most cases, which suggests that accelerometer data is highly effective for capturing translational movements and force-related features. The gyroscope, while slightly less effective in some cases, still achieves perfect performance in several exercises, indicating its value in capturing rotational dynamics.

In the PwPD1 vs. PwPD1.5 task, the classification performance is generally lower compared to the HC vs. PwPD1, which is expected given the subtler differences between the PwPD1 and PwPD1.5. However, the combination of A+G still achieves the perfect performance in certain exercises, such as *Ex.1*, *Ex.4*, and *Ex.5*, and performs well in others, demonstrating its robustness. The accelerometer shows strong performance in some exercises, but struggles in others, particularly those involving fine motor control or subtle rotational movements. Similarly, the gyroscope exhibits mixed results, excelling in exercises that involve rotational dynamics but performing poorly in tasks that rely more on translational movements or static positions. This underscores the importance of sensor fusion, as the combined use of accelerometer and gyroscope data compensates for the limitations of each individual sensor.

4.3.3. The impact of sampling frequency

The goal is to assess the impact of sampling frequency on the model's performance. The results listed in Table 4 show that the model performs well at the sampling rate 32 Hz, achieving the perfect or near-perfect scores for most of the exercises. This suggests that 32 Hz provides sufficient temporal resolution for capturing the relevant dynamics in most cases. Higher sampling rates (48 Hz and 64 Hz) improve performance for some exercises (e.g., *Ex3*) but are not consistently beneficial. In some cases, e.g., *Ex.1*, *Ex.8*, and *Ex.9*, the higher sampling rates may introduce noise or unnecessary complexity, leading to the slightly reduced performance. This indicates that the higher sampling rates are not always beneficial and should be carefully evaluated based on the specific exercise and task.

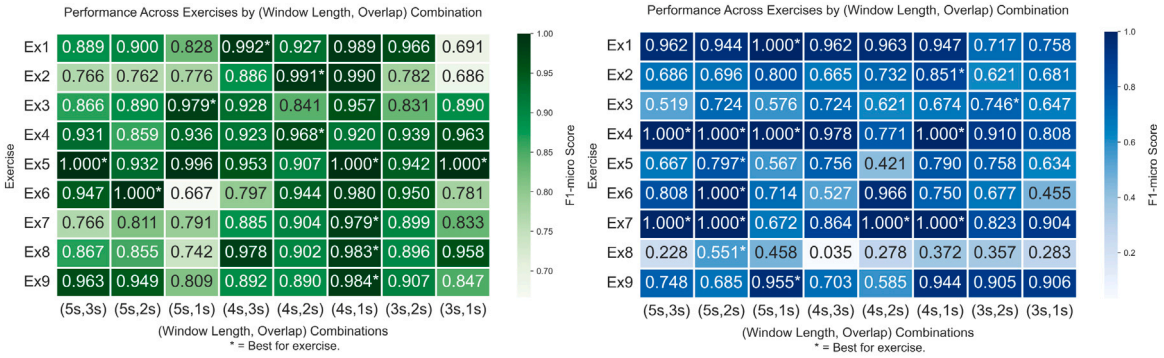
4.3.4. Impact of window length and overlapping

We investigated the impact of window length and overlap on the performance in PD classification tasks through an ablation study (Fig. 3). Our primary results utilized 4 s window with 1 s overlap. However, variations revealed the significant performance differences, particularly, in the dynamic exercises.

As detailed in the experimental protocol, exercise-specific parameters were identified in a separate offline ablation study and were frozen before running the final LOSO experiments. This section presents the analysis that led to the selection of optimal window lengths and overlap parameters, which were subsequently held constant across all main experiments.

For the HC vs. PwPD1 task, longer window lengths, e.g. 5 s, and moderate to high overlaps, e.g. 3 s, enhanced the results, especially in dynamic exercises such as *Ex1* (stand up/walk/sit) and *Ex5* (hold outstretched arms), achieving F1-micro scores 0.992 and 1.000. This configuration captures the full movement cycle, including the gait initiation and transitional dynamics, and effectively addresses the bradykinesia and postural instability. Conversely, static tasks like *Ex2* (sit/rest arms) and *Ex3* (finger tapping) performed better with shorter overlaps, e.g. 1 s, as the static nature of *Ex2* reduces the redundant sampling, while *Ex3* benefits from a longer window to capture the tapping cycles without over-representation of repetitive patterns.

In the PwPD1 vs. PwPD1.5 task, fine motor control exercises, e.g. *Ex4*: touching nose with fingers and *Ex7*: clenching fists, achieved the high scores with longer windows and high overlap. However, static exercises like *Ex8*: standing with palms inward and *Ex9*: standing with palms forward, performed poorly (F1-micro $\leq$ 0.372) due to static nature obscuring nuanced motor differences. Dynamic tasks benefit from the longer windows and high overlap, while static/rhythmic tasks thrive with shorter overlaps. Limited cohort sizes in PwPD1.5 underscore the need for larger datasets to yield robust results.



(a) HC vs. PwPD1 (b) PwPD1 vs. PwPD1.5

Fig. 3. Frame-based F1-micro of the best-found models, using different window lengths and overlaps, for HC vs. PwPD1, and PwPD1 vs. PwPD1.5.

4.4. Model evaluation, statistical analysis, model calibration and confidence analysis

In this section, we present a statistical evaluation of the best-found model for each exercise (from Fig. 3), focusing on two critical classification tasks: HC vs. PwPD1 and PwPD1 vs. PwPD1.5. These tasks were prioritized because they represent the most clinically valuable and challenging objectives for early diagnosis and precise staging. The robustness of these optimal models, initially identified through our ablation studies, is now rigorously validated using statistical methods.

4.4.1. Statistical analysis

Our evaluation of frame-based and subject-based approaches, using ROC curves, PR curves, and confusion matrices, focused on exercise-aware segmentation, as depicted in Figs. C.5 and C.4. We applied statistical validation using Wilson score intervals (95% CI) and bootstrap resampling (n = 1000 for frame-level and n = 100 for subject-level), as summarized in Tables 5 and 6 showing the best-performing models. These visualizations demonstrate the strong discriminative power of our methodology across the classification tasks.

The ROC curves for HC vs. PwPD1 consistently showed near-perfect AUC scores (0.98–1.00), indicating the model's high confidence in distinguishing between these groups. The PR curves further illustrated the substantial precision–recall trade-offs, especially for dynamic exercises like *Ex5*: glass-filling task, where both the precision and recall reached 1.00 (95% CI: 0.936–1.00). For PwPD1 vs. PwPD1.5, while limb-movement tasks, e.g. *Ex7*, maintained high performance (F1 = 1.00, CI: 0.916–1.00), postural tasks revealed broader confidence intervals, highlighting the difficulties in differentiating subtle symptoms of varying stages.

Distinct model behavior emerged between the classification tasks. For HC vs. PwPD1, exercises displayed minimal training-validation gaps (≤0.075 F1-micro), with *Ex5* achieving perfect alignment (gap = 0.000), suggestive of excellent generalization. Exercises *1*, *7*, *8*, and *9* demonstrated stability (gaps ≤0.019) due to clear kinematic distinctions between HC and PwPD1, characterized by noticeable bradykinesia and tremor patterns. Conversely, PwPD1 vs. PwPD1.5 exhibited greater variability, with exercises *2*, *3*, *6*, and *7* maintaining small gaps (≤0.018), indicative of stage-specific features. For instance, exercise 6 captured subtle differences in postural sway variability. However, larger gaps (≥0.114) in exercises *1*, *4*, *5*, *8*, and *9* — peaking at 0.282 for *Ex5* and 0.125 for *Ex8*/*Ex9* — reflect the clinical reality that axial symptoms, such as postural instability, are more challenging to differentiate in the early PD stage.

Statistical validation reinforced these observations, showing the robust performance metrics for HC vs. PwPD1 with the tight confidence intervals (widths < 0.1) and high precision/recall (0.92–1.00 across most exercises). Notably, *Ex5* achieved a perfect F1-score (1.00, CI: 0.985–1.00), demonstrating its sensitivity to the PD-specific motor impairments. In contrast, PwPD1 vs. PwPD1.5 revealed variable task efficacy: the limb-movement tasks were performing reliably, while postural tasks, such as *Ex8*, showed significant degradation in specificity (0.20, CI: 0.105–0.348), indicating diagnostic uncertainty. Interestingly, *Ex4* exhibited contrary efficacy, performing well for PwPD1.5 classifications (F1 = 1.00) compared to its lower performance in the initial detection phase.

The limited sample size for PwPD 1.5 (n = 3) necessitates caution when interpreting the results. While high performance was noted for *Ex4* and *Ex7* in this group, the wider CI for specific exercises suggest the potential overfitting. Future studies should aim at validating these findings on larger datasets, particularly in transitional stages like PwPD 1.5, where symptom heterogeneity is significant.

These results highlight the need for task-specific optimization in the PD assessment. For real-time applications, exercises with minimal training-validation gaps, e.g. *Ex5* for HC vs. PwPD1, and efficient segmentation (4 s windows, 1 s overlap) are ideal due to balancing the latency with accuracy. Dynamic tasks, e.g.*Ex1* and *Ex5*, benefit from longer windows (4–5 s) to capture the complete movement cycles, while the static tasks, e.g. *Ex2* and *Ex8*, perform well with shorter overlaps (1 s) to mitigate redundancy. For research replication, prioritizing exercises with robust CI performance (*Ex5*, *Ex7*) and transparently reporting stage-specific limitations, particularly in small cohorts, is recommended.

4.4.2. Model calibration and confidence analysis

The calibration performance across both the classification tasks revealed the substantial variation in model reliability, with subject-based predictions generally demonstrating better calibration than frame-based approaches across most exercises.

For the HC vs. PwPD1 classification task, models exhibited generally good to excellent calibration. *Ex5* demonstrated exceptional calibration in both the frame-based (ECE = 0.007, Brier = 0.001) and the subject-based (ECE = 0.007, Brier = 0.000) approaches, representing nearly perfect probability calibration. Exercises *1, 3, 7, 8*, and *9* showed good to excellent calibration, with the ECE scores ranging from 0.037 to 0.065. *Ex6* displayed the poorest calibration in this task (frame-based ECE = 0.101, subject-based ECE = 0.076), though still within the moderate calibration range after the subject-level aggrega-

Table 5
Statistical validation of exercise performance for ***HC vs. PwPD1*** classification using Wilson score intervals (95% CI) with bootstrap resampling (n = 1000 iterations).

