VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 1, January-March 2026 *pp:350-365*

A Simple and Reproducible Machine Learning Pipeline for Parkinson's Disease Detection Using Smartwatch-Based Inertial Signals

Quan Vu[1*], **Manh-Cuong Nguyen**[1], **Duc-Tan Tran**[2], **Vijender Kumar Solanki**[3]

[1]Department of Biomedical Engineering, Institute of Control Engineering, Le Quy Don Technical University, Hanoi, Vietnam; [2]Faculty of Electrical and Electronic Engineering, Phenikaa School of Engineering, Phenikaa University, Hanoi, Vietnam; [3]Department of Artificial Intelligence and Data Science, Stanley College of Engineering & Technology for Women, Hyderabad, Telangana, India

Keywords: *Parkinson disease; wearable sensors; inertial signals; machine learning; tremor detection*

Journal Info:
Submitted:
December 19, 2025
Accepted:
March 26, 2026
Published:
March 31, 2026

Abstract Parkinson's disease (PD) is a progressive neurological disorder characterized by motor symptoms such as tremor, rigidity, and bradykinesia. Wearable inertial sensors enable non-invasive and cost-effective assessment of motor abnormalities in real-world settings. Despite recent advances in deep learning, many existing approaches rely on complex architectures with limited interpretability and inconsistent evaluation protocols. This study proposes a simple and reproducible classical machine learning pipeline for PD detection using smartwatch-based inertial signals from the PADS dataset. Spectral and statistical features were extracted from accelerometer and gyroscope signals, and LASSO-based feature selection was applied within a nested subject-level cross-validation framework to prevent data leakage. Several classifiers, including Logistic Regression, Support Vector Machine, Random Forest, CatBoost, and Multi-Layer Perceptron, were evaluated. The proposed pipeline achieved 79.26% balanced accuracy, 87.32% accuracy, and an F1-score of 0.92 for PD vs. healthy control classification, while 67.15% balanced accuracy was obtained for the more challenging PD vs. differential diagnosis task. Feature analysis showed that PD vs. healthy control discrimination is dominated by tremor-related spectral and amplitude features, whereas variability-related features are more relevant for differential diagnosis. These results demonstrate that competitive performance can be achieved using a simple and interpretable pipeline, providing a practical alternative to more complex deep learning approaches.

***Correspondence author email address:** quanvu42@lqdtu.edu.vn;spesinfo@yahoo.com
DOI: 10.21015/vtse.v14i1.2373

1 Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting millions of individuals worldwide and is primarily characterized by motor impairments such as tremor, bradykinesia, and rigidity [1, 2]. Early and objective detection of these motor symptoms is critical for disease management, treatment planning, and long-term monitoring.

However, recent wearable sensing technologies have made it possible to apply continuous non-invasive motion signal acquisition to assess PD continuously and non-invasively [3–5]. Specifically, wrist-worn inertial measurement units (IMUs), like those found on smartwatches, can offer a viable and scalable approach to recording upper-limb motor behavior both in the clinical and real-world contexts [6–9]. The feasibility of

This work is licensed under a Creative Commons Attribution 3.0 License.

wearable-based methods in the monitoring of motor symptoms in non-laboratory environments has been demonstrated by large-scale datasets like the Parkinson's Disease Smartwatch (PADS) dataset [12, 20] and by real-world research like WATCH-PD [21].

Tremor is one of the unique and measurable PD symptoms. The tremor of Parkinsonism usually falls within the range of 37Hz and can be recorded effectively with the help of wrist-worn inertial sensors with high accuracy and reliability [7, 10, 13, 14]. Tremor-related oscillations are more readily visible in a broad spectrum of short-period upper limb tasks than gait-based methods which need well-structured locomotion, and are thus better targeted in smartwatch-based monitoring and in the development of digital biomarkers.

The use of deep learning (DL) techniques in detecting PD using wearable sensor data has become more common over the last few years [13–17]. Multi-scale and attention-based architectures are advanced architectures that have demonstrated high predictive accuracy by automatically learning discriminative representations on raw signals. Nevertheless, these methods are frequently based on feature extraction implicitly, are computationally expensive, and might be not interpretable. Moreover, variations in assessment procedures, specifically the lack of strong subject-level isolation, may result in optimistic evaluation estimates and restrict reproducibility.

Besides that, recent findings bring up the issue of generalization and practical implementation. On controlled datasets, performance might not be reliably transferred to independent cohorts or free-living conditions, as has been shown in experiments on datasets like BioStamp RC2.1 [23].The importance of strong and interpretable models that should be applied in real-world practice settings is also emphasized in clinical validation studies as well as in [24, 25].

Elsewhere, classical machine learning algorithms, applied together with well-designed feature representations and on-the-nose validation strategies offer an orthogonal means of accomplishing interpretability, reproducibility, and computational efficiency. In particular, subject-level nested cross-validation must be imposed to ensure fairness in evaluation as well as prevent data leakage in wearable sensing systems.

The paper provides a PD detection machine learning pipeline, which is easy to recreate and builds on the inertial signals recorded through a smartwatch on the PADS dataset,[12], to identify PD. The framework is based on explicit feature construction, subject level evaluation, and feature analysis to learn more about the model behaviour of different tasks in classification.

The main contributions of this work are as follows:

- Producible and leakage-free evaluation system founded on nested subject-level cross-validation of wearable-based PD detection.
- A representation of spectral and temporal features of upper-limb motion described in interpretable features.
- The general evaluation of classical machine learning models based on the same evaluation protocol.
- A multi-level analysis of feature importance, which gives information about task-dependent and frequency-dependent motor patterns.
- A systematic comparison to recent deep learning methods, emphasizing the performance-interpretability-computational complexity trade-off.

2 Related Work

2.1 Wearable-Based Parkinson's Disease Detection

The concept of determining the condition of Parkinson disease (PD) using wearable inertial sensors has been actively studied over the last few years [3–5]. Initial research centered around accelerometer measurements in order to measure the amplitude and frequency properties of tremors, revealing that Parkinsonian tremor has a tendency to focus in the 3-7 Hz frequency range. These spectral characteristics make it effective to discriminate PD and healthy controls based on frequency-domain analysis analysis [7, 10, 13, 14].

Simultaneously, wearable sensing has been considered in the measurement of upper- limb and dexterity-related PD, such as finger tapping and characterization of hand movement patterns [10, 11]. These investigations reinforce the overall viability of wearable motion sensing in the detection of PD-related motor abnormalities in non-gait.

In addition to tremor-based research, a significant amount of literature has been done on gait assessment through wearable devices on the lower limbs or trunk. These methods have shown that the variability of gait and rhythmical patterns can be effective biomarkers of progression of the Parkinson disease aggravation [18, 19]. Nevertheless, gait-based approaches typically are based on guided walking activities and multi-sensor fittings, which can restrict their utility in unrestricted and natural monitoring.

Later studies have combined machine learning methods with manually designed time- and frequency-domain features of inertial data. Support Vector Machines and random forest as well as Logistic Regression, are all classical classifiers that have been extensively used on statistical descriptors and spectral power features. Although such studies had reported encouraging classification performance, many of them were carried out on a relatively small group of subjects which may restrict generalizability, usually less than 100 subjects in size, as cited in the literature as[6, 8, 9, 11].

