Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2143–2153
https://doi.org/10.54941/ahfe1008440

Toward Robust Patient-Level Parkinson's Disease Classification: Evaluating Stability Across Participant Compositions Using Wearable-Sensor and Questionnaire Information

Ruka Hashimoto[1] and Masaomi Kimura[1,2]

[1]Shibaura Institute of Technology, Tokyo, Japan
[2]Universiti Malaysia Kelantan, Kelantan, Malaysia

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by heterogeneous motor symptoms that vary across individuals and over time. Wearable inertial sensors provide a low-burden means of capturing motor characteristics, but reliable classification requires consideration of performance variability across participant compositions. Building on our previous temporal–frequency integrated deep learning framework, this study evaluated the robustness of patient-level PD classification and examined whether complementary questionnaire information improves classification stability. Using the Parkinson's Disease Smartwatch dataset, participants with PD and healthy controls (HC) were evaluated by stratified five-fold participant-level cross-validation. Movement signals were classified using a model integrating temporal and wavelet-based frequency features, and segment-level probabilities were averaged to obtain patient-level predictions. A 30-item non-motor symptom questionnaire was independently classified using logistic regression, and movement-based and questionnaire-based outputs were integrated with equal weighting to obtain probabilities of PD. The movement-based model achieved an accuracy of 0.854 ± 0.029, while the questionnaire-based classifier achieved 0.864 ± 0.020. Integration increased accuracy to 0.927 ± 0.027. In particular, Recall(HC) improved from 0.791 ± 0.107 for movement-only classification to 0.896 ± 0.036 after integration, indicating reduced fold-level variability, while the change in PD classification performance was comparatively modest. Evaluations using alternative participant partition seeds showed similar overall tendencies without substantial increases in variability. These results suggest that integrating complementary motor and non-motor information can improve both classification performance and stability across participant compositions, supporting more robust patient-level PD classification.

Keywords: Parkinson's disease, Wearable sensors, Deep learning, Patient-level Classification, Multimodal integration

INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder whose prevalence has increased with population aging worldwide (GBD 2016

Received August 26, 2026; Revised September 28, 2026; Accepted October 7, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Parkinson's Disease Collaborators, 2018). Its major motor symptoms include resting tremor, rigidity, bradykinesia, and postural instability, which can substantially impair activities of daily living. These motor manifestations remain central to clinical diagnosis; however, clinicopathological studies have shown that diagnosis based on clinical assessment is not always accurate (Hughes et al., 1992). Because no curative treatment has yet been established, early and objective assessment is important for appropriate symptom management and intervention. Therefore, objective methods that can quantitatively capture disease-related movement characteristics are of growing interest. Wearable inertial sensors provide a relatively low-burden means of measuring acceleration and angular velocity during motor tasks and have been increasingly investigated for objective monitoring of PD-related motor features, including tremor, bradykinesia, and gait disturbances (Moreau et al., 2023). The Parkinson's Disease Smartwatch (PADS) dataset, for example, provides wrist-worn sensor recordings from participants with PD, healthy controls (HC) and differential diagnoses (DD), enabling data-driven analysis of disease-related motor patterns (Varghese et al., 2024). Deep learning models applied to this dataset demonstrated promising performance by extracting temporal and spatial characteristics from wearable-sensor signals (Yu et al., 2024). These studies demonstrate the potential of wearable-sensor-based approaches for objective PD classification.

However, reliable PD classification requires consideration not only of model architecture but also of variability in disease-related motor characteristics and the design of the evaluation procedure. Moreau et al. (2023) highlighted substantial inter-individual variability in PD symptoms as well as temporal fluctuations within individuals, underscoring the importance of objective sensor-based assessment that accounts for such variability. Therefore, predictions obtained from short segments of a motor task may be affected by local movement fluctuations and may not fully represent the overall motor characteristics of a participant. Previous wearable-sensor studies have addressed this issue by integrating predictions obtained from multiple short events to derive subject-level decisions. For example, Atri et al. (2022) aggregated multiple short-event predictions to obtain daily subject-level classifications. In addition, when multiple records are obtained from the same participant, evaluation must ensure strict separation between participants in the training and test sets. Chaibub Neto et al. (2019) demonstrated that record-wise splitting of repeated digital-health measurements can lead to identity confounding and overestimation of predictive performance, and recommended subject-wise partitioning for diagnostic machine-learning applications. These findings indicate that patient-level evaluation should consider both the integration of local predictions and appropriate participant-level data separation when assessing whether a model provides reliable decisions for unseen individuals.

