Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 424–434
https://doi.org/10.54941/ahfe1008275

Appropriate Reliance in Wearable AI for Parkinson's Differential Assessment: Calibration, Selective Referral, and Subgroup Reliability

Mohammad Mahdi Nazari[1], Negar Heidari Matin[1], and Yaser Banad[2]

[1]Christopher C. Gibbs College of Architecture, University of Oklahoma, Norman, OK 73019, USA
[2]School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019, USA

ABSTRACT

Smartwatch-based machine learning for Parkinson's disease (PD) is often evaluated as forced-choice classification against healthy controls, although clinically similar disorders create the harder decision boundary. This retrospective study evaluates a human-centered selective-prediction framework on the Parkinson's Disease Smartwatch (PADS) dataset. The primary task included 276 participants with PD and 114 with differential diagnoses (DD); 79 healthy controls formed a secondary benchmark. From bilateral accelerometer and gyroscope recordings across 11 motor tasks, 3,366 participant-level time-domain, spectral, magnitude, and wrist-asymmetry features were derived, with demographic variables excluded from model inputs. Regularized logistic regression, Extra Trees, and radial-basis-function support-vector models were selected by nested five-fold outer and three-fold inner cross-validation and sigmoid calibrated. We evaluated discrimination, calibration, selective referral, subgroup uncertainty burden, class-conditional split conformal prediction, sensor/task ablation, and leave-one-DD-subtype-out shift. For PD versus DD, calibrated balanced accuracy was 0.674 and AUROC 0.714; Brier score and expected calibration error improved to 0.180 and 0.047. Referring the 20% least-confident cases reduced retained-case error from 30.5% to 25.0%. At 90% nominal conformal coverage, empirical coverage was 94.6%, but 76.7% of outputs were ambiguous. Under unseen DD subtypes, confidence referral did not consistently remove false-PD predictions, whereas conformal ambiguity increased. The results support evaluating wearable AI as a referral-and-review system rather than a forced-choice classifier.

Keywords: Parkinson's disease, Wearable technology, Human-centered design, Selective prediction, Conformal prediction, Probability calibration, Human-AI collaboration

INTRODUCTION

Wearable inertial sensors offer a way to quantify motor behavior using consumer devices and have been investigated for the assessment of Parkinson's disease (PD) and related movement disorders (Varghese et al., 2021). Recent reviews describe diverse machine-learning approaches to PD assessment and continuing challenges in validation and generalizability (Shokrpour

Received October 14, 2026; Revised November 21, 2026; Accepted November 30, 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/

et al., 2025). However, distinguishing PD from healthy controls does not establish the ability to distinguish PD from disorders with overlapping motor signs. The Parkinson's Disease Smartwatch (PADS) resource is valuable because it contains PD, healthy controls, and multiple differential diagnoses (DD), together with bilateral smartwatch accelerometry and gyroscope recordings collected under a standardized examination protocol (Varghese et al., 2024a, 2024b). The original PADS analysis showed that PD-versus-DD discrimination is more difficult than PD-versus-healthy classification (Varghese et al., 2024a).

This harder boundary motivates a different evaluation question: can a model recognize uncertainty in its predictions and defer cases for human review? Such an evaluation requires more than AUROC. Probability calibration concerns whether predicted probabilities agree with observed outcomes (Guo et al., 2017), while selective prediction addresses the trade-off between retained predictions, their reliability, and the burden placed on human experts (Feng et al., 2023). Conformal prediction provides prediction sets with coverage guarantees under specified assumptions, including exchangeability; these guarantees do not automatically hold under diagnostic shift (Angelopoulos and Bates, 2023). Related clinical work has examined subgroup-specific uncertainty in medical imaging (Sun et al., 2025) and combined calibration, conformal prediction, and cost-aware deferral for sepsis triage under temporal shift (Kwon and Kim, 2026). Within remote PD assessment, device-type and handedness biases reported in computer-based tasks further motivate examining technical and participant-level differences (Tumpa et al., 2025). These findings support evaluating subgroup reliability and referral burden, but do not establish the effectiveness or safety of such approaches for smartwatch-based PD differential assessment.