Moreover, various papers have shown that it is actually possible to detect Parkinsonian tremor in the real world with wearable accelerators, which points to the possibility of monitoring continuously even when not in a clinical setting [5, 7, 8]. In more recent studies, clinical-scale trials like WATCH-PD have also confirmed that wearable sensors can be used to screen and monitor early-stage Parkinson's disease remotely and screened at the earliest stage possible [21].

Gyroscopes have relatively lower coverage as compared to accelerators, even though they dominate the literature. Gyroscopes can be used to measure rotational velocity, which can be used to complement other measurements (especially in tremor characterization and pronation-supination dynamics). Some studies have found better discriminative performance when using gyroscope signals, but extensive validation of this modality has not been carried out.

2.2 Deep Learning Approaches for PD Classification

Recent advances in deep learning have led to the adoption of convolutional neural networks, recurrent architectures, and hybrid attention-based models for PD detection [7, 13–15]. These approaches aim to learn hierarchical representations directly from raw inertial signals, potentially capturing complex temporal dependencies beyond handcrafted features.

It has been claimed that deep architectures perform well on several studies, such as multi-scale and attention-based architectures that have been used to capture frequency patterns related to tremors, which are important to seismic sequences and events prediction systems, among others [16, 17]. Specifically, frequency-sensitive and channel-attention models have been demonstrated to improve detection results through concentrating on clinically significant frequency bands.

But these models generally need huge training datasets, significant computing resources and hyperparameter sweeping to prevent overfitting. Notably, reported performance differs greatly according to the evaluation protocol. Research using segment-level cross-validation or subject-dependent splits tends to announce better accuracy than that using subject-independent or nested cross-validation schemes, making it difficult to compare studies fairly.

Furthermore, more intricate architecture is not always associated with better clinical utility. Interpretable feature-based lightweight models can potentially compete with popular performance, and provide benefits in transparency, computational efficiency, and deployability on wearable systems.

2.3 Studies Using the PADS Dataset

Recently, the Parkinson's Disease Smartwatch (PADS) dataset has become among the largest publicly accessible cohorts of smartwatch-based movement disorder analysis, with 469 participants of Parkinson's disease, healthy controls, and differential diagnoses among them [12]. The dataset due to its clinical diversity and the subject-level validation design offers a realistic mark of assessing the wearable-based diagnostic models.

The initial large-scale experiments by Varghese et al. gave the baseline performance on the basis of handcrafted motion features founded on smartwatch inertial signals. The movement-only models that used a subject-level nested cross-validation model had balanced accuracy values of approximately 78.99% in the classification of PD vs. HC and 69.18% in the classification of PD vs. DD classification [20]. These

results highlighted the feasibility of diagnosing using smartwatch and the inherent complexity of the issue of distinguishing PD and other movement disorders using only motion signals.

In order to improve the performance, additional researches incorporated clinical questionnaire and non-motor symptom data to ensemble learning models to increase PD vs. HC balanced accuracy to 91.16% higher. However, at 72.42%, PD vs. DD discrimination did not change significantly, statistically, at 72.42% [20].

Parallel Concurrently, more recent experiments with independent datasets (e.g. BioStamp RC2.1) have demonstrated lower performance than with controlled datasets, emphasizing the difficulty of cross-sensing setup and population generalization [23]. This implies that sound evaluation must be characterized by both good accuracy and consistency across datasets and conditions.

On the contrary, the method proposed in the current research is limited to smartwatch inertial signals, which do not need any questionnaire or clinical inputs. Our movement-only pipeline, based on the PSD-based spectral descriptors, and statistical motion features with LASSO-based feature selection in a nested cross-validation model, yields a balanced accuracy of 79.26% for PD vs. HC, comparable to prior movement-only baselines.

Importantly, this level of performance is achieved using simple and interpretable models such as Logistic Regression. This finding suggests that carefully designed feature representations and leakage-free evaluation protocols can compensate for increased model complexity, providing a more practical and transparent alternative to deep learning approaches.

2.4 Identified Research Gaps

Despite considerable progress, several gaps remain in smartwatch-based PD research:

- Limited large-scale and reproducible comparisons between classical machine learning and deep learning models under consistent nested cross-validation protocols.
- Underexplored generalization across datasets and real-world deployment scenarios, despite evidence from recent clinical and cross-dataset studies [21, 23–25].
- Insufficient focus on PD vs. DD classification, which is clinically more relevant and inherently more challenging than PD vs. HC.
- Lack of emphasis on computational efficiency and interpretability for real-world deployment on wearable and mobile platforms.

Therefore, a systematic and interpretable evaluation of wearable-based PD detection using carefully engineered spectral and statistical features under strict subject-level validation is required to bridge these gaps.

3 Background

3.1 Spectral Representation of Tremor Signals

Parkinsonian tremor is characterized by rhythmic oscillatory motion predominantly occurring within the 3–7 Hz frequency range. Wrist-worn inertial sensors capture these oscillations through two complementary modalities: accelerometers measure linear acceleration, whereas gyroscopes record angular velocity. The combination of these modalities provides a more comprehensive description of tremor dynamics.

Time-domain representations of tremor signals are often affected by noise, transient artifacts, and non-stationarity, which may obscure periodic patterns. Frequency-domain analysis, in contrast, highlights oscillatory behaviour by concentrating signal energy within specific bands. The power spectral density, estimated using the Welch method, provides a stable and variance-reduced spectral representation by averaging periodograms across overlapping windowed segments.

Spectral band power within clinically relevant frequency intervals has been widely used to quantify tremor intensity and rhythmic dominance. In addition to narrow-band tremor power in 3–7 Hz, broader spectral measures such as centroid, spread, and entropy capture redistribution of signal energy across frequencies. Moreover, the asymmetry of the left and right limbs can also indicate the mono-lateral dominance of tremor, which is a typical feature of the Parkinson disease.

3.2 Statistical Motion Descriptors

While spectral analysis captures rhythmic structure, amplitude dispersion and movement variability pro-

vide complementary diagnostic information. Statistical descriptors extracted from time-domain signals characterize signal intensity and irregularity without requiring explicit frequency decomposition.

Commonly used measures include mean amplitude, standard deviation, median value, maximum absolute amplitude, and aggregated signal energy. Window-based statistics computed over subsegments further capture temporal variability within a recording session.

These descriptors are computationally friendly, understandable, and suited to real time wearable implementation. Statistical measures when used in conjunction with spectral measures allow concurrently define rhythmic tremor elements and the larger motor variation patterns.

3.3 Regularization-Based Feature Selection

High-dimensional feature spaces are created by multi-sensor, multi-axis and multi-task recordings. Without appropriate dimensionality control, redundant or correlated features may increase overfitting risk and reduce generalization performance.

Principled embedded feature selection methods are given by regularization-based methods. The least absolute shrinkage and selection operator (LASSO) uses a penalty of the form of the L_1 which imposes sparsity by penalizing a coefficient that is less informative into zero. Compared with filter-based selection approaches, LASSO integrates feature selection directly into the modelling process, balancing predictive power and model simplicity. The importance of such optimization in high-dimensional biological and signal data is well-documented; for instance, Rahu et al. [26] demonstrated that leveraging sophisticated optimization and profiling techniques is critical for accurate predictive modeling in complex protein sequences.

In the present work, the LASSO-based selection is incorporated into a cross-validation framework which is nested. The feature selection is only done on training folds and the selected set is applied to the respective test fold. This design avoids the leak of information and gives unbiased estimations of performance.

Table 1. Summary of participant distribution across diagnostic groups in the PADS dataset.