However, patient-level aggregation of wearable-sensor predictions does not necessarily eliminate performance variability across participants. Wearable inertial sensors capture motor characteristics during a limited

assessment period, and their predictions may therefore be affected by inter-individual differences in motor manifestations. In addition to movement recordings, the PADS dataset includes a self-administered questionnaire consisting of 30 binary-response items concerning PD-related non-motor symptoms (Varghese et al., 2024). Because these questionnaire responses represent information that is different from the motor characteristics captured by wearable sensors, they may provide complementary participant-level information for PD classification. Varghese et al. (2024) previously demonstrated that combining smartwatch-based movement data with non-motor symptom questionnaire information using classifier stacking improved PD-versus-HC classification compared with either modality alone. Building on this multimodal perspective, the present study focuses not only on classification performance but also on class-specific fold-level variability across different participant compositions. In addition, rather than training a meta-classifier for multimodal integration, we combine patient-level movement and questionnaire probabilities using a simple equal-weight integration strategy.

In our previous work at DEIM 2026, we proposed a deep learning framework that integrates temporal and frequency-domain characteristics of wearable-sensor signals. The framework combines an LSTM-FCN branch for extracting temporal features with a convolutional branch that processes time–frequency representations obtained using the Morlet continuous wavelet transform. Segment-level prediction probabilities are then aggregated to obtain participant-level decisions. However, our previous work primarily focused on model design and patient-level classification performance and did not systematically examine the reproducibility of classification performance across different participant partitions or the effect of incorporating complementary participant-level information.

Building on this framework, the present study focuses on the robustness of patient-level PD classification from two perspectives: reproducibility across participant partitions and multimodal information integration. First, we evaluate the movement-based model using five-fold participant-level cross-validation and further repeat the evaluation using alternative partition seeds to examine whether the overall classification tendency remains consistent across different participant compositions. We then independently classify participants using the non-motor symptom questionnaire and integrate the resulting questionnaire-based probability with the patient-level probability obtained from the movement-based model. Finally, movement-only, questionnaire-only, and integrated classifications are compared to examine whether incorporating complementary questionnaire information improves not only average classification performance but also class-specific stability across folds.

Through these analyses, this study aims to clarify whether patient-level classification performance remains reproducible across different participant partitions and whether integrating complementary questionnaire information can improve both classification performance and fold-level stability, thereby providing a more comprehensive evaluation of robustness in wearable-sensor-based PD classification.

METHODS

Dataset and Data Preparation

The Parkinson's Disease Smartwatch (PADS) dataset was used in this study (Varghese et al., 2024). The dataset contains wearable-sensor recordings and questionnaire information collected from 276 participants with Parkinson's disease (PD), 79 healthy controls (HC), and 114 participants with differential diagnoses (DD). Because this study focuses on PD-versus-HC classification, only participants with PD and HC were included in the analysis.

The wearable-sensor data consisted of acceleration and angular velocity signals recorded from both wrists while participants performed 11 different motor tasks. For each participant, the movement data were organized as a multivariate time series with 132 channels corresponding to combinations of motor tasks, wrists, sensor types, and measurement axes. The data were loaded while preserving the original channel-wise structure and converted into a time-by-channel representation for model input. The time-series signals were then divided into overlapping segments using a window length of 200 time steps and a stride of 50 time steps. To prevent data leakage, all training, validation, and test partitions were created at the participant level so that data from the same participant were never assigned to multiple subsets. Standardization parameters were estimated using only the original training data and subsequently applied to the training, validation, and test data.

During training, online data augmentation consisting of gravity removal, time warping, and axis rotation was probabilistically applied to samples from both PD and HC participants. Applying augmentation to both classes was intended to prevent augmentation-specific characteristics from becoming associated with a particular class label and being used as unintended cues for classification. The augmentation was applied with a probability of 0.5 during training. Class imbalance was handled separately using a Weighted Random Sampler, thereby separating the purpose of data augmentation from class balancing. Further details of the participant-level partitioning procedure are described in the robustness evaluation section.