Accordingly, this study evaluates a participant-level, movement-only framework for PD versus DD. We examine whether PD-versus-DD performance is lower than the PD-versus-healthy benchmark, whether post hoc calibration improves probability quality, whether referral of low-confidence cases reduces retained-case error, and whether leave-one-DD-subtype-out testing exposes limitations not visible in random participant-level cross-validation. The contribution is an integrated, human-centered retrospective evaluation rather than a state-of-the-art accuracy claim: discrimination, calibration, selective referral, subgroup uncertainty burden, class-conditional conformal prediction, compact sensor/task ablation, and diagnostic-subtype shift are analyzed in one workflow. This evaluation characterizes the reliability and workload implications of a potential referral-and-review system without establishing prospective clinical safety or benefit.

THEORY AND METHODS

Study Cohort and Signal Representation

PADS v1.0.0 was used as a retrospective, de-identified secondary dataset. The primary analysis contained 390 participants (276 PD and 114 DD); 79 healthy controls were reserved for a secondary PD-versus-healthy benchmark. The DD group comprised 60 other movement disorders, 28 essential tremor,

15 atypical parkinsonism, and 11 multiple sclerosis cases. Demographic variables, questionnaire responses, medication information, and handedness were excluded from model predictors and were used only where specified for post-prediction auditing.

The official preprocessed movement files contained 132 channels and 976 time points per channel, representing 11 retained motor tasks, two wrists, acceleration and gyroscope/rotation sensing, and three axes at 100 Hz. For each axis signal, vector magnitude, and paired left-right magnitude representation, 17 descriptive, dynamic, and spectral summaries were computed. The feature set included central tendency and dispersion, root-mean-square amplitude, robust spread, peak-to-peak range, skewness, kurtosis, mean absolute first difference, centered zero-crossing rate, dominant frequency from 2–12 Hz, relative power in 0.5–3, 3–7, 7–12, and 12–20 Hz, spectral entropy over 0.5–20 Hz, and bilateral contrasts. This produced 3,366 participant-level predictors.

Learning Pipeline and Analytical Assumptions

Model development used five-fold stratified outer cross-validation. Within each outer training set, a three-fold stratified inner search selected the model family, univariate feature count, and hyperparameters among regularized logistic regression, Extra Trees, and radial-basis-function support-vector models. Median imputation and ANOVA-F SelectKBest feature selection were fitted only on the relevant training partition, with k in {50, 100, 200}; class weights were balanced. The selected estimator was sigmoid-calibrated using training-only cross-validation. Out-of-fold predictions were pooled only after every participant had been predicted by a model that had not been trained on that participant.

The main analytical assumptions and conventions were: (A1) participant records were treated as independent observational units; (A2) preprocessing steps that could learn from data were restricted to training partitions to prevent leakage; (A3) protected or demographic attributes were not model inputs; (A4) subgroup results were descriptive safety audits rather than powered fairness claims; (A5) retrospective risk-coverage operating points were descriptive and are not prospective deployment thresholds; and (A6) class-conditional conformal validity relies on exchangeability, an assumption intentionally violated in the held-out-diagnostic-subtype stress test.