(a) Per-class frame-level and subject-level performance metrics

	TSI	W/O	Arch	Precision	Recall	Specificity	F1	Precision	Recall	Specificity	F1
Frame				HC				PD1.0			
1	CWT	4s,3s	MobileNetV2	0.990 (0.964–0.997)	1.000 (0.981–1.000)	0.957 (0.855–0.988)	0.995	1.000 (0.920–1.000)	0.957 (0.855–0.988)	1.000 (0.981–1.000)	0.978
2	GADF	4s,2s	MobileNetV2	1.000 (0.990–1.000)	0.989 (0.972–0.996)	1.000 (0.961–1.000)	0.995	0.960 (0.901–0.984)	1.000 (0.961–1.000)	0.989 (0.972–0.996)	0.979
3	GADF	5s,1s	MobileNetV2	0.975 (0.942–0.989)	1.000 (0.980–1.000)	0.896 (0.778–0.955)	0.987	1.000 (0.918–1.000)	0.896 (0.778–0.955)	1.000 (0.980–1.000)	0.945
4	GADF	4s,2s	MobileNetV2	1.000 (0.986–1.000)	0.964 (0.935–0.980)	1.000 (0.898–1.000)	0.982	0.773 (0.630–0.872)	1.000 (0.898–1.000)	0.964 (0.935–0.980)	0.872
5	GADF	4s,1s	ResNet50V2	1.000 (0.985–1.000)	1.000 (0.985–1.000)	1.000 (0.936–1.000)	1.000	1.000 (0.936–1.000)	1.000 (0.936–1.000)	1.000 (0.985–1.000)	1.000
6	GADF	5s,2s	MobileNetV2	1.000 (0.952–1.000)	1.000 (0.952–1.000)	1.000 (0.785–1.000)	1.000	1.000 (0.785–1.000)	1.000 (0.785–1.000)	1.000 (0.952–1.000)	1.000
7	GASF	4s,1s	MobileNetV2	1.000 (0.983–1.000)	0.974 (0.945–0.988)	1.000 (0.934–1.000)	0.987	0.900 (0.799–0.953)	1.000 (0.934–1.000)	0.974 (0.945–0.988)	0.947
8	CWT	4s,1s	MobileNetV2	0.984 (0.959–0.994)	0.996 (0.977–0.999)	0.922 (0.815–0.969)	0.990	0.979 (0.891–0.996)	0.922 (0.815–0.969)	0.996 (0.977–0.999)	0.949
9	GADF	4s,1s	MobileNetV2	0.992 (0.972–0.998)	0.988 (0.966–0.996)	0.968 (0.891–0.991)	0.990	0.953 (0.871–0.984)	0.968 (0.891–0.991)	0.988 (0.966–0.996)	0.961
Subject				HC				PwPD1			
1	CWT	4s,3s	MobileNetV2	1.000 (0.816–1.000)	1.000 (0.816–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.816–1.000)	1.000
2	GADF	4s,2s	MobileNetV2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.824–1.000)	1.000
3	GADF	5s,1s	MobileNetV2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.824–1.000)	1.000
4	GADF	4s,2s	MobileNetV2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.824–1.000)	1.000
5	GADF	4s,1s	ResNet50V2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.510–1.000)	1.000	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.824–1.000)	1.000
6	GADF	5s,2s	MobileNetV2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.824–1.000)	1.000
7	GASF	4s,1s	MobileNetV2	1.000 (0.816–1.000)	1.000 (0.816–1.000)	1.000 (0.510–1.000)	1.000	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.816–1.000)	1.000
8	CWT	4s,1s	MobileNetV2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.510–1.000)	1.000	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.824–1.000)	1.000
9	GADF	4s,1s	MobileNetV2	1.000 (0.824–1.000)	1.000 (0.824–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.824–1.000)	1.000

(b) Micro and macro average performance metrics

	TSI	W/O	Arch	$Precision_{mic}$	$Recall_{mic}$	$F1_{mic}$	$Precision_{mac}$	$Recall_{mac}$	$F1_{mac}$	ROC	AP
Frame-level											
1	CWT	4s,3s	MobileNetV2	0.992 (0.983–0.998)	0.992 (0.983–0.998)	0.992 (0.983–0.998)	0.994 (0.987–0.999)	0.983 (0.959–0.999)	0.988 (0.972–0.999)	0.998 (0.993–1.000)	0.992 (0.976–1.000)
2	GADF	4s,2s	MobileNetV2	0.991 (0.985–0.997)	0.991 (0.985–0.997)	0.991 (0.986–0.997)	0.984 (0.969–0.995)	0.993 (0.988–0.997)	0.988 (0.978–0.996)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
3	GADF	5s,1s	MobileNetV2	0.979 (0.965–0.992)	0.979 (0.965–0.990)	0.979 (0.967–0.992)	0.985 (0.974–0.993)	0.958 (0.928–0.986)	0.970 (0.947–0.988)	0.997 (0.989–1.000)	0.991 (0.971–1.000)
4	GADF	4s,2s	MobileNetV2	0.968 (0.953–0.981)	0.968 (0.953–0.981)	0.968 (0.953–0.981)	0.914 (0.871–0.952)	0.977 (0.967–0.987)	0.940 (0.909–0.967)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
5	GADF	4s,1s	ResNet50V2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
6	GADF	5s,2s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
7	GASF	4s,1s	MobileNetV2	0.979 (0.967–0.990)	0.979 (0.967–0.990)	0.979 (0.967–0.990)	0.960 (0.931–0.982)	0.985 (0.975–0.992)	0.971 (0.951–0.987)	0.993 (0.982–1.000)	0.921 (0.822–0.998)
8	CWT	4s,1s	MobileNetV2	0.983 (0.971–0.993)	0.983 (0.971–0.992)	0.983 (0.973–0.992)	0.982 (0.963–0.994)	0.967 (0.939–0.990)	0.974 (0.954–0.990)	0.993 (0.981–1.000)	0.980 (0.950–0.999)
9	GADF	4s,1s	MobileNetV2	0.984 (0.975–0.994)	0.984 (0.973–0.994)	0.984 (0.975–0.994)	0.976 (0.955–0.993)	0.980 (0.962–0.994)	0.978 (0.962–0.992)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
Subject-level											
1	CWT	4s,3s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
2	GADF	4s,2s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
3	GADF	5s,1s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
4	GADF	4s,2s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
5	GADF	4s,1s	ResNet50V2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
6	GADF	5s,2s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
7	GASF	4s,1s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
8	CWT	4s,1s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
9	GADF	4s,1s	MobileNetV2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)

Metrics are micro-averages and are reported using a batch size of 64 and 32 sampling frequency. Experiment details: (A + G), Batch size = 64.

tion. The subject-based approach consistently improved the calibration reliability, reducing ECE by average 24% compared to frame-based predictions. This improvement was particularly notable for *Ex6*, where the subject-level aggregation reduced ECE from 0.101 to 0.076.

The PwPD1 vs. PwPD1.5 classification imposed greater calibration challenges, with greater variability in the performance across exercises. *Ex7* achieved perfect calibration (ECE = 0.000, Brier = 0.000) in both the approaches, while Exercises *1, 4, 6,* and *9* maintained good to excellent calibration (ECE = 0.039–0.076). However, Exercises *2, 3, 5*, and *8* demonstrated poor calibration, with ECE scores ranging from 0.130 to 0.276. Notably, Exercise *8* showed the poorest calibration performance across both the tasks (frame-based ECE = 0.256, subject-based ECE = 0.276), indicating substantial miscalibration that persisted the despite subject-level aggregation. The greater calibration variability in this task suggests increased difficulty in distinguishing between the PD severity stages compared to healthy versus disease classification (see Table 7).

4.4.3. Concluding remarks

In conclusion, the results of the ablation study emphasize the importance of combining accelerometer and gyroscope data to achieve robust performance across a wide range of exercises.rticularly effective for tasks involving translational movements and static positions, while the gyroscope excels at capturing rotational dynamics and fine motor control. For HC vs. PwPD1, the dataset is highly separable, and all the sensors' configurations perform well. However, for PwPD1 vs. PwPD1.5, the subtler differences between the classes necessitate the combined use of both sensors to achieve reliable performance. Higher sampling rates do not consistently lead to better performance, suggesting that 32 Hz may be sufficient for most tasks.

Based on the calibration metrics and the results corroborated by rigorous statistical validation including the bootstrap resampling and Wilson score intervals, the best-performing exercises for each classification task were clearly identified. For the HC vs. PwPD1.0 task, *Ex5* (Hold the outstretched arms position) demonstrated excellent calibration and near-perfect statistical reliability (F1 = 1.00, CI: 0.985–1.00), followed by exercises *7, 8, 9* which also showed the strong calibration and consistent performance. In more challenging PwPD1.0 vs. PwPD1.5 classification, *Ex7* (Clench and unclench fists on the table) achieved perfect calibration and statistical robustness (F1 = 1.00), while *Ex1* (Stand up/walk/sit) and 4 (Touch nose with fingers) maintained good calibration reliability despite the limited cohort size. The consistent, strong performance of *Ex5*, *Ex5*, and *Ex5* across both tasks, validated through both calibration metrics and statistical analysis, confirms that these specific movement patterns provide the most reliable probability estimates for clinical decision support. Furthermore, the superior calibration observed in the subject-based predictions reinforces their clinical utility for individual patient assessments, while the complementary frame-based analysis ensures comprehensive model evaluation for real-world deployment.

The current study foundation in established clinical practice is demonstrated by the direct alignment of our nine exercises with specific items from the MDS-UPDRS Part III motor examination [46], as detailed in Table 2. This mapping ensures that the movement features extracted by the proposed model correspond directly to the cardinal motor signs of PD that neurologists assess qualitatively. For instance, the high performance of models analyzing *Ex3 (Finger Tapping)* and *Ex4 (Hand Movements)* indicates a robust capacity to digitally quantify the bradykinesia and amplitude decrement that a clinician would score in

Table 6
Statistical validation of exercise performance for **PwPD1 vs. PwPD1.5** classification using Wilson score intervals (95% CI) with bootstrap resampling (n = 1000 iterations).

(a) Per-class frame-level and subject-level performance metrics

	TSI	W/O	Arch	Precision	Recall	Specificity	F1	Precision	Recall	Specificity	F1
Frame				PwPD1				PwPD1.5			
1	GADF	5s,1s	InceptionV3	1.000 (0.722–1.000)	1.000 (0.722–1.000)	1.000 (0.566–1.000)	1.000	1.000 (0.566–1.000)	1.000 (0.566–1.000)	1.000 (0.722–1.000)	1.000
2	GADF	4s,1s	vgg16	0.790 (0.674–0.873)	0.980 (0.895–0.996)	0.705 (0.558–0.818)	0.875	0.969 (0.843–0.994)	0.705 (0.558–0.818)	0.980 (0.895–0.996)	0.816
3	GADF	3s,2s	InceptionV3	0.751 (0.681–0.811)	0.830 (0.763–0.881)	0.635 (0.544–0.717)	0.789	0.737 (0.643–0.814)	0.635 (0.544–0.717)	0.830 (0.763–0.881)	0.682
4	GADF	5s,2s	InceptionV3	1.000 (0.851–1.000)	1.000 (0.851–1.000)	1.000 (0.741–1.000)	1.000	1.000 (0.741–1.000)	1.000 (0.741–1.000)	1.000 (0.851–1.000)	1.000
5	MTF	5s,2s	InceptionV3	0.719 (0.592–0.819)	1.000 (0.914–1.000)	0.579 (0.422–0.721)	0.837	1.000 (0.851–1.000)	0.579 (0.422–0.721)	1.000 (0.914–1.000)	0.733
6	GASF	5s,2s	InceptionV3	1.000 (0.722–1.000)	1.000 (0.722–1.000)	1.000 (0.701–1.000)	1.000	1.000 (0.701–1.000)	1.000 (0.701–1.000)	1.000 (0.722–1.000)	1.000
7	GADF	5s,2s	vgg16	1.000 (0.910–1.000)	1.000 (0.910–1.000)	1.000 (0.916–1.000)	1.000	1.000 (0.916–1.000)	1.000 (0.916–1.000)	1.000 (0.910–1.000)	1.000
8	GADF	5s,2s	InceptionV3	0.522 (0.405–0.637)	0.921 (0.792–0.973)	0.200 (0.105–0.348)	0.667	0.727 (0.434–0.903)	0.200 (0.105–0.348)	0.921 (0.792–0.973)	0.314
9	GADF	5s,1s	vgg16	1.000 (0.901–1.000)	0.921 (0.792–0.973)	1.000 (0.883–1.000)	0.959	0.906 (0.758–0.968)	1.000 (0.883–1.000)	0.921 (0.792–0.973)	0.951
Subject				PwPD1				PwPD1.5			
1	GADF	5s,1s	InceptionV3	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.439–1.000)	1.000	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.510–1.000)	1.000
2	GADF	4s,1s	vgg16	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.439–1.000)	1.000	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.510–1.000)	1.000
3	GADF	3s,2s	InceptionV3	0.800 (0.376–0.964)	1.000 (0.510–1.000)	0.667 (0.208–0.939)	0.889	1.000 (0.342–1.000)	0.667 (0.208–0.939)	1.000 (0.510–1.000)	0.800
4	GADF	5s,2s	InceptionV3	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.439–1.000)	1.000	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.510–1.000)	1.000
5	MTF	5s,2s	InceptionV3	0.750 (0.301–0.954)	1.000 (0.439–1.000)	0.667 (0.208–0.939)	0.857	1.000 (0.342–1.000)	0.667 (0.208–0.939)	1.000 (0.439–1.000)	0.800
6	GASF	5s,2s	InceptionV3	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.439–1.000)	1.000	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.510–1.000)	1.000
7	GADF	5s,2s	vgg16	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000
8	GADF	5s,2s	InceptionV3	0.500 (0.188–0.812)	1.000 (0.439–1.000)	0.000 (0.000–0.561)	0.667	0.000 (0.000–0.000)	0.000 (0.000–0.561)	1.000 (0.439–1.000)	0.000
9	GADF	5s,1s	vgg16	1.000 (0.510–1.000)	1.000 (0.510–1.000)	1.000 (0.439–1.000)	1.000	1.000 (0.439–1.000)	1.000 (0.439–1.000)	1.000 (0.510–1.000)	1.000