Baseline Classification Model

The movement-based classifier used in this study was based on the temporal–frequency integrated deep learning framework introduced in our previous work at DEIM 2026 (Hashimoto and Kimura, 2026). The model consists of parallel temporal and frequency-domain feature-extraction components. For temporal feature extraction, an LSTM-FCN architecture was used. The convolutional layers capture local waveform patterns from the sensor signals, while the bidirectional LSTM learns temporal dependencies across the input sequence. In parallel, the frequency-domain component applies a Morlet-based continuous wavelet transform (CWT) to generate time–frequency representations, which are subsequently processed by residual convolutional layers. Spectral characteristics, including spectral flatness measure (SFM) and band power, are also incorporated to complement the learned frequency-domain representation. After feature extraction,

the temporal and frequency-domain representations are normalized and integrated before being passed to the final fully connected classification layer. During training, focal loss was used to reduce the influence of easily classified samples and place greater emphasis on difficult-to-classify samples.

In this study, this framework was used as the movement-based classifier for patient-level PD-versus-HC classification. The resulting movement-based predictions were subsequently evaluated across different participant partitions and integrated with questionnaire-based predictions, as described in the following sections.

Patient-Level Evaluation

Because predictions from individual short segments may be influenced by local movement fluctuations, patient-level evaluation was used to obtain a single classification result for each participant by aggregating information across multiple segments. Specifically, for each segment, the movement-based model produced a probability of belonging to the PD class. The segment-level probabilities from the same participant were then averaged to obtain a single patient-level probability. To determine the final classification, participants were classified as PD if their patient-level probability was 0.5 or greater, and as HC otherwise. Threshold calibration using the validation set (selecting the threshold that maximized balanced accuracy per fold) was also examined as an alternative; however, it did not consistently outperform the fixed threshold and instead produced unstable estimates across folds, likely reflecting the limited number of validation participants per fold (approximately 50–60). A fixed threshold of 0.5 was therefore adopted throughout this study to ensure stable and reproducible classification. Participants with a patient-level probability equal to or greater than this threshold were classified as PD, whereas the remaining participants were classified as HC. By aggregating predictions across multiple segments, this procedure reduces the influence of isolated segment-level predictions and enables classification based on movement information represented across the participant's recordings. In this study, this patient-level evaluation framework was used to assess the movement-based model across different participant partitions and also provided the movement-based probability used for subsequent integration with questionnaire-based predictions.

Questionnaire-Based Classification and Multimodal Integration

In addition to the movement data, the PADS dataset includes a self-administered questionnaire consisting of 30 binary-response items concerning PD-related non-motor symptoms. The questionnaire covers multiple symptom domains, including cognitive, psychological, sleep-related, and autonomic characteristics. In this study, these responses were used as complementary participant-level information that represents characteristics different from the motor information captured by wearable sensors. For questionnaire-based classification, the 30 questionnaire responses for each participant were used as input to a logistic regression classifier. Because the questionnaire consisted

of a relatively low-dimensional set of binary variables, we hypothesized that differences in response patterns between PD and HC could be reasonably represented by a linear combination of the questionnaire items in log-odds space. Logistic regression was therefore selected as a simple model for estimating the probability of PD from these responses. For each fold, the classifier was trained using only the participants assigned to the training set, thereby preventing information from the validation and test participants from being used during model fitting. The linear output of the logistic regression model was transformed by the sigmoid function to obtain the probability of belonging to the PD class for each participant. For multimodal integration, the patient-level probability obtained from the movement-based model and the probability obtained from the questionnaire-based classifier were first converted to logit values using the log-odds transformation. The two logits were then averaged with equal weighting, and the resulting value was transformed back into a probability using the sigmoid function. This yielded a single integrated PD probability for each participant. This integration was designed to combine complementary motor and non-motor information while giving equal contribution to the two modalities.

Robustness Evaluation Under Different Patient Partitions

To evaluate the reproducibility of classification performance across different participant compositions, stratified five-fold cross-validation was performed at the participant level. For each fold, participants were divided into training, validation, and test sets at a ratio of 3:1:1 while approximately preserving the class distribution. All data belonging to the same participant were assigned to only one subset so that no participant was shared across the training, validation, and test sets. For performance evaluation, accuracy, precision, recall, and F1-score were calculated at the patient level. Precision, recall, and F1-score were calculated separately for HC and PD to evaluate class-specific classification performance. The mean and standard deviation of each metric were then calculated across the five test folds to summarize both overall classification performance and fold-level variability. The same evaluation procedure was applied to the movement-only, questionnaire-only, and integrated classifications.