Group	Number of Subjects	Percentage (%)
PD	276	58
HC	114	25
DD	79	17

3.4 Evaluation Under Class Imbalance

Clinical datasets often exhibit imbalanced class distributions. Standard accuracy may overestimate performance in such scenarios. Balanced metrics that incorporate true positive and true negative rates are therefore essential. The distribution of participants across diagnostic groups used in this study is summarized in Table 1

The Matthews correlation coefficient (MCC) provides a comprehensive summary of classification performance by accounting for all confusion matrix components. In contrast to accuracy, MCC is still informative when the classes are skewed and is especially suitable in medical diagnostic problems. Oversampling techniques such as SMOTE can be applied within training folds to mitigate imbalance while preserving unbiased test evaluation.

4 Proposed Methodology

Figure 1 illustrates the complete processing framework adopted in this study. The proposed pipeline transforms raw smartwatch inertial signals into subject-level classification outputs through sequential stages including feature extraction, normalization, feature selection, imbalance handling, and nested cross-validation-based model evaluation.

4.1 Feature Extraction and Preprocessing

The pipeline begins with raw tri-axial accelerometer and gyroscope signals sampled at 100 Hz. Signals were segmented into fixed-length windows of 2.5 seconds to ensure stable spectral estimation and reduce intra-subject variability. Window-level features were aggregated at the subject level prior to classification.

4.2 Frequency-Domain Representation

To characterize oscillatory tremor behaviour, power spectral density was estimated using the Welch method.

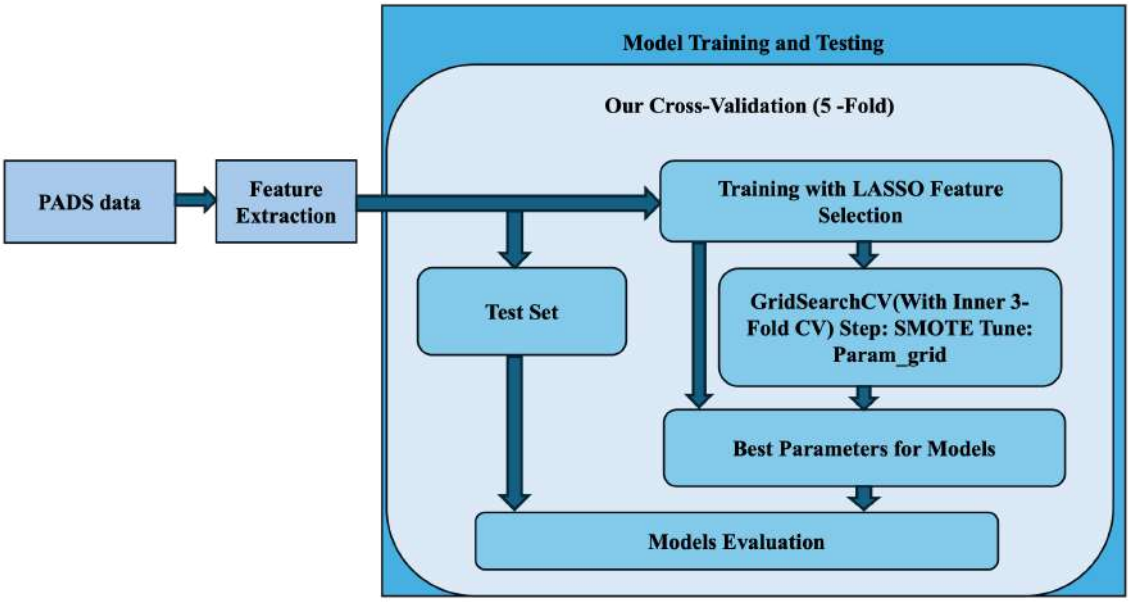

Figure 1. Overview of the proposed machine learning pipeline. After feature extraction, subject-level nested cross-validation is applied. The outer loop splits the data into training and testing folds, while the inner loop performs feature selection and hyperparameter tuning.

For a windowed signal $x(n)$, the PSD is defined as

$$P_{xx}(f) = \frac{1}{K}\sum_{k=1}^{K}\left|\mathcal{F}\left\{x_k(n)\right\}\right|^2, \quad (1)$$

where K denotes the number of overlapping segments and $\mathcal{F}\{\cdot\}$ is the Fourier transform operator.

From the PSD representation, band power values were computed within clinically relevant frequency ranges from 1 to 19 Hz, with particular emphasis on the tremor-dominant band between 3 and 7 Hz. In addition to band energy measures, spectral descriptors including centroid, spread, entropy, and band-power ratios were extracted to characterize energy redistribution across frequencies. The plots in Figure 2 depict the example power spectrum density of the left and right wrists.

4.3 Peak-Based Spectral Descriptors

To capture rhythmic characteristics beyond amplitude concentration, peak-based descriptors were introduced. Spectral peaks were identified as local maxima within the tremor-relevant frequency band subject to a minimum prominence threshold and a minimum inter-peak frequency separation.

Let f_i denote the detected spectral peak frequencies. The average peak frequency (APF) is defined as

$$APF = \frac{1}{M}\sum_{i=1}^{M} f_i, \quad (2)$$

where M is the number of detected peaks.

In the time domain, peak positions were detected using local maxima with amplitude prominence constraints. Inter-peak intervals are defined as

$$\Delta t_i = t_{i+1} - t_i. \quad (3)$$

The average peak interval (API) is computed as

$$API = \frac{1}{M-1}\sum_{i=1}^{M-1}\Delta t_i. \quad (4)$$

API quantifies temporal rhythmic regularity and captures tremor periodicity independently of amplitude magnitude.

4.4 Time-Domain Statistical Features

Complementary statistical descriptors were extracted from each window, including mean, standard deviation, median, signal energy, and maximum amplitude. To capture unilateral tremor dominance, asymmetry indices

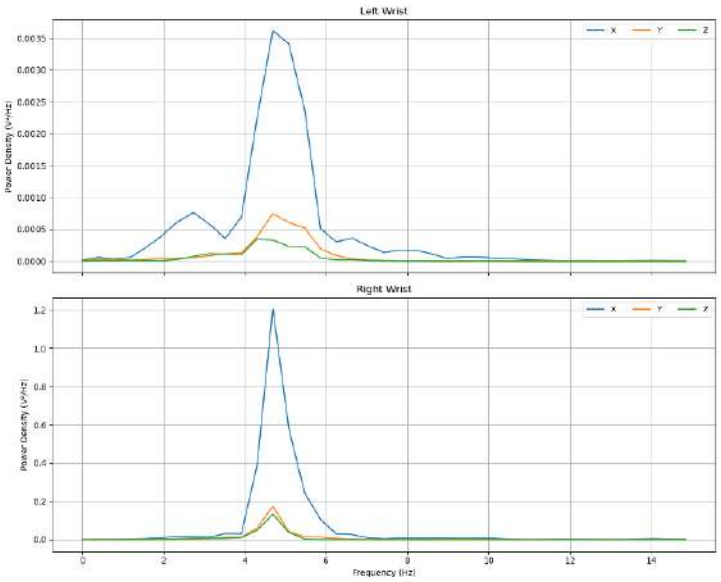

Figure 2. Example power spectral density plots for left and right wrists. The dominant tremor-related peak is concentrated around 4–6 Hz, particularly on the X axis.

between corresponding left and right wrist channels were computed as

$$AI = \frac{|L - R|}{|L| + |R| + \epsilon}, \tag{5}$$

where L and R denote feature values from left and right wrist sensors, respectively, and ϵ is a small constant for numerical stability.