(b) Micro and macro average performance metrics

	TSI	W/O	Arch	$Precision_{mic}$	$Recall_{mic}$	$F1_{mic}$	$Precision_{mac}$	$Recall_{mac}$	$F1_{mac}$	ROC	AP
Frame-level											
1	GADF	5s,1s	InceptionV3	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
2	GADF	4s,1s	vgg16	0.851 (0.803–0.899)	0.851 (0.798–0.894)	0.851 (0.798–0.899)	0.870 (0.821–0.912)	0.845 (0.789–0.896)	0.847 (0.789–0.899)	0.972 (0.937–0.997)	0.973 (0.939–0.996)
3	GADF	3s,2s	InceptionV3	0.746 (0.711–0.784)	0.746 (0.709–0.782)	0.746 (0.711–0.784)	0.745 (0.704–0.784)	0.737 (0.697–0.779)	0.739 (0.696–0.779)	0.834 (0.783–0.878)	0.793 (0.716–0.866)
4	GADF	5s,2s	InceptionV3	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
5	MTF	5s,2s	InceptionV3	0.797 (0.734–0.861)	0.797 (0.728–0.861)	0.797 (0.741–0.854)	0.839 (0.790–0.887)	0.792 (0.730–0.852)	0.789 (0.724–0.851)	0.906 (0.837–0.960)	0.919 (0.858–0.970)
6	GASF	5s,2s	InceptionV3	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
7	GADF	5s,2s	vgg16	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
8	GADF	5s,2s	InceptionV3	0.551 (0.474–0.628)	0.551 (0.474–0.628)	0.551 (0.474–0.628)	0.600 (0.488–0.704)	0.557 (0.494–0.616)	0.511 (0.426–0.593)	0.565 (0.428–0.700)	0.656 (0.509–0.786)
9	GADF	5s,1s	vgg16	0.955 (0.918–0.985)	0.955 (0.918–0.985)	0.955 (0.918–0.985)	0.954 (0.911–0.986)	0.959 (0.922–0.988)	0.955 (0.914–0.985)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
Subject-level											
1	GADF	5s,1s	InceptionV3	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
2	GADF	4s,1s	vgg16	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
3	GADF	3s,2s	InceptionV3	0.857 (0.643–1.000)	0.857 (0.643–1.000)	0.857 (0.643–1.000)	0.886 (0.496–1.000)	0.841 (0.592–1.000)	0.849 (0.528–1.000)	0.750 (0.167–1.000)	0.833 (0.286–1.000)
4	GADF	5s,2s	InceptionV3	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
5	MTF	5s,2s	InceptionV3	0.833 (0.583–1.000)	0.833 (0.583–1.000)	0.833 (0.583–1.000)	0.861 (0.500–1.000)	0.833 (0.541–1.000)	0.830 (0.497–1.000)	0.889 (0.400–1.000)	0.917 (0.500–1.000)
6	GASF	5s,2s	InceptionV3	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
7	GADF	5s,2s	vgg16	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
8	GADF	5s,2s	InceptionV3	0.500 (0.250–0.750)	0.500 (0.250–0.750)	0.500 (0.250–0.750)	0.333 (0.148–0.667)	0.500 (0.200–0.667)	0.389 (0.151–0.625)	0.778 (0.200–1.000)	0.867 (0.333–1.000)
9	GADF	5s,1s	vgg16	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)

Metrics are micro-averages and are reported using a batch size of 64 and 32 sampling frequency. Experiment details: (A + G), Batch size = 64.

Table 7
Calibration metrics of the best-performing model per exercise.

	HC vs. PwPD1.0				PwPD1.0 vs. PwPD1.5			
	ECE_F	$Brier_F$	ECE_S	$Brier_S$	ECE_F	$Brier_F$	ECE_S	$Brier_S$
Ex 1	0.053	0.024	0.041	0.010	0.052	0.007	0.041	0.004
Ex 2	0.061	0.030	0.054	0.019	0.149	0.143	0.130	0.052
Ex 3	0.056	0.021	0.065	0.014	0.167	0.189	0.169	0.148
Ex 4	0.088	0.050	0.052	0.022	0.060	0.008	0.042	0.004
Ex 5	0.007	0.001	0.007	0.000	0.160	0.152	0.192	0.141
Ex 6	0.101	0.025	0.076	0.013	0.076	0.020	0.069	0.013
Ex 7	0.041	0.031	0.045	0.010	0.000	0.000	0.000	0.000
Ex 8	0.037	0.024	0.050	0.012	0.256	0.304	0.276	0.255
Ex 9	0.057	0.020	0.044	0.012	0.045	0.031	0.039	0.011

F stands for frame-based, while *S* stands for subject-based results.

MDS-UPDRS items 3.4 and 3.5. Similarly, the postural sway and tremor captured during *Ex5 (Postural Tremor)* are the objective correlates of the impairments rated in item 3.15. While the proposed deep learning model outputs a classification probability rather than a direct 0–4 MDS-UPDRS score, there is a clear qualitative relationship: a higher model confidence for a PD classification corresponds to the presence of more severe motor impairments that would be reflected in a higher (worse) score on the corresponding MDS-UPDRS items. This linkage underscores that our system is not merely a black-box classifier, but an objective tool that measures the well-defined clinical constructs, thereby strengthening its potential for integration into the clinical assessment pipelines.

4.5. Comparison with related works

We provide a comparative analysis with the state-of-the-art methods to provide insights about the performance and complexity of the proposed approach. Table 8 lists information on similar studies, including the objectives, size of data sets, exercises, and their methods. In Table 9, a side-by-side comparison is performed, with detailed comparisons, identifying trends and differences, showcasing how this work aligns with and advances the state-of-the-art research.

As shown in Table 9, this work demonstrates consistently high performance for certain exercises, such as *Ex.1* and *Ex.5*, due to their distinct and well-defined motor patterns. This observation aligns with the findings reported in the study [64], where the authors achieved

Table 8
Datasets description of similar studies.

Work	Task	Dataset # (Male/Female)	#Exercises	Methods
Two IMU on wrists (one each) [28]	HC/PD	26 HC (20/6), age range: [55, 84] and [22, 45] 32 PD (16/16), age range: [55, 84]. PD1-1.5 (3/5) PD2-2.5 (11/9) PD3 (2/2)	11	• SF = 416 Hz, STFT + 1D-CNN • LOSO + K-fold cross-validation (2), where 2 HC and 2 PwPD • are used for testing. • 5 cross-validation. • No signal segmentation.
One IMU is placed on the dorsal side of the dominant hand, or video or image data [33]	HC/PD	83 subjects (44/39), age range: [22, 84], with average age: 58.5 24 HC 42 PD (stage 1, 1.5, 2, 2.5, 3)	15	• Signal segmentation (3 s frames and 50% overlapping). • kfold cross-validation. • Dimensionality reduction + RF, XGBoost.
One IMU is placed on the dorsal side of the dominant hand and/or video [31]	HC/PD PD/ET	24 HC (15/9). 42 PD (19/23) 8 PD1 (1/7), 26 PD2 (15/11) 7 PD3 (2/5), 1 PD4 (1/NA) 13 ET (4/9)	15	• SF = 100 Hz. • Signal segmentation (3-sec frames). • Logistic regression, XGBoost, RF, SVM, • Gaussian process classifiers. • 10-fold cross-validation. • Voting and concatenation decision.
Four IMU sensors (wrist and dorsum) [64]	HC/PD	21 HC 33 PD	11	• SF = 416 Hz and downsampled to 64 Hz. • Signal segmentation (4-sec frames and 1-sec overlapping). • Limited duration of 11 s. • Logistic regression, DT, RF, XGBoost, NB, MLP, KNN. • Tested two frequency ranges.
Four IMU sensors (wrist and dorsum) [30]	HC/PD1 HC/PD2 PD1/PD2	31 HC (17/4) 14 PD (4/10) 40 PD2 (2/10) 10 PD3 (4/6) 1 PD4 (1/NA)	11	• Signal segmentation (3-sec frames). • 5-fold cross-validation. • Zero overlap between train and validation dataset. • Gini-based feature importance from RF and Factor analysis. • RF, LightGBM.
Two IMU on wrists (one each) [65]	HC/PD, HC/PD1/ PD2/PD3, HC/PD1 HC/PD2 HC/PD3 PD1/PD2 PD1/PD3 PD2/PD3	30 HC, 55 PD, 17 PD1 15 PD1.5 12 PD2 6 PD2.5 5 PD3	11	• SF = 416 Hz and downsampled to 64 Hz. • Signal segmentation (4-sec frames and 1-sec overlapping). • Limited duration of 11 s • 774 features + Feature importance (ANOVA). • 5-fold cross-validation. • 70% train and 30% validation. • Zero overlap between train and validation dataset. • ML: Logistic regression, DT, RF, XGBoost, NB, MLP, KNN.

perfect F1-score of 1.000 using data from four IMU sensors placed on wrists and dorsums. In contrast, other exercises, such as *Ex.2* and *Ex.7*, tend to yield lower performance, primarily because they involve more subtle or repetitive movements that are harder to distinguish. This low performance underscores the challenges in accurately classifying exercises with less pronounced motor patterns, highlighting the need for more refined methodologies to improve the detection in such cases.

From Table 9, it is evident that the proposed approach consistently achieves a perfect F1-score of 1.000 across most exercises, demonstrating its robustness and ability to generalize effectively, even for challenging tasks. For instance, in *Ex.1*, the proposed method matches the performance of the four IMU sensor-based approaches (Wrists + Dorsums), achieving the identical perfect F1-score. Similarly, in *Ex.5*, the proposed method surpasses the results of the study by [65], which used the left and right IMU wrist sensors and achieved an F1-score of 0.867.

In contrast, other methods, such as Video + IMU on Dominant Hand's Dorsum + RF [31], exhibit the variability in performance with F1-score ranging from 0.920 to 0.960 for HC vs. PwPD classification tasks. For instance, in *Ex.4*, this approach achieves an F1-score of 0.930, while the proposed method attains a perfect score of 1.000. Similarly, in *Ex.9*, the proposed method helps to achieve an F1-score 0.910, significantly outperforming the Left + Right Wrist Sensor approach, which achieves a score of 0.714. This consistent superiority highlights the effectiveness of the proposed methodology in handling the diverse and challenging tasks, further underscoring its robustness and generalizability.

While the proposed method delivers top-tier performance, its complexity should be considered against other approaches. Both the four-IMU-based approach and video and dorsum-worn-IMU studies rely on multiple sensors and advanced data fusion, which increases both the hardware and the computational demands. In comparison, the systems using just two wrist sensors offer a good balance, achieving strong results with fewer sensors and simpler algorithms. On the other hand, single-sensor methods, such as "One IMU on Dominant Hand's Dorsum + RF", are less resources-demanding, but underperform due to limited data. Multisensor fusion consistently boosts the performance, for example, combining sensors and video achieves an F1-score of 0.960, surpassing the single-sensor-based. Similarly, fusing data from both wrists yields an F1-score of up to 0.933, highlighting the advantages of multi-sensor integration.

4.5.1. Head-to-head comparison

To provide a stronger head-to-head comparison, we compared a classical RF model on our data under the same settings (preprocessing, LOSO using Grid search for a hyper parameter tuning (refer to Table B.12)). For each window of the recorded signal, a total of 20 features are extracted to capture its temporal structure and spectral characteristics (see Table B.12).

It is important to mention that we used the same segmentation parameters used in the best-found models reported in Tables 5 and 6. The results are reported in Table B.13 along with the searcher parameters space.

Across both the classification problems, the RF baseline provides a strong classical reference point, as listed below:

Table 9
F1-score comparative study of this work and the state-of-the-art.