In addition, the participant-level partitioning procedure was repeated using different random seeds while maintaining the same 3:1:1 training–validation–test ratio and class-stratification procedure. This changed the combinations of participants assigned to each subset while preserving the overall evaluation framework. The resulting patient-level classification performance was compared across the different partition seeds to examine whether the observed performance tendency was reproducible and not dependent on a particular participant partition.

RESULTS

Movement-Based Patient-Level Classification Performance

To examine patient-level classification performance based on movement information, we first evaluated the movement-based model using stratified five-fold cross-validation. The fold-wise results are shown in Table 1.

Table 1: Fold-wise patient-level classification performance of the movement-based model.

Fold	Accuracy	Precision (HC)	Recall (HC)	F1-Score (HC)	Precision (PD)	Recall (PD)	F1-Score (PD)
1	0.852	0.619	0.929	0.743	0.975	0.830	0.897
2	0.867	0.750	0.643	0.692	0.896	0.935	0.915
3	0.883	0.688	0.846	0.759	0.955	0.894	0.923
4	0.800	0.529	0.692	0.600	0.907	0.830	0.867
5	0.867	0.647	0.846	0.733	0.953	0.872	0.911
Mean ± Std	0.854 ± 0.029	0.647 ± 0.074	0.791 ± 0.107	0.705 ± 0.057	0.937 ± 0.030	0.872 ± 0.040	0.903 ± 0.020

The movement-based model achieved a mean patient-level accuracy of 0.854 ± 0.029 across the five folds. Precision(PD), Recall(PD), and F1-score(PD) were 0.937 ± 0.030, 0.872 ± 0.040, and 0.903 ± 0.020, respectively. In contrast, the corresponding HC metrics showed greater variation across folds. Recall(HC) ranged from 0.643 to 0.929, with a mean of 0.791 ± 0.107, while Precision(HC) and F1-score(HC) were 0.647 ± 0.074 and 0.705 ± 0.057, respectively. In particular, Recall(HC) varied substantially across folds, whereas Recall(PD) remained comparatively stable. For example, Recall(HC) decreased to 0.643 in Fold 2 but reached 0.929 in Fold 1. These results indicate that although the movement-based model achieved consistent overall classification performance, class-specific performance, particularly for HC participants, remained sensitive to the composition of the evaluation fold.

Therefore, the movement-based results suggest that patient-level performance should be examined not only in terms of mean accuracy but also in terms of class-specific variability across participant compositions. This observation provides the basis for subsequently examining whether complementary questionnaire information can improve both classification performance and fold-level stability.

Effect of Integrating Movement and Questionnaire Information

To examine the effect of incorporating complementary questionnaire information, patient-level classification performance was compared among the movement-only, questionnaire-only, and integrated conditions. The mean and standard deviation across the five folds are shown in Table 2.

Table 2: Patient-level classification performance for movement, questionnaire, and integrated information (Mean ± Std).

Input	Class	Accuracy	Precision	Recall	F1-Score
Movement	HC	0.854 ± 0.029	0.647 ± 0.073	0.791 ± 0.107	0.705 ± 0.057
	PD	0.854 ± 0.029	0.937 ± 0.030	0.872 ± 0.040	0.903 ± 0.020
Questionnaire	HC	0.864 ± 0.020	0.640 ± 0.031	0.879 ± 0.078	0.740 ± 0.046
	PD	0.864 ± 0.020	0.962 ± 0.023	0.859 ± 0.011	0.908 ± 0.013
Movement + Questionnaire	HC	0.927 ± 0.027	0.805 ± 0.081	0.896 ± 0.036	0.846 ± 0.054
	PD	0.927 ± 0.027	0.969 ± 0.011	0.936 ± 0.030	0.952 ± 0.018