4.5 Standardization

All features were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma}, \tag{6}$$

where μ and σ were computed exclusively on training data within each outer fold.

4.6 Model Training and Feature Selection

High-dimensional feature spaces may introduce redundancy and overfitting risk. Therefore, LASSO-based feature selection was embedded within the training process. The LASSO optimization problem is defined as

$$\min_{\beta} \left(\|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right), \tag{7}$$

where λ controls the sparsity penalty. Only features with non-zero coefficients were retained. Importantly, feature selection was performed exclusively within training folds during the inner cross-validation loop to prevent information leakage.

In order to deal with the imbalance in the classes, SMOTE oversampling was performed only on training data in each outer fold. Test data and validation were not interfered. Logistic Regression (LR), Support Vector Machine with radial basis kernel (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP), and CatBoost (CB) were considered as the following classifiers.Hyperparameters were optimized through grid search in the inner loop.

4.7 Nested Cross-Validation and Evaluation Metrics

In order to guarantee the performance estimation without any biases, a nested cross-validation structure was implemented. The outer loop used 5-fold subject-level partitioning, which guaranteed that no data of the same subject was used in both the training and testing sets. Within each outer fold, a 3-fold inner cross-validation loop was used for hyperparameter tuning and LASSO optimization.The nested subject-level cross-validation strategy proposed inFigure 3 is aimed at preventing data leakage.

Performance metrics were computed exclusively on outer test folds and included balanced accuracy (BAcc), area under the ROC curve (AUROC), F_1-score, and Matthews correlation coefficient (MCC). MCC is defined as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{8}$$

Mean and standard deviation across outer folds were reported to assess model stability.

5 Experimental Results and Discussion

5.1 Overall Classification Performance

The proposed framework was evaluated on two clinically relevant binary classification tasks under a nested 5×3 subject-level cross-validation protocol. All performance metrics were computed exclusively on outer test folds to ensure unbiased generalization estimates. Table 2 reports the classification performance of all evaluated models across the two diagnostic tasks.

Task A – PD vs. HC

For the PD vs. HC discrimination task, the proposed movement-only pipeline achieved strong and stable

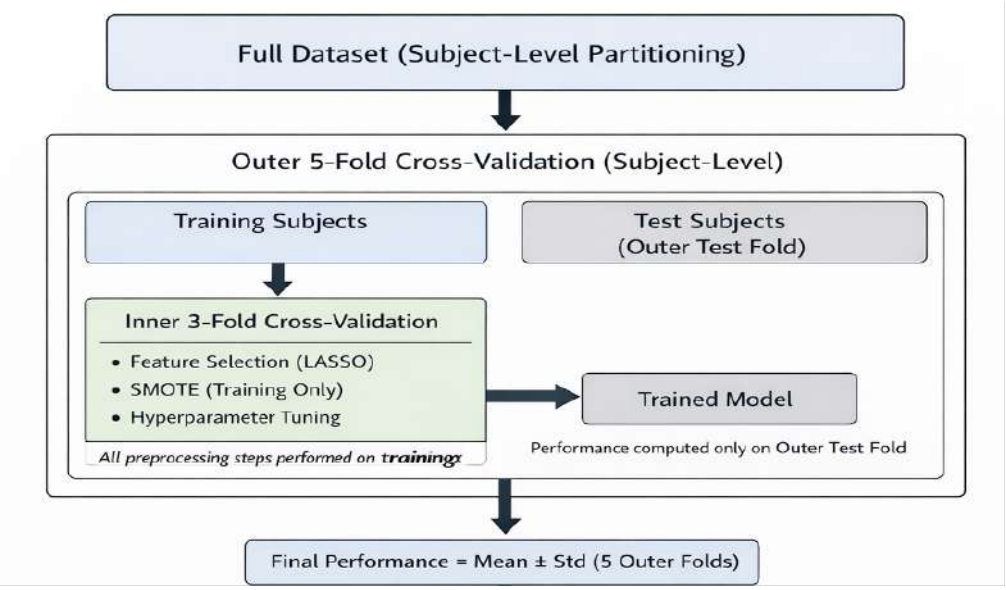

Figure 3. Nested subject-level cross-validation framework for leakage-free hyperparameter tuning and unbiased performance estimation.

performance across models. Logistic Regression obtained the highest balanced accuracy of 79.26 ± 5.50%, with an overall accuracy of 87.32 ± 1.99% and AUROC of 89.77 ± 4.29%. The corresponding MCC value of 61.91 ± 7.14 indicates substantial agreement beyond chance.

SVM and MLP demonstrated comparable performance, achieving balanced accuracies of 73.85 ± 3.85% and 75.46 ± 8.43%, respectively. Random Forest yielded slightly lower balanced accuracy of 67.85 ± 4.26%, suggesting that nonlinear ensemble modelling did not provide additional benefit over simpler linear decision boundaries in this feature space.

Task B – PD vs. DD

The PD vs. DD classification task proved substantially more challenging. Balanced accuracy values ranged from 65.58 ± 3.28% for Logistic Regression to 67.15 ± 2.76% for MLP. AUROC values remained in the moderate range, and MCC values were notably lower than in Task A, reflecting reduced separability between PD and other movement disorders.

Remarkably, there were not significant differences in the performance between models in this task. Random Forest had the best overall accuracy of 77.43 ± 3.22% and MCC of 40.16 ± 9.53, but balanced accuracy was similar to other models.This convergence suggests that model capacity is not the primary limiting factor; rather, intrinsic overlap in motor characteristics between PD and differential diagnoses likely constrains classification performance. Performance variability was moderate across tasks, which means generalization was consistent in the case of subject-level separation. Notably, all findings were derived without using non-motor or questionnaire data. There is an evident performance difference between PD vs. HC and PD vs. DD classification. Although tremor-related characteristics and asymmetry scores have a high discriminative ability between PD and healthy individuals, the separation of PD and other movement disorders is inherently more challenging because of the presence of overlapping motor phenotypes.

It is worth noting that the overall performance of Logistic Regression was the best out of the reviewed classical models in Task A, with SVM and MLP showing relatively similar results. This implies that engineered feature space offers a lot of discriminative information, and that further model complexity among the classical family of experiments under test offers few additional benefits at the same nested subject-level validation protocol.

Since Logistic Regression achieved the strongest overall performance in the PD vs. HC task and served as the

Table 2. Classification performance across tasks under nested subject-level cross-validation. Results are reported as mean ± standard deviation across outer folds.

Task	Model	Acc (%)	BAcc (%)	F_1-score	AUROC	MCC
4*PD vs. HC	LR	87.32 ± 1.99	79.26 ± 5.50	92.01 ± 1.16	89.77 ± 4.29	61.91 ± 7.14
	SVM	84.50 ± 1.72	73.85 ± 3.85	90.31 ± 1.23	88.65 ± 4.73	52.86 ± 4.51
	MLP	85.63 ± 4.27	75.46 ± 8.43	91.05 ± 2.59	88.40 ± 3.71	55.52 ± 14.30
	RF	83.66 ± 3.09	67.85 ± 4.26	90.15 ± 1.99	86.81 ± 5.79	47.35 ± 9.30
4*PD vs. DD	LR	73.58 ± 2.50	65.58 ± 3.28	50.38 ± 5.45	72.67 ± 2.34	33.37 ± 6.16
	SVM	73.33 ± 3.06	66.18 ± 3.41	51.67 ± 5.25	72.95 ± 3.19	33.75 ± 6.83
	MLP	73.58 ± 2.66	67.15 ± 2.76	53.32 ± 4.32	72.47 ± 1.93	35.17 ± 6.06
	RF	77.43 ± 3.22	66.26 ± 4.32	50.25 ± 8.04	73.67 ± 5.27	40.16 ± 9.53

Table 3. Sensitivity and specificity of the Logistic Regression model under subject-level nested cross-validation.