Work	1	2	3	4	5	6	7	8	9
HC vs. PwPD									
IMU on the dominant hand's dorsum + RF [31]	0.840	0.840	–	0.870	0.880	0.840	–	0.880	0.870
Video + RF [31]	0.830	0.880	–	0.920	0.870	0.910	–	0.880	0.830
Video + IMU on dominant hand's dorsum + RF [31]	0.960	0.940	–	0.930	**0.940**	**0.940**	–	**0.940**	0.920
One IMU on dominant hand's dorsum + XGB [31]	0.830	0.850	–	0.870	0.860	0.830	–	0.880	0.850
Video +XGBoost [31]	0.820	0.880	–	0.920	0.870	0.900	–	0.880	0.810
Video + IMU on dominant hand's dorsum + XGBoost [31]	0.940	0.930	–	0.920	0.900	0.910	–	0.900	0.860
One IMU on the dorsal part of the dominant hand [33]	0.880	0.910	–	0.930	0.880	0.860	–	0.820	0.850
Video [33]	0.780	–	–	0.840	–	0.880	–	–	–
Two IMU wrist sensors + Modified STFT + 1D-CNN [28]	0.780	0.770	0.810	0.850	0.920	0.870	–	0.820	0.820
Two IMU wrist sensors +Standard STFT + 1D-CNN [28]	0.800	0.760	0.700	0.850	0.860	0.820	–	0.850	0.760
Left wrist sensor [65]	0.800	0.813	0.750	0.813	0.867	0.875	0.933	0.867	0.733
Right wrist sensor [65]	0.800	0.800	0.867	0.813	0.875	0.813	0.800	0.733	0.800
Left + Right wrist sensor [65]	0.800	0.867	0.867	0.867	0.867	0.800	8.670	0.933	**0.933**
4 sensors (wrists + dorsums) [64]	**1.000**	**1.000**	**1.000**	**1.000**	0.929	0.857	**1.000**	0.929	0.867
HC vs. PwPD1									
Four IMU sensors (wrists and dorsum) [30]	0.740	0.680	0.620	0.79	0.760	0.700	0.600	0.690	0.520
Left wrist sensor [65]	**1.000**	0.889	0.889	**1.000**	**1.000**	0.667	**1.000**	0.889	0.889
Right wrist sensor [65]	0.750	0.875	**1.000**	0.875	0.889	0.889	0.875	0.875	0.889
Left + Right wrist sensor [65]	0.875	**1.000**	**1.000**	0.875	0.875	0.889	0.875	0.875	0.889
This work (exercise-aware segmentation)	**0.992**	**0.991**	**0.979**	**0.968**	**1.000**	**1.000**	**0.979**	**0.983**	**0.984**
HC vs. PwPD2									
Four IMU sensors (wrists and dorsum) [30]	0.760	0.720	0.790	0.760	0.830	0.820	0.840	0.860	0.860
Left wrist sensor [65]	0.750	0.750	0.833	0.833	0.917	0.750	0.727	0.909	0.818
Right wrist sensor [65]	0.750	0.750	0.750	0.583	**1.000**	0.750	0.917	0.750	**1.000**
Left + Right wrist sensor [65]	0.750	0.750	0.750	0.750	0.917	0.750	**1.000**	0.727	0.727
This work (4s window and 1s overlap)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.950**	**1.000**	**1.000**	0.980
PwPD1 vs. PwPD2									
Four IMU sensors (wrists and dorsum) [30]	0.710	0.720	0.710	0.610	0.680	0.780	0.690	0.630	0.690
Left wrist sensor [65]	0.750	0.750	0.667	0.750	0.875	0.750	0.714	0.875	0.875
Right wrist sensor [65]	0.750	0.714	0.750	**1.000**	0.875	0.750	**1.000**	**1.000**	0.875
Left + Right wrist sensor [65]	0.750	0.750	0.750	**1.000**	0.875	0.750	**1.000**	**1.000**	–
This work (exercise-aware segmentation)	**1.000**	**1.000**	**0.860**	**1.000**	**1.000**	**0.860**	**1.000**	0.670	**1.000**
HC vs. PwPD1 vs. PwPD2 vs. PwPD3									
Left wrist sensor [65]	0.600	0.688	0.625	0.667	0.800	0.667	0.786	0.714	0.733
Right wrist sensor [65]	0.533	0.733	0.467	0.667	0.667	0.733	0.667	0.533	0.667
Left + Right wrist sensor [65]	0.600	0.733	0.467	0.733	**0.867**	0.733	0.786	0.643	0.714
This work (4s window and 1s overlap for multiclassification task)	**0.980**	**0.950**	**0.930**	**1.000**	0.820	**0.980**	**0.890**	**0.950**	**0.910**

- For HC vs. PwPD1, RF already reach reasonably high frame-level micro- and macro-F1 (F1-micro $\in [0.93 - 1]$), with the tight confidence intervals, demonstrating that the relatively simple hand-crafted descriptors capture the substantial disease-related information.
- For PwPD1 vs. PwPD1.5, RF also attain the essentially perfect discrimination, indicating that the classical features can separate the subtle disease stages when the model is well configured, although some configurations fail to detect PwPD1.5 at all and yield poor specificity for one class.

The TSI-based models, particularly those using the GADF/GASF/CWT representations with MobileNetV2, ResNet50V2, or InceptionV3, generally match or surpass the best RF configurations on both task, as listed below:

- In HC vs. PwPD1, many TSI setups show the frame-level precision, recall, specificity, and F1 all $\geq$0.97 for both the classes (often 1.00), with micro- and macro-F1 and ROC/AP values extremely close to 1.00.
- Similarly, for PwPD1 vs. PwPD1.5, multiple TSI configurations achieve perfect or near-perfect per-class metrics and aggregate scores, again with narrow confidence intervals.

These results suggest that TSI representations enable the deep networks to exploit richer temporal–spatial structure than the engineered RF features, giving a clear advantage in terms of best-case performance and robustness when the representation–architecture combination is well chosen. However, the TSI pipeline is more configuration-sensitive and computationally demanding than the RF approach. Suboptimal TSI choices (e.g. certain GADF settings with specific windowing) show markedly reduced performance, with macro F1 dropping to around 0.5–0.8 and clear imbalances between precision and recall for one of the classes, whereas many RF settings remain stably high once a reasonable feature set and hyperparameters are fixed. In practice, RF offers simplicity, interpretability of features, and relatively stable behavior once tuned, making it an informative and strong baseline. The TSI-based deep models, in contrast, require careful architectural selection and hyperparameter tuning, but when properly configured, they yield consistently superior performance across HC vs. PwPD1 and PwPD1 vs. PwPD1.5, illustrating the added discriminative power of learned TSI representations over purely hand-crafted features.

The proposed approach offers a simple and cost-effective method for monitoring PD progression around the clock. In contrast, most research works relied on the complex testbeds involving multiple modalities or sensors attached to various body parts, which could be bothersome for the patients and doctors.

4.6. Limitations and future work

Although our left-wrist monitoring approach is clinically feasible, it has several limitations. First, the single-site sensing cannot fully capture the axial or bilateral symptoms characteristic of PD progression [66]. Second, the inherent diagnostic ambiguity of Hoehn and Yahr stage 1.5, representing a transitional phase between the unilateral and bilateral symptoms [46], resulted in a limited cohort size (n = 3) that may affect the statistical power [67]. Third, current approach of equally weighting all three axes in the RGB conversion, while preventing bias, may not optimally represent axis-specific movement patterns. Axis-wise scalar weighting has no effect because the per-segment normalization to $[-1, 1]$ mathematically cancels any multiplicative scaling before the TSI transformation. The resulting GASF/GADF images depend only on the normalized samples, making weighted and unweighted signals identical. Any residual global scaling is further neutralized by CNN invariances (ReLU, convolution, batch normalization), preventing axis weights from influencing the model's predictions. Finally, the computational demands of our server-based analysis platform currently preclude the development of real-time applications in clinical settings. Another limitation is that we applied and validated the exercise-aware segmentation approach only on two classification tasks: HC vs. PwPD 1 and PwPD 1 vs. PwPD 1.5. We note here that the diagnosis and classification of 1–2 early stages is the most challenging for doctors as it is not a trivial task to distinguish them. Further experiments will be conducted on other classification tasks.

To address these limitations, we plan to pursue several research directions. We will expand the data collection to include additional body sites (right wrist, ankles, and trunk) to capture a more comprehensive view of full-body symptomology. Model optimization via TensorFlow Lite conversion and 8-bit quantization will enable efficient deployment on edge devices. This allows real-time analysis. Large-scale validation studies will be conducted through multi-center collaborations [68]. To address class imbalance and enhance model robustness, we will investigate generative adversarial networks,as data augmentation, to improve representation of early and intermediate PD stages. We will also explore multimodal TSI fusion approaches and investigate adaptive weighting schemes for different movement axes to improve the detection of specific motor symptoms via the DL models with learnable axis weights that can automatically discover optimal weighting patterns across different PD symptom manifestations. These developments will contribute to a more comprehensive and extensible framework for PD monitoring.

5. Conclusions

In this research, we have proposed an intelligent framework for the PD detection and severity assessment using a single wrist-worn sensor and time-series imaging (TSI) algorithms. Within the proposed framework, we process the signals from accelerometers and gyroscopes, transforming them into images, and employ transfer learning with convolutional neural networks (CNNs). Our approach achieves an average F1-score of 0.986 in distinguishing between the HC and PwPD 1. Notably, for functional tasks, such as holding outstretched arms and filling a glass with water, the system achieved perfect classification (F1-score is 1.0), highlighting its significant clinical relevance. For staging PD severity (PwPD 1 vs. PwPD 1.5), the system maintained strong performance (F1 = 0.878), with dynamic exercises, such as walking and nose-touching, achieving an F1-score of 1.0. The Gramian Angular Difference Field (GADF) paired with ResNet proved to be the most effective in capturing the PD motor patterns. Optimized parameters, such as 32 Hz sampling rate and task-specific windowing (5s/3s for dynamic movements, 4s/1s for real-time use), ensured the robust detection without unnecessary computational overhead.

To ensure statistical reliability, we incorporated 95% confidence intervals, per-exercise confusion matrices, and calibration analyses (ECE, Brier score), confirming that the model's predictions are both stable and well-calibrated. A head-to-head comparison with a classical Random Forest baseline further demonstrated the clear performance advantage of the proposed TSI-based approach. This work advances the accessible, real-time PD monitoring, offering a scalable tool for early PD diagnosis and progression tracking.

CRediT authorship contribution statement

Mohammed Hammoud: Writing – original draft, Investigation, Data curation. **Aleksei Shcherbak:** Project administration, Methodology, Investigation. **Ekaterina Bril:** Validation, Software, Methodology. **Maksim Semenov:** Writing – original draft, Visualization, Software, Data curation, Conceptualization. **Oleg Sergiyenko:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Andrey Somov:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4F0002 and the agreement No. 139-10-2025-033.

Appendix A. The results of models trained with various architectures and TSI algorithms using segmentation parameters of (4s, 1s).

See Tables A.10 and A.11.

Appendix B. Random forest model - implementation details and performance

See Tables B.12 and B.13.

Appendix C. The best-performing models (after segmentation aware)

C.1. ROC, PR plots and confusion matrices

See Fig. C.4.