The movement-based model achieved a mean accuracy of 0.854 ± 0.029, whereas the questionnaire-based classifier achieved 0.864 ± 0.020. When the movement-based and questionnaire-based predictions were integrated, the mean accuracy increased to 0.927 ± 0.027. A similar improvement was observed in the class-specific metrics. For HC participants, F1-score increased from 0.705 ± 0.057 for the movement-only condition and 0.740 ± 0.046 for the questionnaire-only condition to 0.846 ± 0.054 after integration. Recall(HC) also increased from 0.791 ± 0.107 and 0.879 ± 0.078, respectively, to 0.896 ± 0.036. For PD participants, the integrated model achieved a Precision(PD) of 0.969 ± 0.011, Recall(PD) of 0.936 ± 0.030, and F1-score(PD) of 0.952 ± 0.018. These values were higher overall than those obtained from either modality alone, indicating that the integrated model improved classification performance for both classes. To further examine the fold-level stability of the integrated model, its classification performance in each of the five folds is shown in Table 3. In particular, fold-wise analysis showed that the improvement after integration was accompanied by reduced variability in HC classification. Recall(HC) for the integrated model ranged from 0.846 to 0.929 across the five folds, whereas that of the movement-based model ranged from 0.643 to 0.929. Accordingly, the standard deviation of Recall(HC) decreased from 0.107 in the movement-only condition to 0.036 in the integrated condition. Recall(PD) also remained relatively stable across folds, ranging from 0.894 to 0.979. These results suggest that incorporating questionnaire information not only improved average classification performance but also contributed to more stable class-specific performance across different participant compositions. In addition, evaluations using alternative participant partition seeds showed similar overall performance tendencies, with no substantial increase in fold-level variability. This result suggests that the observed classification performance was not specific to a single participant partition and that the integrated model exhibited a certain degree of robustness to changes in participant composition.

Table 3: Fold-wise patient-level classification performance of the integrated model.

Fold	Accuracy	Precision (HC)	Recall (HC)	F1-Score (HC)	Precision (PD)	Recall (PD)	F1-Score (PD)
1	0.967	0.929	0.929	0.929	0.979	0.979	0.979
2	0.917	0.800	0.857	0.828	0.956	0.935	0.945
3	0.900	0.706	0.923	0.800	0.977	0.894	0.933
4	0.900	0.733	0.846	0.786	0.956	0.915	0.935
5	0.950	0.857	0.923	0.889	0.978	0.957	0.968
Mean ± Std	0.927 ± 0.027	0.805 ± 0.081	0.896 ± 0.036	0.846 ± 0.054	0.969 ± 0.011	0.936 ± 0.030	0.952 ± 0.018

DISCUSSION

The results of this study highlight the importance of evaluating wearable-sensor-based PD classification not only in terms of average performance but also in terms of class-specific variability across participant compositions.

In the movement-only condition, the model achieved a mean accuracy of 0.854, while Recall(HC) showed relatively large variation across folds, with a mean and standard deviation of 0.791 ± 0.107. In contrast, Recall(PD) remained comparatively stable. One possible reason for this difference is the smaller number of HC participants in the dataset, meaning that a small number of misclassified HC participants can produce relatively large changes in recall. In addition, movement characteristics captured during a limited assessment period may differ substantially among participants. Therefore, the observed fold-level variability is likely to reflect a combination of participant composition, class imbalance, and inter-individual differences in movement characteristics rather than a single factor. Integrating questionnaire information with movement-based predictions improved both overall classification performance and class-specific stability. The integrated model increased mean accuracy from 0.854 for the movement-only condition and 0.864 for the questionnaire-only condition to 0.927. In particular, Recall(HC) increased to 0.896 ± 0.036, while its standard deviation decreased substantially compared with the movement-only condition. A similar improvement was observed in F1-score for both HC and PD. These results suggest that the questionnaire information provided complementary information that was not fully captured by movement signals alone. Wearable sensors primarily characterize motor behavior during specific assessment tasks, whereas the questionnaire represents participant-reported non-motor characteristics across different symptom domains. Combining these two sources of information may therefore compensate for cases in which movement-based predictions alone are less reliable. This tendency is consistent with Varghese et al. (2024), who also reported improved PD-versus-HC classification by combining smartwatch and questionnaire information. However, the present study extends this perspective by explicitly examining class-specific fold-level variability and reproducibility across different participant partitions using a different movement model and integration strategy.

In addition, evaluations using alternative participant partition seeds showed similar overall performance tendencies without substantial increases in variability. This suggests that the improvement observed after integrating movement and questionnaire information was not limited to a specific participant partition. Although these results do not establish robustness under all possible population or evaluation conditions, they indicate that the integrated approach exhibited a certain degree of stability against changes in participant composition. Taken together with the fold-wise analysis, this finding supports the importance of evaluating both average performance and its variability when assessing the reliability of patient-level classification.