Task	Sensitivity	Specificity
PD vs. HC	0.9384 ± 0.0243	0.6467 ± 0.1266
PD vs. DD	0.4640 ± 0.0907	0.8477 ± 0.0496

primary interpretable model in the subsequent analyses, its diagnostic sensitivity and specificity are reported separately in Table 3.

Table 3 shows a clear contrast between the two tasks. For PD vs. HC classification, Logistic Regression achieved high sensitivity but lower specificity, indicating that the model was highly effective at detecting Parkinson's disease cases but more prone to misclassifying healthy controls as PD. In contrast, for PD vs. DD classification, specificity was higher while sensitivity decreased substantially, suggesting that the model was more effective at rejecting non-PD cases than identifying PD within a clinically similar population. This asymmetric behavior highlights the task-dependent nature of movement-based diagnosis.

5.2 Sensor Ablation Analysis

To evaluate the contribution of different sensor modalities, an ablation study was conducted comparing accelerometer-only, gyroscope-only, and combined sensor configurations.

For Task A, gyroscope-based features consistently outperformed accelerometer-only features, indicating that rotational motion characteristics provide strong discriminative information for Parkinsonian tremor. Sensor fusion further improved performance, demonstrating complementary contributions between amplitude-based and rotational features.

For Task B, the performance differences between sensor configurations were smaller, reflecting the increased difficulty of distinguishing PD from other movement disorders. Nevertheless, sensor fusion remained the most effective configuration, suggesting that combining multiple motion modalities enhances robustness when discriminative signals are subtle.

Figure 4 presents the sensor ablation results for both PD vs. HC and PD vs. DD tasks. A clear performance gain is observed when combining accelerometer and gyroscope signals, suggesting that each modality captures distinct aspects of tremor dynamics. This result highlights the importance of multimodal sensing for improving classification robustness.

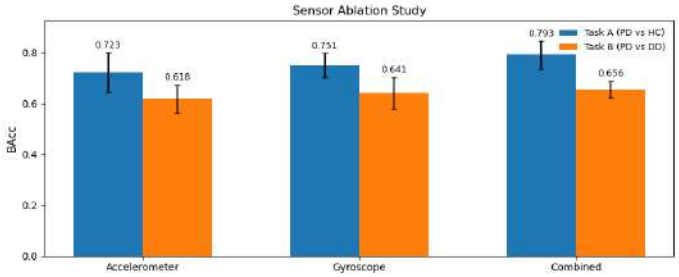

Figure 4. Sensor ablation results for Task A (PD vs. HC) and Task B (PD vs. DD). The combined sensor configuration consistently achieves the highest balanced accuracy, highlighting the complementary nature of accelerometer and gyroscope signals.

5.3 Error Analysis via Confusion Matrix

To further analyze classification behaviour, confusion matrices were examined for both tasks using the best-performing model.

In the case of the PD vs. HC task, the model is very sen-

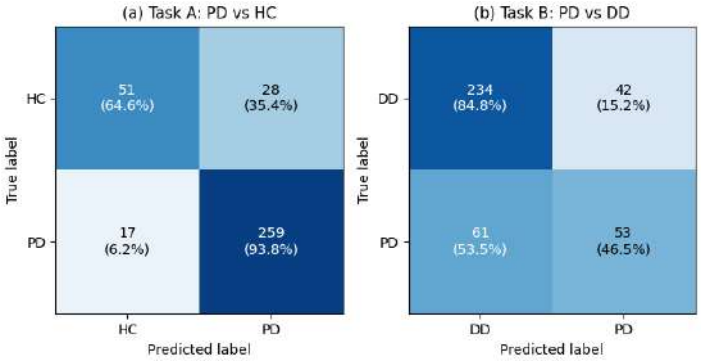

Figure 5. Normalized confusion matrices for (a) PD vs. HC and (b) PD vs. DD classification. Each cell represents the proportion of predictions per class. Errors are concentrated between clinically similar categories, particularly in PD vs. DD.

sitive to the cases of the Parkinson disease, meaning that tremor-related features are well represented. Majority of misclassifications is seen in normal control samples which are at times falsely predicted as PD. This implies a partial overlap of low amplitude or physiological motion patterns between healthy and persons with Parkinson disease.

However, with the PD vs. DD task, the model has a different pattern of errors. Misdiagnoses are mainly linked to the cases of Parkinson being predicted as the other movement disorders and thus less sensitive. This means that it is significantly more difficult to draw the line between PD and closely related clinical conditions because the two groups could display similar motor traits.

Figure 5 illustrates the normalized confusion matrices for both classification tasks. While clear separation is observed between PD and healthy controls, classification errors are more pronounced in the PD vs. DD setting. These results suggest that the primary limitation arises from feature overlap between pathological conditions rather than from insufficient model capacity.

These observations motivate a deeper investigation of feature-level contributions, as discussed in the following subsection.

5.4 Model Behaviour and Feature Contribution

The highest rate of the classifier was observed in the PD vs. HC task with the best performance of the Logistic Regression with SVM and MLP with similar yet slightly lower results. In the case of the PD vs. DD task, the difference between the performance of the models was comparatively minor, and there were no specific classifiers that perform better than others do in all cases. This indicates that it is the representation of the features that mainly dictate the performance of classification and not the complexity of the decision boundary.

The best decoded features to classification are shown in Figure 6. This representation allows straightforward interpretation of the effect of particular contexts of motion on model decisions by storing all structural information. It is important to note that the most significant features are not uniformly distributed but those specific task sensor configurations, which means that the discriminative information is context-specific.

Figure 7 depicts the cumulative importance of the base-feature in each channel. The heatmap shows that relatively few categories of features drive the performance of the models, with many contributing a little. This implies that there is a redundancy in the feature space and dimensionality reduction could be achievable without causing a significant loss in discriminative power.

Task A – PD vs. HC

In the case of PD vs. HC, spectral and amplitude-related features are the most influential decoded features. In particular, features associated with spectral centroid, spectral spread, and band-power ratios between low-frequency (13 Hz) and tremor-related (37 Hz) components consistently rank among the top contributors.

A representative example is the feature:

TouchNose_LeftWrist_Acceleration_Z_ratio_pd_vs_high_3_7_over_7_12

which is a redistribution of spectral energy in frequency bands. This trend implies that Parkinsonian tremor brings about a typical change in energy focus to certain frequency bands.