Table A.10
Average subject-based testing results of LOSO for binary classification, using various TSI algorithm, exercises, and architecture.

		f1-micro									Specificity								
TSI	Architecture/Exercise	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
HC vs. PD1 (Subject-based)																			
GADF	InceptionV3	0.86	0.96	0.96	0.91	0.86	**1.00**	0.90	0.95	**1.00**	0.86	0.96	0.96	0.91	0.86	1.00	0.90	0.95	1.00
	MobileNetV2	0.91	**1.00**	**1.00**	**1.00**	0.86	**1.00**	0.86	0.77	**1.00**	0.91	1.00	1.00	1.00	0.86	1.00	0.86	0.77	1.00
	ResNet50V2	**1.00**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.90	0.95	**1.00**	1.00	1.00	1.00	0.96	1.00	1.00	0.90	0.95	1.00
	vgg16	0.86	0.91	0.91	**1.00**	0.82	0.91	**1.00**	**1.00**	**1.00**	0.86	0.91	0.91	1.00	0.82	0.91	1.00	1.00	1.00
GASF	InceptionV3	0.91	**1.00**	0.91	0.91	0.91	**1.00**	0.81	**1.00**	0.96	0.91	1.00	0.91	0.91	0.91	1.00	0.81	1.00	0.96
	MobileNetV2	0.91	**1.00**	0.96	0.78	0.95	0.96	**1.00**	0.86	**1.00**	0.91	1.00	0.96	0.78	0.95	0.96	1.00	0.86	1.00
	ResNet50V2	**1.00**	0.87	**1.00**	0.96	**1.00**	**1.00**	**1.00**	0.95	**1.00**	1.00	0.87	1.00	0.96	1.00	1.00	1.00	0.95	1.00
	vgg16	0.64	0.78	**1.00**	**1.00**	0.95	0.78	0.81	0.95	**1.00**	0.64	0.78	1.00	1.00	0.95	0.78	0.81	0.95	1.00
MTF	InceptionV3	0.86	0.96	**1.00**	**1.00**	0.86	**1.00**	0.90	**0.86**	**0.96**	0.86	0.96	1.00	1.00	0.86	1.00	0.90	0.86	0.96
	MobileNetV2	0.68	0.78	0.96	0.83	0.77	0.70	0.76	0.68	**0.96**	0.68	0.78	0.96	0.83	0.77	0.70	0.76	0.68	0.96
	ResNet50V2	**1.00**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	**0.95**	0.82	**0.96**	1.00	1.00	1.00	0.96	1.00	1.00	0.95	0.82	0.96
	vgg16	0.68	0.74	0.78	0.70	0.82	0.61	0.90	0.36	0.70	0.68	0.74	0.78	0.70	0.82	0.61	0.90	0.36	0.70
RP	InceptionV3	0.86	0.65	0.96	**0.91**	0.86	0.96	0.90	**0.91**	0.83	0.86	0.65	0.96	0.91	0.86	0.96	0.90	0.91	0.83
	MobileNetV2	0.86	0.61	0.87	**0.91**	0.86	0.91	0.90	0.82	0.65	0.86	0.61	0.87	0.91	0.86	0.91	0.90	0.82	0.65
	ResNet50V2	**1.00**	0.61	**1.00**	0.87	**0.91**	**1.00**	**0.95**	0.86	**1.00**	1.00	0.61	1.00	0.87	0.91	1.00	0.95	0.86	1.00
	vgg16	0.77	**0.78**	0.78	0.78	0.82	0.78	0.81	0.86	0.78	0.77	0.78	0.78	0.78	0.82	0.78	0.81	0.86	0.78
CWT	InceptionV3	0.95	0.83	0.87	0.78	**0.91**	**1.00**	**0.95**	0.95	0.91	0.95	0.83	0.87	0.78	0.91	1.00	0.95	0.95	0.91
	MobileNetV2	**1.00**	**0.96**	0.91	0.83	0.82	0.96	0.81	**1.00**	**1.00**	1.00	0.96	0.91	0.83	0.82	0.96	0.81	1.00	1.00
	ResNet50V2	**1.00**	**0.96**	**1.00**	**1.00**	0.86	**1.00**	0.90	0.95	**1.00**	1.00	0.96	1.00	1.00	0.86	1.00	0.90	0.95	1.00
	vgg16	0.77	0.83	0.78	0.78	0.73	0.78	0.81	0.41	0.96	0.77	0.83	0.78	0.78	0.73	0.78	0.81	0.41	0.96
HC vs. PD1.5 (Subject-based)																			
GADF	InceptionV3	**1.00**	0.86	0.95	0.81	0.95	**1.00**	**0.85**	0.95	**1.00**	1.00	0.86	0.95	0.81	0.95	1.00	0.85	0.95	1.00
	MobileNetV2	0.70	0.86	0.90	0.81	0.76	0.95	**0.85**	0.71	0.86	0.70	0.86	0.90	0.81	0.76	0.95	0.85	0.71	0.86
	ResNet50V2	**1.00**	**1.00**	**1.00**	**0.95**	0.86	**1.00**	0.70	**1.00**	**1.00**	1.00	1.00	1.00	0.95	0.86	1.00	0.70	1.00	1.00
	vgg16	**1.00**	**1.00**	**1.00**	0.86	**1.00**	0.90	**0.85**	0.95	0.86	1.00	1.00	1.00	0.86	1.00	0.90	0.85	0.95	0.86
GASF	InceptionV3	0.95	0.90	**0.95**	0.90	**0.95**	**1.00**	**0.85**	0.76	**1.00**	0.95	0.90	0.95	0.90	0.95	1.00	0.85	0.76	1.00
	MobileNetV2	0.90	0.90	0.76	0.90	0.71	**1.00**	0.70	0.76	0.76	0.90	0.90	0.76	0.90	0.71	1.00	0.70	0.76	0.76
	ResNet50V2	**1.00**	0.81	0.90	0.86	0.90	**1.00**	0.80	0.90	0.90	1.00	0.81	0.90	0.86	0.90	1.00	0.80	0.90	0.90
	vgg16	**1.00**	**0.95**	0.86	**1.00**	0.67	**1.00**	0.70	**0.95**	**1.00**	1.00	0.95	0.86	1.00	0.67	1.00	0.70	0.95	1.00
MTF	InceptionV3	0.95	0.81	0.95	0.86	0.90	0.95	0.85	0.71	**0.95**	0.95	0.81	0.95	0.86	0.90	0.95	0.85	0.71	0.95
	MobileNetV2	0.90	0.81	0.90	0.86	0.52	0.86	0.85	0.67	0.57	0.90	0.81	0.90	0.86	0.52	0.86	0.85	0.67	0.57
	ResNet50V2	**1.00**	**0.86**	**1.00**	0.81	**0.95**	**1.00**	**0.90**	**0.90**	0.90	1.00	0.86	1.00	0.81	0.95	1.00	0.90	0.90	0.90
	vgg16	0.95	0.57	0.57	**1.00**	0.62	0.86	0.80	0.14	0.24	0.95	0.57	0.57	1.00	0.62	0.86	0.80	0.14	0.24
RP	InceptionV3	0.90	0.67	**0.95**	0.81	0.81	**1.00**	0.90	0.86	0.81	0.90	0.67	0.95	0.81	0.81	1.00	0.90	0.86	0.81
	MobileNetV2	0.90	**0.90**	**0.95**	0.76	**0.86**	**1.00**	**0.95**	0.76	0.67	0.90	0.90	0.95	0.76	0.86	1.00	0.95	0.76	0.67
	ResNet50V2	**1.00**	0.86	0.86	**0.90**	**0.86**	**1.00**	0.85	**0.90**	**0.90**	1.00	0.86	0.86	0.90	0.86	1.00	0.85	0.90	0.90
	vgg16	0.85	0.38	0.86	0.86	**0.86**	0.86	0.85	0.86	0.86	0.85	0.38	0.86	0.86	0.86	0.86	0.85	0.86	0.86
CWT	InceptionV3	0.90	0.86	0.81	0.81	0.90	**1.00**	0.85	0.90	**1.00**	0.90	0.86	0.81	0.81	0.90	1.00	0.85	0.90	1.00
	MobileNetV2	0.85	**0.95**	0.90	0.76	0.86	0.86	0.85	0.62	0.81	0.85	0.95	0.90	0.76	0.86	0.86	0.85	0.62	0.81
	ResNet50V2	**1.00**	0.86	**1.00**	0.71	**0.95**	**1.00**	**0.90**	0.90	0.95	1.00	0.86	1.00	0.71	0.95	1.00	0.90	0.90	0.95
	vgg16	0.85	**0.95**	0.86	**0.86**	0.67	0.86	0.85	**0.95**	0.38	0.85	0.95	0.86	0.86	0.67	0.86	0.85	0.95	0.38
PD1 vs. PD1.5 (Subject-based)																			
GADF	InceptionV3	**1.00**	0.71	**0.71**	**1.00**	0.83	0.71	0.83	**0.50**	0.71	1.00	0.71	0.71	1.00	0.83	0.71	0.83	0.50	0.71
	MobileNetV2	0.71	0.57	0.43	0.57	0.33	0.71	0.67	**0.50**	0.71	0.71	0.57	0.43	0.57	0.33	0.71	0.67	0.50	0.71
	ResNet50V2	**1.00**	0.86	**0.71**	**1.00**	**1.00**	**0.86**	0.83	**0.50**	0.86	1.00	0.86	0.71	1.00	1.00	0.86	0.83	0.50	0.86
	vgg16	0.86	**1.00**	0.57	0.71	0.83	**0.86**	**1.00**	0.33	**1.00**	0.86	1.00	0.57	0.71	0.83	0.86	1.00	0.33	1.00
GASF	InceptionV3	**1.00**	**0.71**	**0.71**	**1.00**	0.67	**0.86**	**0.83**	**0.67**	**0.86**	1.00	0.71	0.71	1.00	0.67	0.86	0.83	0.67	0.86
	MobileNetV2	0.57	0.29	0.43	0.43	0.00	0.71	0.33	0.50	0.43	0.57	0.29	0.43	0.43	0.00	0.71	0.33	0.50	0.43
	ResNet50V2	**1.00**	0.57	**0.71**	**1.00**	0.67	**0.86**	**0.83**	0.50	0.71	1.00	0.57	0.71	1.00	0.67	0.86	0.83	0.50	0.71
	vgg16	**1.00**	0.29	0.43	0.57	**0.83**	0.71	0.50	0.17	0.57	1.00	0.29	0.43	0.57	0.83	0.71	0.50	0.17	0.57
MTF	InceptionV3	0.86	**0.57**	**0.86**	**1.00**	**1.00**	**0.86**	0.50	0.50	**0.71**	0.86	0.57	0.86	1.00	1.00	0.86	0.50	0.50	0.71
	MobileNetV2	0.71	0.29	0.29	0.57	0.50	0.43	0.50	**0.67**	0.57	0.71	0.29	0.29	0.57	0.50	0.43	0.50	0.67	0.57
	ResNet50V2	**1.00**	0.43	0.43	0.86	0.67	**0.86**	0.33	0.50	**0.71**	1.00	0.43	0.43	0.86	0.67	0.86	0.33	0.50	0.71
	vgg16	0.43	**0.57**	0.57	0.43	0.33	0.43	**1.00**	0.33	0.43	0.43	0.57	0.57	0.43	0.33	0.43	1.00	0.33	0.43
RP	InceptionV3	**1.00**	0.57	**0.86**	0.86	**0.67**	**0.57**	**0.83**	0.50	**0.71**	1.00	0.57	0.86	0.86	0.67	0.57	0.83	0.50	0.71
	MobileNetV2	0.43	0.43	0.71	0.71	0.50	**0.57**	0.17	0.50	**0.71**	0.43	0.43	0.71	0.71	0.50	0.57	0.17	0.50	0.71
	ResNet50V2	**1.00**	0.57	0.71	**1.00**	0.50	**0.57**	0.50	**0.67**	0.43	1.00	0.57	0.71	1.00	0.50	0.57	0.50	0.67	0.43
	vgg16	0.57	**0.71**	0.57	0.57	0.50	**0.57**	0.50	**0.67**	0.43	0.57	0.71	0.57	0.57	0.50	0.57	0.50	0.67	0.43
CWT	InceptionV3	**1.00**	0.43	**0.71**	0.71	**0.83**	**0.86**	**0.67**	0.50	**0.71**	1.00	0.43	0.71	0.71	0.83	0.86	0.67	0.50	0.71
	MobileNetV2	0.71	0.43	0.57	0.57	0.33	0.43	0.50	**0.67**	0.43	0.71	0.43	0.57	0.57	0.33	0.43	0.50	0.67	0.43
	ResNet50V2	0.86	**0.71**	0.57	**0.86**	0.33	0.71	**0.67**	**0.67**	0.43	0.86	0.71	0.57	0.86	0.33	0.71	0.67	0.67	0.43
	vgg16	0.57	0.57	0.57	0.57	0.50	0.57	0.50	0.50	0.57	0.57	0.57	0.57	0.57	0.50	0.57	0.50	0.50	0.57

Results are aggregated across subjects, each performing 9 exercises, with final subject-level labels determined by majority voting of predictions from all the frames.