Metrics, Calibration, and Selective Referral

Balanced accuracy was the primary discrimination metric because the PD and DD classes were unequal in size. Let TP, TN, FP, and FN denote the usual confusion-matrix counts. Balanced accuracy was as below:

$$BA = \frac{1}{2}\left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right] \tag{1}$$

Probability quality was assessed by the Brier score, with calibrated PD probability p_i and binary target y_i in {0,1}.

$$Brier = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(p_i - y_i)^2 \tag{2}$$

and by expected calibration error (ECE) over adaptive quantile bins B_m.

$$ECE = \sum_{m=1}^{M}\left(\frac{|B_m|}{n}\right)\left|acc(B_m) - conf(B_m)\right| \tag{3}$$

The sigmoid-calibrated model used a fold-specific decision threshold equal to the PD prevalence in the corresponding outer training data. For outer fold f, with training counts $N_{PD,f}$ and $N_{DD,f}$.

$$t_f = \frac{N_{PD,f}}{N_{PD,f} + N_{DD,f}} \tag{4}$$

This prevalence-based threshold avoids treating 0.5 as universally appropriate in the imbalanced cohort. Retrospective referral confidence was defined as the class-normalized distance from the relevant decision boundary:

$$c_i = \frac{p_i - t}{1 - t} if\ p_i \geq t;\ \ c_i = \frac{t - p_i}{t} if\ p_i < t \tag{5}$$

Cases were ranked by c_i, and the least-confident 0%, 5%, 10%, 15%, 20%, or 30% were referred. For a retained-confidence threshold τ, selective risk was the observed classification error among retained cases:

$$R_{sel}(\tau) = [\Sigma_i\ I(c_i \geq \tau)\ I(\hat{y}_i \neq\ y_i)] / [\Sigma_i I(c_i \geq \tau)] \tag{6}$$

Subgroup Audit and Class-Conditional Conformal Prediction

At the 80% retained-coverage operating point, recorded gender, age (<65 or >=65 years), and BMI (<25, 25–29.9, or >=30 kg/m²) were audited. Handedness was not stratified because only 32 participants were left-handed; race and ethnicity were unavailable. If r_g denotes the referral rate of subgroup g, uncertainty-burden gap (UBG) was

$$UBG = max_g(r_g) - min_g(r_g) \tag{7}$$

A separate class-conditional split-conformal analysis used a proper-training subset and a 25% calibration subset within each outer fold. For candidate class y, the nonconformity score was $1 - p_y(x)$. With n_y calibration examples in class y and $\alpha = 0.10$, the finite-sample higher order statistic and prediction set were defined as:

$$\begin{aligned} s_i(y) = 1 - p_y(x_i), k_y = ceil[(n_y + 1)(1 - \alpha)];\ \ q_y = s^{(y)}_{(ky)}; \Gamma(x) \\ = \{y : 1 - p_y(x) \leq q_y\} \end{aligned} \tag{8}$$

where $s^{(y)}_{(ky)}$ denotes the k_y-th ordered calibration score for class y. Singleton sets were treated as eligible for automation; two-label or empty sets were referrals. Class-conditional thresholds were used so that minority-class coverage would not be hidden by marginal coverage (Angelopoulos and Bates, 2023; Feng et al., 2023).

Ablation, Task Importance, and Diagnostic Shift

Paired outer folds compared all signals with acceleration-only, rotation-only, left-wrist-only, and right-wrist-only feature sets using the same regularized logistic procedure. Grouped permutation importance measured the decrease in outer-test balanced accuracy after jointly permuting all features belonging to one motor task. For diagnostic-shift stress testing, each DD subtype was excluded in turn; the model was trained on PD plus the remaining DD subtypes and evaluated only on the unseen subtype. A training-only calibration split supplied conventional confidence-referral and conformal thresholds. Because this shift breaks exchangeability, conformal coverage under the stress test is descriptive rather than guaranteed.

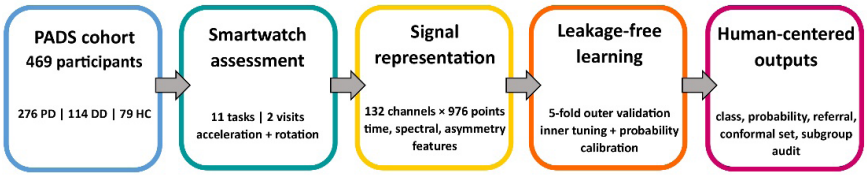


Figure 1: Participant-level PADS workflow from bilateral smartwatch signals to leakage-free learning, calibration, referral, subgroup auditing, conformal prediction, and diagnostic-shift evaluation.