The aggregated heatmap further confirms that spectral dispersion measures and band-power ratios play a dominant role across channels. Moreover, the descriptors that are amplitude-related like windowed standard deviation and summed absolute values are always significant. The results of these studies suggest that the discrimination between PD and HC is highly conditioned by

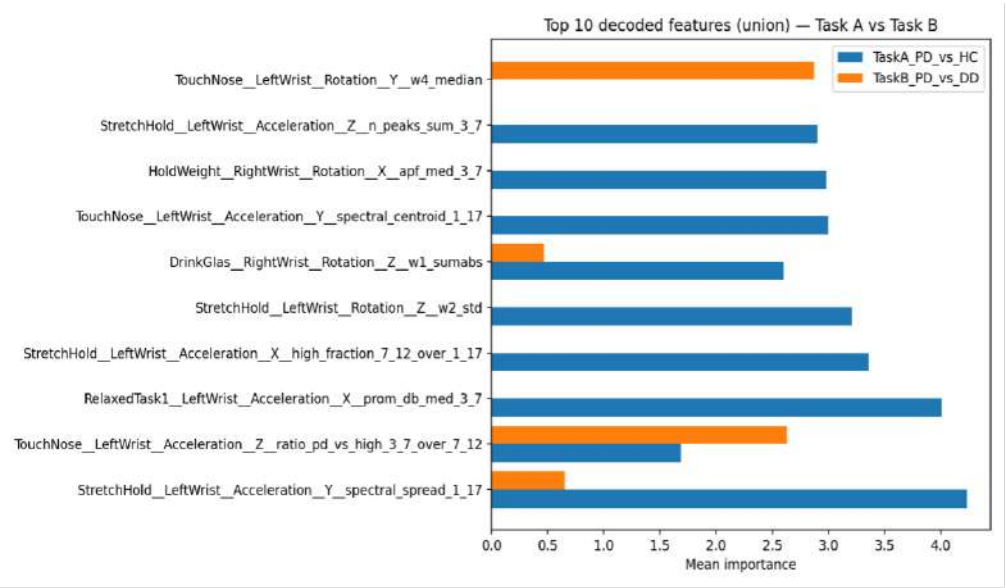

Figure 6. Top 10 decoded features ranked by mean importance across outer folds for Task A (PD vs. HC) and Task B (PD vs. DD). Each feature preserves task, wrist, sensor modality, axis, and descriptor information.

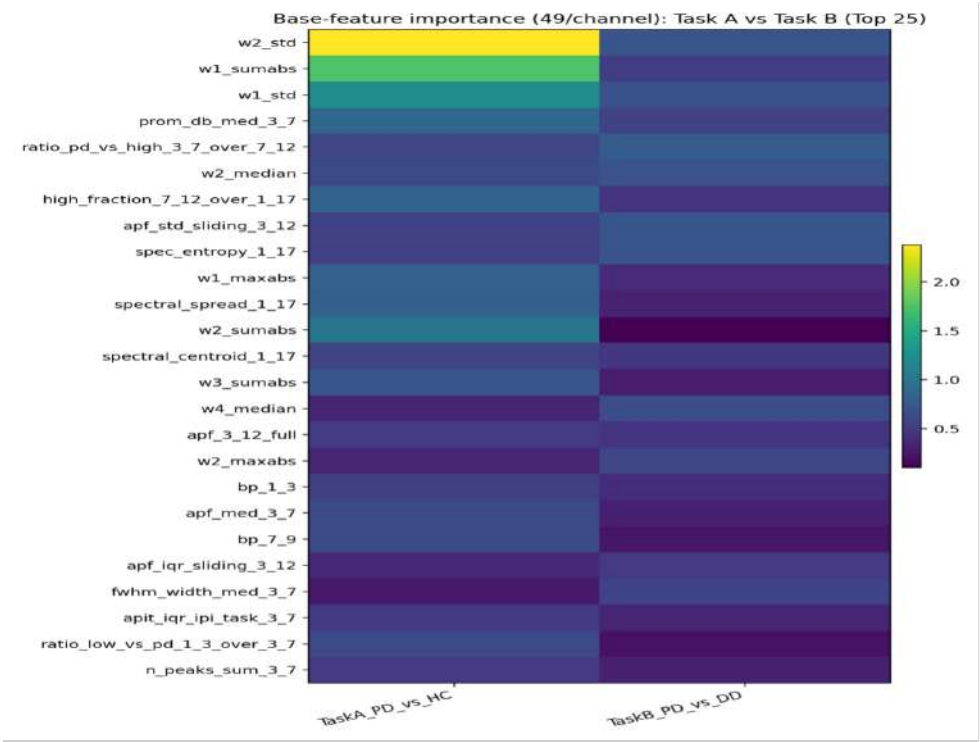

Figure 7. Aggregated base-feature importance heatmap for Task A (PD vs. HC) and Task B (PD vs. DD), showing the top 25 feature categories after averaging importance across channels and outer folds.

the intensity of global tremors and the distribution of energy in frequency domains.

Task B – PD vs. DD

In contrast, PD vs. DD classification exhibits a different importance structure. The top decoded features are

more closely related to temporal variability and oscillatory stability rather than simple energy redistribution.

For example, features such as:

TouchNose_LeftWrist_Rotation_Y_w4_median

and sliding-window frequency variability descriptors (e.g., APF-based measures) appear among the most influential features. These characteristics isolate the momentary changes in motion patterns, indicating that temporal irregularity is of substance in separating the Parkinson disease and other movement disorders.

The significance of band-power ratios is less in comparison with Task A and characteristics associated with spectral entropy, peak variability, and window-level dynamics are more pronounced.This shift indicates that PD vs. DD discrimination relies on more subtle characteristics of movement stability rather than dominant tremor signatures.

Cross-Task Interpretation

There is evident difference in the discriminative processes in the two tasks. The classification of PD vs. HC is mainly subject to strong spectral and amplitude cues that are related to tremor dominance and the classification of PD vs. DD relies more on variability, irregularity, and fine-grained temporal dynamics.

Collectively, the analysis of decoded features and feature categories together prove that the proposed feature representation reflects both general and situation-specific trends in motor behavior. These findings also confirm that the complexity of classifiers is less relevant in determining the performance of a model than feature representation and design.

These pattern feature-level differences are in agreement with the patterns of misclassification as identified in the confusion matrix analysis.

In the case of the PD vs. HC task, the majority of errors involve healthy control samples being falsely categorized as PD and this is because the characteristics of tremor are the ones that dominate and could also be found in low-amplitude physiological movement.

Conversely, in the case of the PD vs. DD task, misclassifications have a higher rate of associating cases of Parkinson disease with other movement disorders. It means that the proposed feature representation is effective to indicate the existence of abnormal motor activity, but it is less effective to distinguish between clinically similar pathological conditions with overlapping motion characteristics.

These results indicate that the classification errors are mostly due to overlapping of features as opposed to lack of model capacity, which supports the role of feature design in movement-based Parkinson disease analysis.

5.5 Comparison with Previous PADS Studies

Table 4 compares the proposed movement-only framework with previous studies on the PADS dataset under a consistent subject-level nested cross-validation protocol. The proposed approach achieved 79.26% balanced accuracy for the PD vs. HC task, closely matching the previously reported movement-only baseline of 78.99%, indicating that the performance gain is not due to overfitting or data leakage but reflects comparable discriminative capability under the same evaluation setting.

The major finding of Table 4 is the difference between movement-only and multimodal methods. Although more multimodal techniques that utilize questionnaire or clinical metadata have significantly higher performance (e.g., 91.16% balanced accuracy), the performance is attained with a different input condition which involves non-sensor information. Thus, direct comparison must be viewed with a grain of salt, because the multimodal environment offers more diagnostic data, than motion data.