Several limitations of this study should also be considered. First, the number of HC participants was smaller than the number of PD participants, which may have increased the sensitivity of HC-specific metrics to individual misclassifications. Second, this study focused on PD-versus-HC classification using a single dataset, and the observed stability may not directly generalize to other populations or to more difficult differential-diagnosis settings.

Third, the questionnaire information was integrated using a relatively simple equal-weight combination of prediction probabilities. Further investigation of alternative integration strategies and external validation will therefore be necessary to determine whether the observed improvements can be maintained under broader evaluation conditions. Fourth, although a fixed classification threshold of 0.5 was adopted based on the instability observed with per-fold validation-based calibration, alternative calibration strategies that account for validation set size may warrant further investigation in larger datasets.

CONCLUSION

This study investigated the robustness of wearable-sensor-based PD classification by examining patient-level performance across different participant compositions and the effect of integrating movement and questionnaire information. Using a temporal–frequency integrated deep learning model, movement-based classification was first evaluated under stratified five-fold cross-validation, and complementary non-motor questionnaire information was subsequently incorporated at the patient level.

The movement-based model achieved a mean accuracy of 0.854, while class-specific analysis revealed relatively large fold-level variability in HC classification. Integrating movement and questionnaire information increased the mean accuracy to 0.927 and improved classification performance for both HC and PD participants. In particular, Recall(HC) increased from 0.791 ± 0.107 in the movement-only condition to 0.896 ± 0.036 after integration, indicating reduced fold-level variability in HC classification. Evaluations using alternative participant partition seeds also showed similar overall performance tendencies, suggesting that the observed results were not limited to a single participant partition.

These findings suggest that evaluating averaged classification performance alone may be insufficient for assessing the robustness of wearable-sensor-based PD classification. Considering class-specific variability across participant compositions and incorporating complementary participant-level information may contribute to more stable patient-level classification. Future work will investigate alternative multimodal integration strategies and evaluate the proposed framework across broader participant populations and more challenging classification settings, including differential diagnosis.

REFERENCES

Atri R, Urban K, Marebwa B, Simuni T, Tanner C, Siderowf A, Frasier M, Haas M, Lancashire L. (2022) Deep learning for daily monitoring of Parkinson's disease outside the clinic using wearable sensors. *Sensors*, 22(18), 6831.

Chaibub Neto E, Pratap A, Perumal TM, Tummalacherla M, Snyder P, Bot BM, Trister AD, Friend SH, Mangravite L, Omberg L. (2019) Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *npj Digital Medicine*, 2, 99.

GBD 2016 Parkinson's Disease Collaborators. (2018) Global, regional, and national burden of Parkinson's disease, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, 17(11), 939–953.

Hashimoto R, Kimura M. (2026) Deep learning-based Parkinson's disease classification by integrating time-series and frequency features. In: The 18th Forum on Data Engineering and Information Management (DEIM 2026).

Hughes AJ, Daniel SE, Kilford L, Lees AJ. (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(3), 181–184.

Johnson S, Kantartjis M, Severson J, Dorsey R, Adams JL, Anderson D, et al. (2024) Wearable sensor-based assessments for remotely screening early-stage Parkinson's disease. *Sensors*, 24(17), 5637.

Moreau C, Rouaud T, Grabli D, Benatru I, Remy P, Marques AR, Drapier S, Mariani LL, Roze E, Devos D, Dupont G, Bereau M, Fabbri M. (2023) Overview on wearable sensors for the management of Parkinson's disease. *npj Parkinson's Disease*, 9, 153.

Pollard T, Moody BE, Lehman L, Gow B, Fernandes C, Xie C, Johnson A, Mark RG, Heldt T. (2026) PhysioNet as a global platform for biomedical research. Nature Health. https://doi.org/10.1038/s44360-026-00096-z.

Varghese J, Brenner A, Fujarski M, van Alen CM, Plagwitz L, Warnecke T. (2024) Machine learning in the Parkinson's disease smartwatch (PADS) dataset. npj Parkinson's Disease, 10, 9.

Varghese J, Brenner A, Plagwitz L, van Alen C, Fujarski M, Warnecke T. (2024) PADS - Parkinsons Disease Smartwatch dataset (version 1.0.0). PhysioNet. https://doi.org/10.13026/m0w9-zx22.

Yu J, Meng K, Liang T, Liu H, Wang X. (2024) Improved deep learning for Parkinson's diagnosis based on wearable sensors. Electronics, 13(23), 4638.