RESULTS

Primary Discrimination and Calibration

Across the five outer folds, the inner search selected RBF support-vector models twice and regularized logistic regression three times. Calibrated outer-fold balanced accuracy ranged from 0.613 to 0.731. For the pooled PD-versus-DD analysis, sigmoid calibration improved probability quality while leaving ranking discrimination essentially unchanged: Brier score fell from 0.202 to 0.180, ECE from 0.157 to 0.047, and log loss from 0.591 to 0.541, whereas AUROC changed from 0.723 to 0.714 (Table 1). The calibrated model achieved sensitivity 0.725 and specificity 0.623. The easier PD-versus-healthy benchmark reached balanced accuracy 0.791 and AUROC 0.877.

Table 1: Primary PD-versus-DD out-of-fold performance.

Metric	Raw	Calibrated	95% CI, Calibrated
Balanced accuracy	0.688	0.674	0.624–0.725
Sensitivity	0.754	0.725	0.674–0.775
Specificity	0.623	0.623	0.535–0.711
AUROC	0.723	0.714	0.651–0.772
Brier score	0.202	0.180	0.166–0.192
ECE	0.157	0.047	0.043–0.103

Figure 2 further characterizes discrimination and probability reliability in the primary PD-versus-DD task. Raw and sigmoid-calibrated predictions

showed similar discrimination, with AUROC values of 0.723 and 0.714, respectively, indicating that calibration did not materially alter participant ranking. In contrast, the calibrated reliability curve moved substantially closer to the ideal identity line, consistent with reductions in Brier score from 0.202 to 0.180 and expected calibration error from 0.157 to 0.047. Calibration therefore improved the reliability of model confidence rather than the underlying discrimination between PD and differential diagnoses.

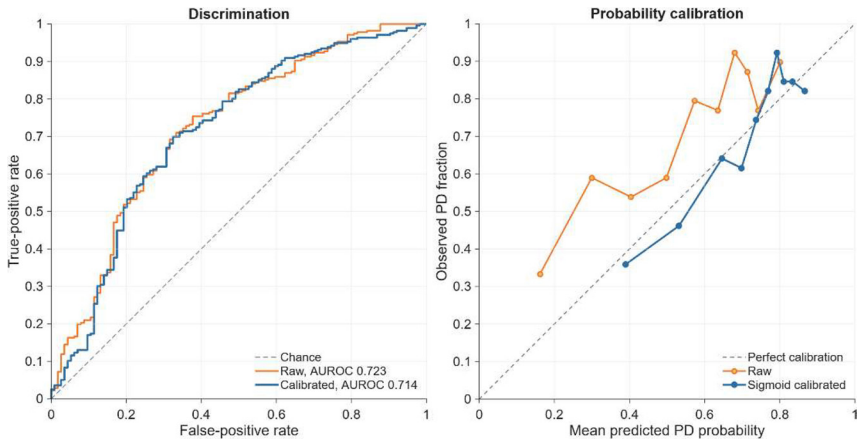


Figure 2: ROC discrimination and probability calibration for PD versus differential diagnoses.

Figure 3 provides a participant-level view of the remaining classification difficulty. Calibrated PD probabilities were generally higher among participants with PD, but substantial overlap between the PD and DD distributions remained, particularly around the decision threshold. Using the calibrated decision rule, 200 of 276 participants with PD were correctly classified and 76 were classified as DD, whereas 71 of 114 DD participants were correctly classified and 43 were classified as PD. These counts correspond to sensitivity of 0.725 and specificity of 0.623. The overlapping probability distributions illustrate why forced-choice prediction remains imperfect and motivate selective referral when movement features do not provide strong separation between the two diagnostic classes.