On the contrary, the suggested framework is entirely based on the inertial signals obtained by wearables, which makes it better adapted to the context of continuous and real-life monitoring where there might be no auxiliary clinical data. This design option is consistent with the recent real-life research like the WATCH-PD, which focus on the viability and significance of sensor-only surveillance in free setting [21]. Within this limitation, the obtained performance is competitive to the previous movement-only baselines, suggesting that the suggested feature-representation and assessment protocol is able to represent effectively discriminative motion patterns.

Besides the classical machine learning methods, re-

Table 4. Comparison of the proposed movement-only framework with previous studies on the PADS dataset under subject-level nested cross-validation.

Study	Input Data	Feature Selection	Validation Protocol	PD vs. HC (BAcc %)	PD vs. DD (BAcc %)
Varghese et al. [6]	Inertial sensors	Embedded / ensemble	Nested CV (subject-level)	78.99	69.18
Varghese et al. [20]	Inertial + questionnaire	Embedded	Nested CV (subject-level)	91.16	72.42
Proposed method	Inertial sensors only	LASSO (nested CV)	Nested CV (subject-level)	79.26	67.15

cent literature has also delved into deep learning models to detect Parkinson disease using wearable inertial data. For example, Meng et al. [7] proposed deep neural network architectures evaluated on the same PADS dataset. In contrast, Wang et al. [16] investigated a different dataset and reported strong performance, achieving 92.39% accuracy and an AUC of 0.9797 using a multiscale frequency-sharing channel attention network.

Although these findings show that deep learning has a high potential, there are several significant distinctions that must be taken into account. To begin with, the accuracy and AUC are reported mainly in these studies, but the current study focuses on balanced accuracy (BACC), which is more suitable in imbalanced data such as PADS. This has led to a limited direct numerical comparison.

Second, deep learning models are based on implicit feature learning via intricate structures and this can decrease the level of interpretability. Conversely, the suggested framework uses a clear feature engineering approach with a LASSO-based selection, which allows the task-, sensor-, and frequency-specific contributions to be analyzed in detail.

Third, despite the reported usage of nested cross-validation by studies, there are variations in the details of the implementation, specifically on the subject-level grouping, which could affect the performance reported. Ensuring strict subject-level separation is critical in wearable sensing applications to avoid data leakage across repeated measurements from the same individual.

Furthermore, recent cross-dataset studies indicate that performance achieved on controlled datasets such as PADS may not generalize directly to independent cohorts. For example, models evaluated on the BioStamp RC2.1 dataset show reduced performance, highlighting the challenges of generalization across sensor configurations and patient populations [23]. This implies that wearable-based diagnostic systems should be considered highly robust and reproducible, as well as raw accuracy.

Notably, the further discussion of sensitivity and specificity demonstrates that it is not only the balanced accuracy that shows the differences in performance. In the case of the PD vs. HC task, high sensitivity and lower specificity of the proposed model implies that it is sensitive enough to detect cases of Parkinson disease but is prone to a greater confusion with healthy controls. This action is in line with the dependence of tremor-related characteristics that can also be present in the low-amplitude physiological movement.

In the case of the PD vs. DD task, the total performance is lower in all the studies, including the proposed method. The model is less sensitive and more specific, meaning that it is more difficult to detect the disease of Parkinson in a population of the same type of disease with accuracy. The observation is in line with previous research and demonstrates the inherent difficulty of the problem of motion-based differential diagnosis.

These results indicate that the main constraint is the feature separability and not the model complexity. Movement-based features are useful in identifying abnormal motor behavior, although they are not as effective as differentiating various pathological conditions sharing common motion features. It is worth noting that the frequency-conscious schemes used by modern deep learning systems are aligned with the relevance of spectral characteristics of tremors found in the present

study, thus indicating that both methods represent the same underlying motor patterns irrespective of the complexity of the model.

6 Conclusion and Future Work

In this work, we introduced a machine learning framework for detecting Parkinson's disease using smartwatch-based inertial signals from the PADS dataset. The design of the pipeline emphasizes both reproducibility and interpretability. By combining spectral and statistical features with LASSO-based feature selection, and evaluating the model using subject-level nested cross-validation, we ensured that the results remain reliable and free from data leakage.

The experimental findings show that models based solely on movement data can still achieve performance comparable to previously reported inertial baselines when evaluated under the same protocol. For the PD vs. HC classification task, the framework achieved a balanced accuracy of 79.26% and an overall accuracy of 87.32%. Interestingly, even relatively simple models such as Logistic Regression performed competitively, suggesting that the quality of feature representation plays a more critical role than model complexity.

On further examination of the contribution of features, it can be found that varying tasks depend on varying motion properties underlying them. The frequency components of tremors and the features based on amplitude seem to be the primary discriminators in the PD vs. HC environment. Conversely, the variability and irregularity of the motion patterns with time are more crucial in differentiating PD and DD. This implies that various mechanisms of motor control are manifested in various classification situations.

Concurrently, the outcomes indicate that the PD vs. DD classification is more difficult. Even though the performance remains similar across the folds, a smaller balanced accuracy is observed as compared to PD vs. HC. This probably indicates the similarity in the motor symptoms between movement disorders, indicating that the constraint is more on separability of features than on the type of classifier.

While some recent deep learning approaches report higher accuracy on the same dataset, these gains often come at the cost of increased model complexity and reduced interpretability. On the other hand, the suggested solution represents an easier and more transparent option with competitive outcomes. It is also noteworthy that the patterns of feature importance here are also consistent with frequency-aware representations of deep learning studies, which suggests that the two methods can be extracting the same underlying signal properties.

Generalizability is also another significant factor. Recent cross-dataset studies indicate that models developed on controlled datasets are potentially not very well transferable to independent cohorts because of sensor placement differences, acquisition setting differences, and population differences. This strengthens the argument that accuracy, as well as strength and reproducibility should be considered when developing wearable-based diagnostic systems.

One can say a couple of limitations. First, the analysis is based on structured motor tasks rather than continuous real-world monitoring. Second, only wrist-based inertial signals were used, without incorporating other potentially informative modalities such as speech or clinical scores. Third, although nested cross-validation provides strong internal validation, external validation on independent datasets is still required to confirm generalizability.

In future work, several directions can be explored. Improving feature discriminability may help enhance performance in more challenging tasks such as PD vs. DD classification. Also, the integration of several data modalities might offer a more detailed view of patient conditions. Lastly, hybrid methods combining interpretable feature engineering with lightweight learning models can potentially provide a trade-off between performance and transparency.

The future work is going to be devoted to three directions:

- Increasing the discriminativity of features to increase performance on differential diagnosis.
- Incorporating multimodal data sources to elicit complementary clinical data.
- The use of hybrid methods to integrate interpretable feature engineering and lightweight representation learning models.

Overall,the results indicate that feature-based

machine learning pipelines are a viable and scalable solution to smartwatch-based Parkinson's disease screening, especially in the context of a real-world application.

Author Contributions

Quan Vu: Conceptualization, Methodology, Software, Writing – Original draft preparation. **Manh-Cuong Nguyen**: Investigation, Data curation, Writing – Review and editing. **Duc-Tan Tran**: Supervision, Validation, Writing – Review and editing. **Vijender Kumar Solanki**: Supervision, Review and editing.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

funding: This research received no external funding.

Ethical Approval: This study uses a publicly available dataset and does not involve any new studies with human participants or animals performed by the authors.

Informed Consent: Informed consent was obtained from all participants in the original data collection as reported by the dataset providers.