Table A.11
Average subject-based testing results of LOSO for multi-classification tasks.

		f1-micro									Specificity								
TSI	Arch/Ex	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
HC vs. PD1 vs. PD1.5 vs. PD2 (Subject-based)																			
GADF	InceptionV3	0.65	0.80	**0.90**	0.88	0.69	0.93	**0.84**	0.81	0.89	0.88	0.93	0.97	0.96	0.90	0.98	0.95	0.94	0.96
	MobileNetV2	0.95	0.80	0.86	0.95	0.62	0.93	**0.84**	0.68	**0.92**	0.98	0.93	0.95	0.98	0.87	0.98	0.95	0.89	0.97
	ResNet50V2	**0.98**	**0.85**	0.83	**1.00**	**0.72**	**0.98**	**0.84**	**0.95**	0.84	0.99	0.95	0.94	1.00	0.91	0.99	0.95	0.98	0.95
GASF	InceptionV3	0.95	**0.90**	**0.93**	0.90	0.64	0.75	0.78	**0.89**	0.79	0.98	0.97	0.98	0.97	0.88	0.92	0.93	0.96	0.93
	MobileNetV2	0.95	0.73	0.76	0.98	0.69	0.95	**0.84**	0.84	**0.84**	0.98	0.91	0.92	0.99	0.90	0.98	0.95	0.95	0.95
	ResNet50V2	**0.98**	0.83	0.86	**1.00**	**0.79**	**0.98**	0.76	**0.89**	**0.84**	0.99	0.94	0.95	1.00	0.93	0.99	0.92	0.96	0.95
	vgg16	0.65	0.59	0.52	0.50	0.49	0.60	0.59	0.78	0.66	0.88	0.86	0.84	0.83	0.83	0.87	0.86	0.93	0.89
MTF	InceptionV3	0.80	0.80	0.88	0.88	0.77	0.73	**0.89**	**0.89**	**0.82**	0.93	0.93	0.96	0.96	0.92	0.91	0.96	0.96	0.94
	MobileNetV2	0.93	0.56	0.86	**0.90**	0.51	0.83	0.65	0.70	0.79	0.98	0.85	0.95	0.97	0.84	0.94	0.88	0.90	0.93
	ResNet50V2	**0.98**	**0.88**	**0.93**	0.68	**0.82**	**0.98**	0.84	0.76	0.71	0.99	0.96	0.98	0.89	0.94	0.99	0.95	0.92	0.90
	vgg16	0.63	0.78	0.24	0.48	0.23	0.53	0.57	0.78	0.16	0.88	0.93	0.75	0.83	0.74	0.84	0.86	0.93	0.72
RP	InceptionV3	0.78	0.68	0.69	0.65	**0.72**	**0.80**	**0.84**	**0.73**	**0.89**	0.93	0.89	0.90	0.88	0.91	0.93	0.95	0.91	0.96
	MobileNetV2	0.93	0.66	**0.81**	**0.70**	0.51	0.65	0.51	**0.73**	0.74	0.98	0.89	0.94	0.90	0.84	0.88	0.84	0.91	0.91
	ResNet50V2	**0.98**	**0.85**	**0.81**	**0.70**	0.59	**0.80**	**0.84**	**0.73**	0.74	0.99	0.95	0.94	0.90	0.86	0.93	0.95	0.91	0.91
	vgg16	0.45	0.27	0.45	0.48	0.49	0.48	0.49	0.51	0.50	0.82	0.76	0.82	0.83	0.83	0.83	0.83	0.84	0.83
CWT	InceptionV3	0.83	0.78	0.86	0.73	0.72	**0.93**	**0.86**	**0.76**	0.74	0.94	0.93	0.95	0.91	0.91	0.98	0.95	0.92	0.91
	MobileNetV2	0.93	0.56	0.83	0.70	0.51	0.70	0.68	0.65	0.76	0.98	0.85	0.94	0.90	0.84	0.90	0.89	0.88	0.92
	ResNet50V2	**0.98**	**0.90**	**0.88**	**0.98**	**0.82**	**0.93**	0.81	**0.76**	**0.82**	0.99	0.97	0.96	0.99	0.94	0.98	0.94	0.92	0.94
	vgg16	0.45	0.37	0.45	0.48	0.23	0.48	0.49	0.43	0.34	0.82	0.79	0.82	0.83	0.74	0.83	0.83	0.81	0.78

Table B.12
The feature set and parameter space used with random forest.

(a)

Domain	Features	Count
Time	Mean, Standard deviation, Median, Minimum, Maximum, Range, RMS, Skewness, Kurtosis, Zero-crossing rate	10
Frequency	Spectral centroid, Spectral bandwidth, Dominant frequency 1, Dominant frequency 2, Spectral flatness, Band power 0.5–3 Hz, Band power 3–7 Hz, Band power 7–12 Hz, Band power 12–20 Hz, Total spectral power	10

- n_estimators = [100, 200, 500], max_depth = [None, 5, 10, 20].
- min_samples_split = [2, 5, 10], min_samples_leaf = [1, 2, 4]).
- Score metric for the best-model selection - F1_macro was used.

Table B.13
Random forest performance.

(a)

	Precision	Recall	Specificity	F1	Precision	Recall	Specificity	F1
	H				PD1.0			
1	1.000 (0.981–1.000)	1.000 (0.981–1.000)	1.000 (0.923–1.000)	1.000	1.000 (0.923–1.000)	1.000 (0.923–1.000)	1.000 (0.981–1.000)	1.000
2	0.920 (0.889–0.943)	1.000 (0.990–1.000)	0.663 (0.563–0.750)	0.958	1.000 (0.943–1.000)	0.663 (0.563–0.750)	1.000 (0.990–1.000)	0.797
3	0.990 (0.963–0.997)	1.000 (0.980–1.000)	0.958 (0.860–0.988)	0.995	1.000 (0.923–1.000)	0.958 (0.860–0.988)	1.000 (0.980–1.000)	0.979
4	0.923 (0.888–0.948)	1.000 (0.986–1.000)	0.324 (0.191–0.492)	0.960	1.000 (0.741–1.000)	0.324 (0.191–0.492)	1.000 (0.986–1.000)	0.489
5	0.926 (0.888–0.951)	1.000 (0.985–1.000)	0.643 (0.512–0.755)	0.961	1.000 (0.904–1.000)	0.643 (0.512–0.755)	1.000 (0.985–1.000)	0.783
6	0.987 (0.931–0.998)	1.000 (0.952–1.000)	0.929 (0.685–0.987)	0.994	1.000 (0.772–1.000)	0.929 (0.685–0.987)	1.000 (0.952–1.000)	0.963
8	0.996 (0.977–0.999)	1.000 (0.985–1.000)	0.980 (0.897–0.997)	0.998	1.000 (0.929–1.000)	0.980 (0.897–0.997)	1.000 (0.985–1.000)	0.990
9	0.996 (0.978–0.999)	1.000 (0.985–1.000)	0.984 (0.915–0.997)	0.998	1.000 (0.942–1.000)	0.984 (0.915–0.997)	1.000 (0.985–1.000)	0.992
	PwPD1.0				PwPD1.5			
1	0.667 (0.417–0.848)	1.000 (0.722–1.000)	0.000 (0.000–0.434)	0.800	0.000 (0.000–0.000)	0.000 (0.000–0.434)	1.000 (0.722–1.000)	0.000
2	1.000 (0.929–1.000)	1.000 (0.929–1.000)	1.000 (0.920–1.000)	1.000	1.000 (0.920–1.000)	1.000 (0.920–1.000)	1.000 (0.929–1.000)	1.000
3	1.000 (0.975–1.000)	0.987 (0.954–0.996)	1.000 (0.968–1.000)	0.993	0.983 (0.940–0.995)	1.000 (0.968–1.000)	0.987 (0.954–0.996)	0.991
4	0.880 (0.700–0.958)	1.000 (0.851–1.000)	0.727 (0.434–0.903)	0.936	1.000 (0.676–1.000)	0.727 (0.434–0.903)	1.000 (0.851–1.000)	0.842
5	1.000 (0.914–1.000)	1.000 (0.914–1.000)	1.000 (0.908–1.000)	1.000	1.000 (0.908–1.000)	1.000 (0.908–1.000)	1.000 (0.914–1.000)	1.000
6	1.000 (0.722–1.000)	1.000 (0.722–1.000)	1.000 (0.701–1.000)	1.000	1.000 (0.701–1.000)	1.000 (0.701–1.000)	1.000 (0.722–1.000)	1.000
7	1.000 (0.910–1.000)	1.000 (0.910–1.000)	1.000 (0.916–1.000)	1.000	1.000 (0.916–1.000)	1.000 (0.916–1.000)	1.000 (0.910–1.000)	1.000
8	1.000 (0.908–1.000)	1.000 (0.908–1.000)	1.000 (0.912–1.000)	1.000	1.000 (0.912–1.000)	1.000 (0.912–1.000)	1.000 (0.908–1.000)	1.000
9	1.000 (0.908–1.000)	1.000 (0.908–1.000)	1.000 (0.883–1.000)	1.000	1.000 (0.883–1.000)	1.000 (0.883–1.000)	1.000 (0.908–1.000)	1.000

(b)

Ex	Precision_{micro}	Recall_{micro}	F1_{micro}	Precision_{macro}	Recall_{macro}	F1_{macro}
HC vs. PwPD1.0						
1	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
2	0.931 (0.914–0.946)	0.931 (0.915–0.948)	0.931 (0.914–0.946)	0.950 (0.939–0.962)	0.865 (0.830–0.898)	0.896 (0.867–0.922)
3	0.992 (0.983–0.998)	0.992 (0.983–0.998)	0.992 (0.983–0.998)	0.994 (0.987–1.000)	0.983 (0.961–0.999)	0.988 (0.974–0.998)
4	0.926 (0.905–0.945)	0.926 (0.905–0.945)	0.926 (0.904–0.944)	0.950 (0.936–0.964)	0.750 (0.698–0.806)	0.792 (0.720–0.851)
5	0.934 (0.915–0.954)	0.934 (0.913–0.952)	0.934 (0.915–0.954)	0.953 (0.939–0.967)	0.859 (0.815–0.903)	0.893 (0.854–0.927)
6	0.989 (0.973–1.000)	0.989 (0.973–1.000)	0.989 (0.973–1.000)	0.992 (0.980–1.000)	0.973 (0.923–1.000)	0.982 (0.947–1.000)
8	0.997 (0.992–1.000)	0.997 (0.992–1.000)	0.997 (0.992–1.000)	0.998 (0.994–1.000)	0.992 (0.977–1.000)	0.995 (0.985–1.000)
9	0.997 (0.992–1.000)	0.997 (0.992–1.000)	0.997 (0.992–1.000)	0.998 (0.994–1.000)	0.994 (0.981–1.000)	0.996 (0.988–1.000)
PwPD1.0 vs. PwPD1.5						
1	0.667 (0.500–0.833)	0.667 (0.500–0.833)	0.667 (0.500–0.833)	0.444 (0.332–0.557)	0.556 (0.467–0.633)	0.489 (0.378–0.595)
2	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
3	0.993 (0.985–0.998)	0.993 (0.985–0.998)	0.993 (0.985–0.998)	0.992 (0.983–0.999)	0.993 (0.986–0.999)	0.992 (0.985–0.998)
4	0.909 (0.848–0.970)	0.909 (0.833–0.970)	0.909 (0.833–0.970)	0.930 (0.875–0.977)	0.879 (0.770–0.968)	0.896 (0.801–0.968)
5	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
6	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
7	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
8	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
9	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)