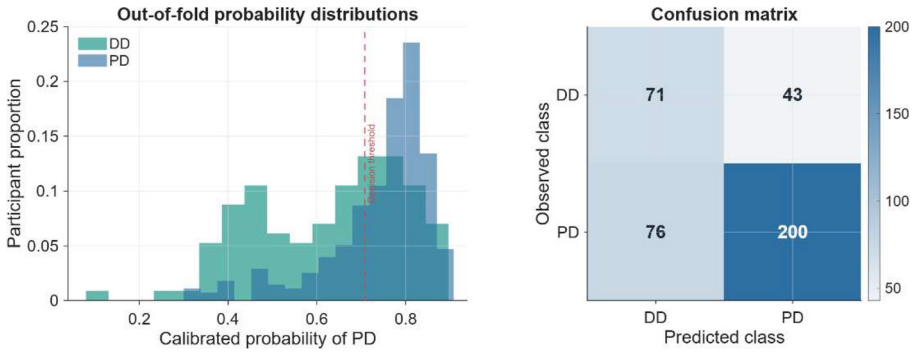


Figure 3: Calibrated probability distributions and confusion matrix for PD versus differential diagnoses.

Given this overlap in calibrated class probabilities, we next examined whether referring the least-confident cases could reduce error among predictions retained for automated decision support.

Selective Referral and Subgroup Uncertainty Burden

Here, the Figure 4, shows the results that demonstrate at 80% retained coverage, 78 of 390 participants were referred. Error among retained cases decreased from 30.5% to 25.0%, an absolute reduction of 5.5 percentage points (18.1% relative), while balanced accuracy increased from 0.674 to 0.703. The improvement changed the error mix: sensitivity rose to 0.809 while specificity decreased slightly to 0.598. Table 2 shows referral-rate gaps were modest-2.7 percentage points by recorded gender, 2.7 by age, and 1.9 by BMI-but similar referral rates did not imply similar calibration quality. Female participants had Brier 0.221 and ECE 0.133 versus 0.157 and 0.061 for male participants.

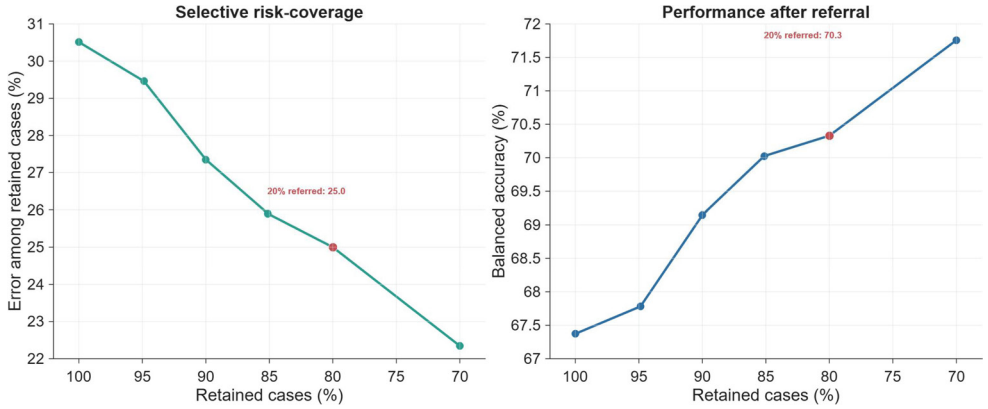


Figure 4: Selective risk-coverage and subgroup uncertainty burden at the 80% overall retained-coverage operating point.

Table 2: Compact subgroup audit at 80% retained coverage.