Data Availability: The Parkinson's Disease Smartwatch (PADS) dataset used in this study is publicly available on PhysioNet at https://physionet.org/content/pads/1.0.0/ (DOI: 10.13026/m0w9-zx22).

Acknowledgment

The authors acknowledge that the artificial intelligence (AI) tool ChatGPT was used solely for language editing and technical refinement of the manuscript. No AI tool was used for data generation, analysis, or scientific conclusions. All AI-assisted content was carefully reviewed and validated by the authors. results, or scientific conclusions. All AI-assisted content was reviewed and validated by the authors.

References

[1] K. R. Chaudhuri, D. G. Healy, and A. H. V. Schapira, "Non-motor symptoms of Parkinson's disease: Diagnosis and management," *Lancet Neurology*, vol. 5, pp. 235–245, 2006.

[2] C. H. Hawkes, K. Del Tredici, and H. Braak, "A timeline for Parkinson's disease," *Parkinsonism and Related Disorders*, vol. 16, pp. 79–84, 2010.

[3] L. Lonini, A. Dai, N. Shawen, T. Simuni, C. Poon, L. Shimanovich, M. Daeschler, R. Ghaffari, J. A. Rogers, and A. Jayaraman, "Wearable sensors for Parkinson's disease: Which data are worth collecting for training symptom detection models," *npj Digital Medicine*, vol. 1, no. 1, p. 64, 2018.

[4] M. H. G. Monje, G. Foffani, J. Obeso, and A. Sanchez-Ferro, "New sensor and wearable technologies to aid in the diagnosis and treatment monitoring of Parkinson's disease," *Annual Review of Biomedical Engineering*, vol. 21, no. 1, pp. 111–143, 2019.

[5] E. Rovini, G. Maremmani, and F. Cavallo, "How wearable sensors can support Parkinson's disease diagnosis and treatment: A systematic review," *Frontiers in Neuroscience*, vol. 11, Art. 555, 2017.

[6] J. Varghese, C. M. V. Alen, M. Fujarski, G. S. Schlake, J. Sucker, T. Warnecke, *et al.*, "Sensor validation and diagnostic potential of smartwatches in movement disorders," *Sensors*, vol. 21, no. 9, p. 3139, 2021.

[7] R. San-Segundo, A. Zhang, A. Cebulla, S. Panev, G. Tabor, K. Stebbins, *et al.*, "Parkinson's disease tremor detection in the wild using wearable accelerometers," *Sensors*, vol. 20, no. 20, p. 5817, 2020.

[8] M. Keba, J. Helmich, G. P. Schifino, and W. Maetzler, "Assessing Parkinson's rest tremor from the wrist with inertial sensors: Validity against clinical ratings and implications for monitoring," *Journal of Clinical Medicine*, vol. 14, no. 6, Art. 2073, 2025.

[9] G. P. Schifino, M. Keba, J. Helmich, and W. Maetzler, "Smart watch sensors for tremor assessment in Parkinson's disease: Development and validation of a wrist-worn algorithm," *Sensors*, vol. 25, no. 14, Art. 4313, 2025.

[10] F. Lin, Z. Wang, H. Zhao, S. Qiu, R. Liu, X. Shi, C. Wang, and W. Yin, "Hand movement recognition and salient tremor feature extraction with wearable devices in Parkinson's patients," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 1, pp. 284–295, 2023.

[11] L. Battista and A. Romaniello, "A wearable tool for continuous monitoring of movement disorders: Clinical assessment and comparison with tremor scores," *Neurological Sciences*, vol. 42, no. 10, pp. 4241–4248, 2021.

[12] J. Varghese *et al.*, "PADS: Parkinson's Disease Smartwatch Dataset (version 1.0.0)," *PhysioNet*, 2024. [Online]. Available: https://physionet.org/content/pads/1.0.0/. DOI: 10.13026/m0w9-zx22s

[13] L. Sigcha, I. Pavon, N. Costa, S. Costa, M. Gago, P. Arezes, J. M. Lopez, and G. Arcas, "Automatic resting tremor assessment in Parkinson's disease using smartwatches and multitask convolutional neural networks," *Sensors*, vol. 21, p. 291, 2021.

[14] A. Papadopoulos, K. Kyritsis, L. Klingelhoefer, S. Bostanjopoulou, K. R. Chaudhuri, and A. Delopoulos, "Detecting parkinsonian tremor from IMU data collected in-the-wild using deep multiple-instance learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2559–2569, 2020.

[15] L. Sigcha, L. Borzi, F. Amato, *et al.*, "Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: Systematic review," *Expert Systems with Applications*, vol. 229, part A, Art. 120541, 2023.

[16] Y. K. Meng, T. Liang, H. Liu, and X. Wang, "Improved deep learning for Parkinson's diagnosis based on wearable sensors," *Electronics*, vol. 13, no. 21, p. 4638, 2024.

[17] M. Wang, J. Chen, J. Du, Y. Wu, J. Mou, and D. Camacho, "Parkinson's disease detection using multi-scale frequency-sharing channel attention network with smartwatch movement recordings," *IEEE Internet of Things Journal*, vol. 12, no. 9, pp. 14567–14577, 2025.

[18] S. Del Din, A. Godfrey, C. Mazza, S. Lord, and L. Rochester, "Free-living monitoring of Parkinson's disease: Lessons from the field," *Movement Disorders*, vol. 31, no. 9, pp. 1293–1313, 2016.

[19] S. Del Din, A. Godfrey, and L. Rochester, "Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and Parkinson's disease: Toward clinical and at home use," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 3, pp. 838–847, 2016.

[20] J. Varghese, A. Brenner, M. Fujarski, C. M. van Alen, L. Plagwitz, and T. Warnecke, "Machine learning in the Parkinson's disease smartwatch (PADS) dataset," *npj Parkinson's Disease*, vol. 10, no. 1, p. 9, 2024.

[21] A. Sanchez-Ferro, J. Rodriguez-Martin, C. Catalan-Matamoros, *et al.*, "Wearable sensor-based assessments for remotely screening early-stage Parkinson's disease," *Sensors*, vol. 24, no. 18, Art. 5637, 2024.

[22] Y. Zhang, H. Liu, J. Wang, *et al.*, "Wearable inertial sensor-based Parkinson's disease detection using hybrid feature learning and machine learning," *Sensors*, vol. 25, no. 4, Art. 4924, 2025.

[23] M. P. Burns, J. R. Matarazzo, A. Patel, *et al.*, "Wearable sensor-based assessment of Parkinson's disease using BioStampRC21 dataset: Challenges in generalization," *Bioengineering*, vol. 12, no. 1, Art. 37, 2024.

[24] J. Lipsmeier, J. Taylor, B. Kilchenmann, *et al.*, "Evaluation of smartphone-based testing to generate exploratory outcome measures in Parkinson's disease," *npj Parkinson's Disease*, vol. 11, no. 1, Art. 953, 2025.

[25] S. Zhan, M. Little, D. R. White, *et al.*, "Digital biomarker validation for Parkinson's disease using wearable sensors in real-world settings," *npj Parkinson's Disease*, vol. 11, no. 1, Art. 1214, 2025.

[26] S. Rahu, G. Ali, S. Tahseen, A. K. Baloch, and A. R. Baloch, "2DCNN_CLA: Accurate prediction of Clathrin proteins using hyperparameter optimization in deep learning and DDE profiles," *VAWKUM Transactions on Computer Sciences*, vol. 12, no. 1, pp. 163–179, 2024.