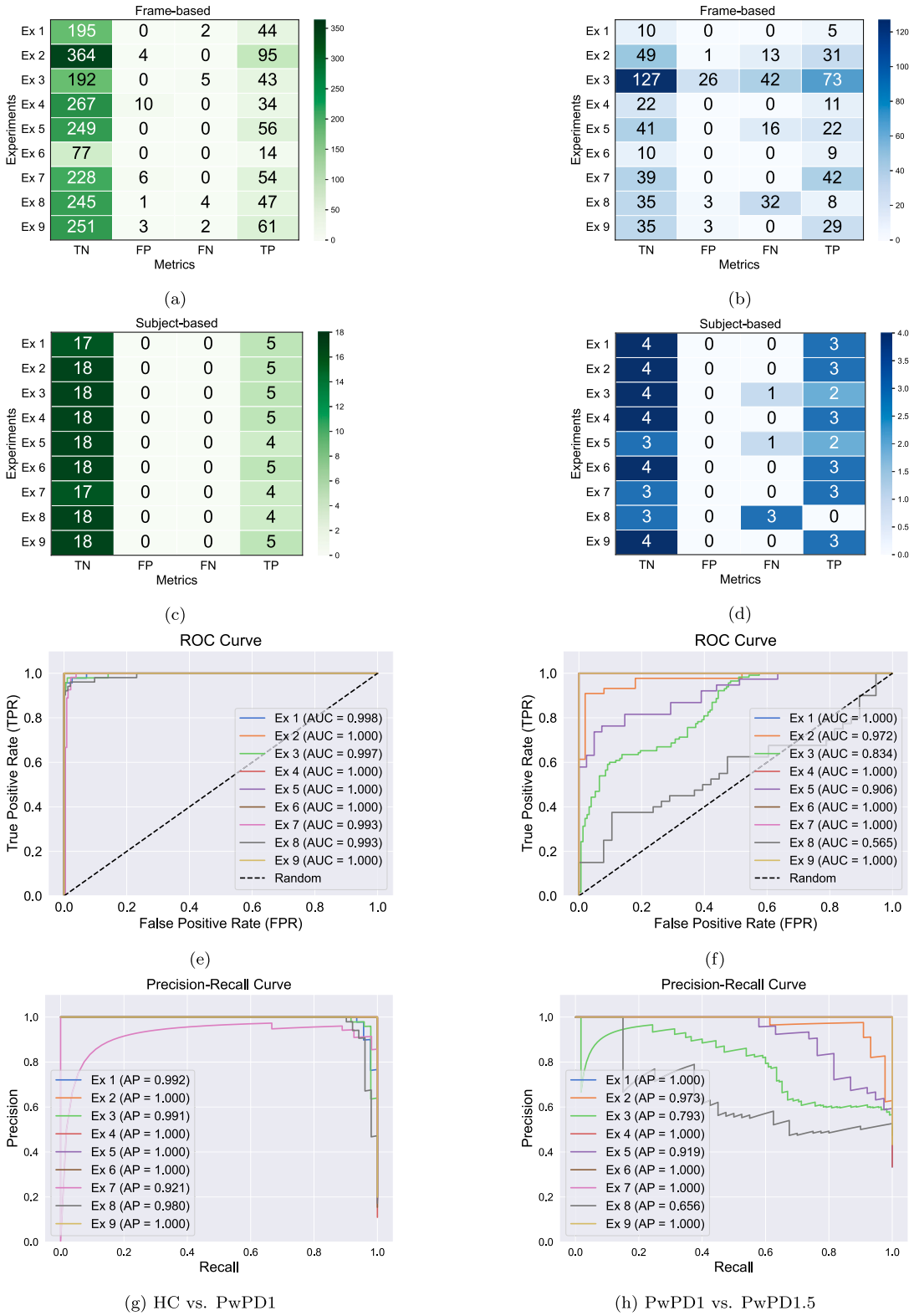


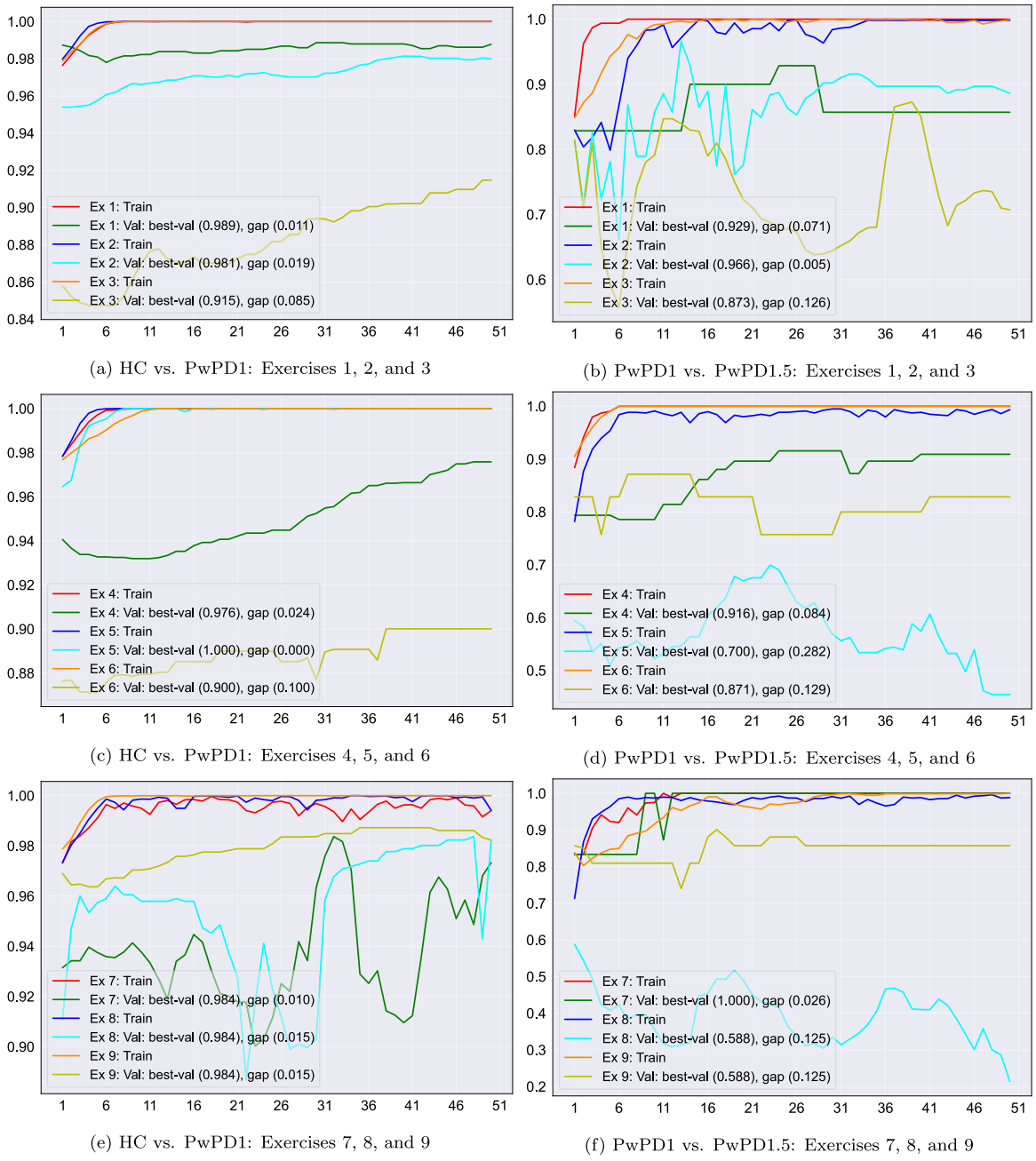
Fig. C.4. ROC, PR curves (frame-based), and Confusion matrices (frame-based and subject-based) for both HC vs. PwPD1 (the first column) and PwPD1 vs. PwPD1.5 (the second column).

C.2. Training/validation curves

See Fig. C.5.

Data availability

Data will be made available on request.



(a) HC vs. PwPD1: Exercises 1, 2, and 3

(b) PwPD1 vs. PwPD1.5: Exercises 1, 2, and 3

(c) HC vs. PwPD1: Exercises 4, 5, and 6

(d) PwPD1 vs. PwPD1.5: Exercises 4, 5, and 6

(e) HC vs. PwPD1: Exercises 7, 8, and 9

(f) PwPD1 vs. PwPD1.5: Exercises 7, 8, and 9

Fig. C.5. Training/Validation curves in terms of F1-score, where the gap represents the training-validation gap.

References

[1] J. Jankovic, Parkinson's disease: clinical features and diagnosis, J. Neurol. Neurosurg. Psychiatry 79 (4) (2008) 368–376.

[2] O.-B. Tysnes, A. Storstein, Epidemiology of Parkinson's disease, J. Neural Transm. 124 (2017) 901–905.

[3] E.R. Dorsey, B.R. Bloem, The Parkinson pandemic—a call to action, JAMA Neurol. 75 (1) (2018) 9–10.

[4] H. Abujrida, E. Agu, K. Pahlavan, Machine learning-based motor assessment of Parkinson's disease using postural sway, gait and lifestyle features on crowd-sourced smartphone data, Biomed. Phys. Eng. Express 6 (3) (2020) 035005, http://dx.doi.org/10.1088/2057-1976/ab39a8.

[5] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, G. Logroscino, Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis, Neurology 86 (6) (2016) 566–576.

[6] D. Buongiorno, I. Bortone, G.D. Cascarano, G.F. Trotta, A. Brunetti, V. Bevilacqua, A low-cost vision system based on the analysis of motor features for recognition and severity rating of Parkinson's Disease, BMC Med. Inform. Decis. Mak. 19 (2019) 1–13.

[7] B. Kashyap, D. Phan, P.N. Pathirana, M. Horne, L. Power, D. Szmulewicz, Objective assessment of cerebellar ataxia: A comprehensive and refined approach, Sci. Rep. 10 (1) (2020) 1–17, http://dx.doi.org/10.1038/s41598-020-65303-7.

[8] X. Cui, Y. Zhou, C. Zhao, J. Li, X. Zheng, X. Li, S. Shan, J.-X. Liu, X. Liu, A multiscale hybrid attention networks based on multiview images for the diagnosis of Parkinson's disease, IEEE Trans. Instrum. Meas. 73 (2024) 1–11, http://dx.doi.org/10.1109/TIM.2023.3315407.

[9] A. Papadopoulos, D. Iakovakis, L. Klingelhoefer, S. Bostantjopoulou, K.R. Chaudhuri, K. Kyritsis, S. Hadjidimitriou, V. Charisis, L.J. Hadjileontiadis, A. Delopoulos, Unobtrusive detection of Parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques, Sci. Rep. 10 (1) (2020) 21370.

[10] E. Samadi, H. Ahmadi, F.N. Rahatabad, Analysis of hand tremor in Parkinson's disease: Frequency domain approach, Front. Biomed. Technol. 7 (2) (2020) 105–111.

[11] J. Varghese, M. Fujarski, T. Hahn, M. Dugas, T. Warnecke, The smart device system for movement disorders: preliminary evaluation of diagnostic accuracy in a prospective study, in: Digital Personalized Health and Medicine, IOS Press, 2020, pp. 889–893.

[12] N. Li, F. Tian, X. Fan, Y. Zhu, H. Wang, G. Dai, Monitoring motor symptoms in Parkinson's disease via instrumenting daily artifacts with inertia sensors, CCF Trans. Pervasive Comput. Interact. 1 (2019) 100–113.

[13] M. Hammoud, E. Kovalenko, A. Somov, E. Bril, A. Baldycheva, Deep learning framework for neurological diseases diagnosis through near-infrared eye video and time series imaging algorithms, Internet Things 24 (2023) 100914, http://dx.doi.org/10.1016/j.iot.2023.100914.

[14] Z. Yin, V.J. Geraedts, Z. Wang, M.F. Contarino, H. Dibeklioglu, J. Van Gemert, Assessment of Parkinson's disease severity from videos using deep architectures, IEEE J. Biomed. Health Informatics 26 (3) (2021) 1164–1176.

[15] R. Sun, Y. Ren, A multi-source heterogeneous data fusion method for intelligent systems in the Internet of Things, Intell. Syst. Appl. 23 (2024) 200424, http://dx.doi.org/10.1016/j.iswa.2024.200424.

[16] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

[17] M.B. Makarious, H.L. Leonard, D. Vitale, H. Iwaki, L. Sargent, A. Dadu, I. Violich, E. Hutchins, D. Saffo, S. Bandres-Ciga, et al., Multi-modality machine learning predicting Parkinson's disease, Npj Parkinson's Dis. 8 (1) (2022) 35.

[18] S. Rupprechter, G. Morinan, Y. Peng, T. Foltynie, K. Sibley, R.S. Weil, L.-A. Leyland, F. Baig, F. Morgante, R. Gilron, et al., A clinically interpretable computer-vision based method for quantifying gait in Parkinson's disease, Sensors 21 (16) (2021) 5437.

[19] M.S.R. Sajal, M.T. Ehsan, R. Vaidyanathan, S. Wang, T. Aziz, K.A.A. Mamun, Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis, Brain Informatics 7 (1) (2020) 1–11.

[20] A. Muro-De-La-Herran, B. Garcia-Zapirain, A. Mendez-Zorrilla, Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications, Sensors 14 (2) (2014) 3362–3394.

[21] M. Hammoud, A. Shcherbak, M. Getahun, O. Istrakova, N. Shindryaeva, O. Zimniakova, E. Bril, M. Semenov, A. Baldycheva, A. Somov, Wearable sensors and machine learning fusion for Parkinson's disease assessment, in: 2024 IEEE International Instrumentation and Measurement Technology Conference, I2MTC, 2024, pp. 1–6, http://dx.doi.org/10.1109/I2MTC60896.2024.10561139.

[22] W. Wang, J. Lee, F. Harrou, Y. Sun, Early detection of Parkinson's disease using deep learning and machine learning, IEEE Access 8 (2020) 147635–147646.

[23] N. Shawen, M.K. O'Brien, S. Venkatesan, L. Lonini, T. Simuni, J.L. Hamilton, R. Ghaffari, J.A. Rogers, A. Jayaraman, Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors, J. Neuroeng. Rehabil. 17 (2020) 1–14, http://dx.doi.org/10.1186/s12984-020-00684-4.

[24] S.-S.M. Ajibade, G.N. Alhassan, A. Zaidi, O.A. Oki, J.B. Awotunde, E. Ogbuju, K.A. Akintoye, Evolution of machine learning applications in medical and healthcare analytics research: A bibliometric analysis, Intell. Syst. Appl. 24 (2024) 200441, http://dx.doi.org/10.1016/j.iswa.2024.200441, URL https://www.sciencedirect.com/science/article/pii/S2667305324001157.

[25] C. Sotirakis, Z. Su, M.A. Brzezicki, N. Conway, L. Tarassenko, J.J. FitzGerald, C.A. Antoniades, Identification of motor progression in Parkinson's disease using wearable sensors and machine learning, Npj Parkinson's Dis. 9 (1) (2023) 142.

[26] L. Mesin, P. Porcu, D. Russu, G. Farina, L. Borzì, W. Zhang, Y. Guo, G. Olmo, A multi-modal analysis of the freezing of gait phenomenon in Parkinson's disease, Sensors 22 (7) (2022) 2613.

[27] L. Lonini, A. Dai, N. Shawen, T. Simuni, C. Poon, L. Shimanovich, M. Daeschler, R. Ghaffari, J.A. Rogers, A. Jayaraman, Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models, NPJ Digit. Med. 1 (1) (2018) 64.

[28] A. Shcherbak, E. Kovalenko, E. Bril, A. Baldycheva, A. Somov, Dominant hand invariant Parkinson's disease detection using 1-D CNN model and STFT-based IMU data fusion, IEEE ISIE 2023 (2023) http://dx.doi.org/10.1109/ISIE51358.2023.10228119.

[29] A. Talitckii, A. Anikina, E. Kovalenko, A. Shcherbak, O. Mayora, O. Zimniakova, E. Bril, M. Semenov, D.V. Dylov, A. Somov, Defining optimal exercises for efficient detection of Parkinson's disease using machine learning and wearable sensors, IEEE Trans. Instrum. Meas. 70 (2021) 1–10, http://dx.doi.org/10.1109/TIM.2021.3097857.

[30] A. Shcherbak, E. Kovalenko, A. Somov, Detection and classification of early stages of Parkinson's disease through wearable sensors and machine learning, IEEE Trans. Instrum. Meas. 72 (2023) 1–9, http://dx.doi.org/10.1109/TIM.2023.3284944.

[31] E. Kovalenko, A. Shcherbak, A. Somov, E. Bril, O. Zimniakova, M. Semenov, A. Samoylov, Detecting the Parkinson's disease through the simultaneous analysis of data from wearable sensors and video, IEEE Sensors J. 22 (16) (2022) 16430–16439, http://dx.doi.org/10.1109/JSEN.2022.3191864.

[32] E. Kovalenko, A. Talitckii, A. Anikina, A. Shcherbak, O. Zimniakova, M. Semenov, E. Bril, D.V. Dylov, A. Somov, Distinguishing between Parkinson's disease and essential tremor through video analytics using machine learning: A pilot study, IEEE Sensors J. 21 (10) (2020) 11916–11925, http://dx.doi.org/10.1109/JSEN.2020.3035240.

[33] A. Talitckii, E. Kovalenko, A. Shcherbak, A. Anikina, E. Bril, O. Zimniakova, M. Semenov, D.V. Dylov, A. Somov, Comparative study of wearable sensors, video, and handwriting to detect Parkinson's disease, IEEE Trans. Instrum. Meas. 71 (2022) 1–10, http://dx.doi.org/10.1109/TIM.2022.3176898.

[34] N. Hatami, Y. Gavet, J. Debayle, Classification of time-series images using deep convolutional neural networks, in: Tenth International Conference on Machine Vision, ICMV 2017, Vol. 10696, SPIE, 2018, pp. 242–249.

[35] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, A. Etemad, Classification of hand movements from EEG using a deep attention-based LSTM network, IEEE Sensors J. 20 (6) (2019) 3113–3122.

[36] N. Phutela, D. Relan, G. Gabrani, P. Kumaraguru, M. Samuel, Stress classification using brain signals based on LSTM network, Comput. Intell. Neurosci. 2022 (2022).

[37] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G.S. Choi, B.-W. On, Heartbeat sound signal classification using deep learning, Sensors 19 (21) (2019) 4819.

[38] P.C. Vakkantula, Speech Mode Classification using the Fusion of CNNs and LSTM Networks, West Virginia University, 2020.

[39] Z. Wang, T. Oates, Imaging time-series to improve classification and imputation, 2015, arXiv preprint arXiv:1506.00327.

[40] L.-M. Vortmann, F. Putze, Combining implicit and explicit feature extraction for eye tracking: Attention classification using a heterogeneous input, Sensors 21 (24) (2021) 8205.

[41] K.P. Thanaraj, B. Parvathavarthini, U.J. Tanik, V. Rajinikanth, S. Kadry, K. Kamalanand, Implementation of deep neural networks to classify EEG signals using gramian angular summation field for epilepsy diagnosis, 2020, arXiv preprint arXiv:2003.04534.

[42] B. Bertalanič, M. Meža, C. Fortuna, Resource-aware time series imaging classification for wireless link layer anomalies, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[43] C.-L. Yang, Z.-X. Chen, C.-Y. Yang, Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images, Sensors 20 (1) (2019) 168.

[44] J. Yan, J. Kan, H. Luo, Rolling bearing fault diagnosis based on Markov transition field and residual network, Sensors 22 (10) (2022) 3936.

[45] R. Bhidayasiri, D. Tarsy, R. Bhidayasiri, D. Tarsy, Parkinson's disease: Hoehn and yahr scale, in: Movement disorders: a video atlas: a video atlas, Springer, 2012, pp. 4–5, http://dx.doi.org/10.1007/978-1-60327-426-5_2.

[46] 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, Mov. Disord.: Off. J. Mov. Disord. Soc. 23 (15) (2008) 2129–2170.

[47] P.Y. Chan, Z.M. Ripin, S.A. Halim, W.N. Arifin, A.S. Yahya, G.B. Eow, K. Tan, J.Y. Hor, C.K. Wong, Motion characteristics of subclinical tremors in Parkinson's disease and normal subjects, Sci. Rep. 12 (1) (2022) 4021.

[48] A. Channa, R.-C. Ifrim, D. Popescu, N. Popescu, A-WEAR bracelet for detection of hand tremor and bradykinesia in Parkinson's patients, Sensors 21 (3) (2021) 981, http://dx.doi.org/10.3390/s21030981.

[49] M.D. Hssayeni, J. Jimenez-Shahed, M.A. Burack, B. Ghoraani, Wearable sensors for estimation of Parkinsonian tremor severity during free body movements, Sensors 19 (19) (2019) 4215.

[50] J. Fujikawa, R. Morigaki, N. Yamamoto, H. Nakanishi, T. Oda, Y. Izumi, Y. Takagi, Diagnosis and treatment of tremor in Parkinson's disease using mechanical devices, Life 13 (1) (2022) 78.

[51] D. Pan, R. Dhall, A. Lieberman, D.B. Petitti, et al., A mobile cloud-based Parkinson's disease assessment system for home-based monitoring, JMIR MHealth UHealth 3 (1) (2015) e3956, http://dx.doi.org/10.3390/s23115212.

[52] P.-K. Yang, B. Filtjens, P. Ginis, M. Goris, A. Nieuwboer, M. Gilat, P. Slaets, B. Vanrumste, Freezing of gait assessment with inertial measurement units and deep learning: effect of tasks, medication states, and stops, J. NeuroEngineering Rehabil. 21 (1) (2024) 24.

[53] R. San-Segundo, A. Zhang, A. Cebulla, S. Panev, G. Tabor, K. Stebbins, R.E. Massa, A. Whitford, F. De la Torre, J. Hodgins, Parkinson's disease tremor detection in the wild using wearable accelerometers, Sensors 20 (20) (2020) 5817.

[54] B. Li, Z. Yao, J. Wang, S. Wang, X. Yang, Y. Sun, Improved deep learning technique to detect freezing of gait in Parkinson's disease based on wearable sensors, Electronics 9 (11) (2020) 1919.

[55] M. Chen, Z. Sun, T. Xin, Y. Chen, F. Su, An interpretable deep learning optimized wearable daily detection system for Parkinson's disease, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 3937–3946.

[56] M. Hammoud, A. Shcherbak, M. Getahun, O. Istrakova, N. Shindryaeva, O. Zimniakova, E. Bril, M. Semenov, A. Baldycheva, A. Somov, Wearable sensors and machine learning fusion for Parkinson's disease assessment, in: 2024 IEEE International Instrumentation and Measurement Technology Conference, I2MTC, IEEE, 2024, pp. 1–6.

[57] L. Aguiar-Conraria, M.J. Soares, The Continuous Wavelet Transform: A Primer, Tech. rep., NIPE-Universidade do Minho, 2011.

[58] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[59] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[60] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.

[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[62] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.

[63] R. Roelofs, N. Cain, J. Shlens, M.C. Mozer, Mitigating bias in calibration error estimation, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 4036–4054.

[64] M. Hammoud, A. Shcherbak, M. Getahun, O. Istrakova, N. Shindryaeva, O. Zimniakova, E. Bril, M. Semenov, A. Baldycheva, A. Somov, Wearable sensors and machine learning fusion for Parkinson's disease assessment, in: 2024 IEEE International Instrumentation and Measurement Technology Conference, I2MTC, 2024, pp. 1–6, http://dx.doi.org/10.1109/I2MTC60896.2024.10561139.

[65] M. Hammoud, A. Shcherbak, O. Istrakova, N. Shindryaeva, E. Bril, R. Passerone, A. Somov, Wrist-worn sensors and machine learning for Parkinson's disease detection: Investigation of binary and multi-classification problem, IEEE Trans. Instrum. Meas. (2025).

[66] M.M. Hoehn, M.D. Yahr, Parkinsonism: onset, progression, and mortality, Neurology 17 (5) (1967) 427.

[67] L.M. Shulman, A.L. Gruber-Baldini, K.E. Anderson, C.G. Vaughan, S.G. Reich, P.S. Fishman, W.J. Weiner, The evolution of disability in Parkinson disease, Mov. Disorders 23 (6) (2008) 790–796.

[68] R.B. Postuma, D. Berg, M. Stern, W. Poewe, C.W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A.E. Lang, et al., MDS clinical diagnostic criteria for Parkinson's disease, Mov. Disorders 30 (12) (2015) 1591–1601, http://dx.doi.org/10.1002/mds.26424.