Group	n	Referred	BA	Brier	ECE
Female	138	21.7%	0.652	0.221	0.133
Male	252	19.0%	0.675	0.157	0.061
Age < 65	178	18.5%	0.710	0.205	0.130
Age >= 65	212	21.2%	0.639	0.158	0.074
BMI < 25	153	20.9%	0.662	0.195	0.076
BMI 25 – 29.9	153	19.6%	0.637	0.189	0.079
BMI >= 30	84	19.0%	0.801	0.135	0.117

Conformal Prediction and Sensor/Task Ablation

At 90% nominal class-conditional conformal coverage, empirical coverage was 94.6% overall, 94.7% for DD, and 94.6% for PD. Coverage was therefore conservative and balanced across the two classes, but the usability

cost was substantial: only 23.3% of outputs were singletons and 76.7% contained both labels. Singleton predictions achieved balanced accuracy 0.780 with 23.1% error. Thus, nominal validity alone did not imply an efficient human-AI workload.

The compact sensor/wrist ablation showed that rotation-only features outperformed acceleration-only features (balanced accuracy 0.656 versus 0.611). Right-wrist-only features achieved balanced accuracy 0.677 and AUROC 0.731, slightly above the all-signal logistic ablation (0.666 and 0.711), whereas left-wrist-only performance was lower (Table 3). Grouped permutation importance ranked HoldWeight as the most influential task (mean balanced-accuracy decrease 0.067 when permuted), followed by DrinkGlas (0.023) and CrossArms (0.016). The HoldWeight standard deviation was 0.057, and fold variation plus multiple comparisons preclude a sensor-reduction conclusion.

Table 3: Paired-fold sensor and wrist ablation.

Feature Set	Balanced Accuracy	AUROC
All signals	0.666	0.711
Acceleration only	0.611	0.663
Rotation only	0.656	0.705
Left wrist only	0.648	0.698
Right wrist only	0.677	0.731

Diagnostic-Subtype Shift

Figure 5 shows the False-PD rates on completely unseen DD subtypes ranged from 39.3% for essential tremor to 54.5% for multiple sclerosis. Conventional confidence referral did not consistently remove these errors: among retained cases, false-PD rate increased to 70.0% for other movement disorders, 70.0% for atypical parkinsonism, and 71.4% for multiple sclerosis. The model could therefore remain confident in a clinically undesirable direction under shift. Conformal non-singleton rates rose to 57.1%-91.7%, making the shift more visible, although most cases were then referred and formal coverage guarantees no longer applied.

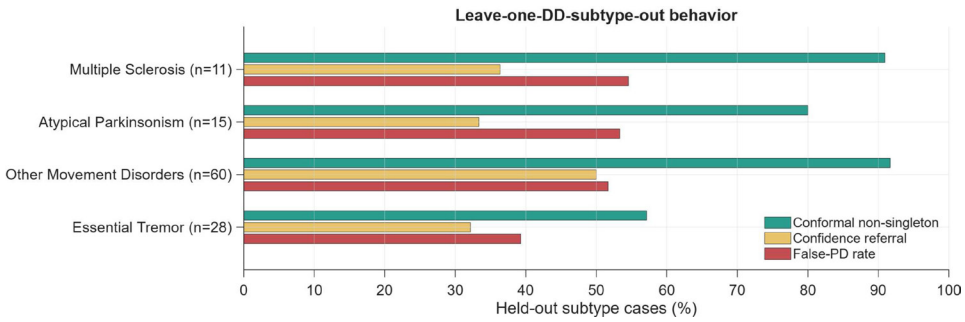


Figure 5: Leave-one-differential-diagnosis-subtype-out stress test. Conventional confidence referral can retain confidently wrong false-PD predictions, whereas conformal ambiguity increases under unseen-subtype shift.

DISCUSSION

The primary finding is not that smartwatch signals are ready for autonomous differential diagnosis. The calibrated PD-versus-DD balanced accuracy of 0.674 is substantially below the PD-versus-healthy benchmark and should instead be interpreted as a realistic boundary at which uncertainty handling becomes important. Within that boundary, calibration improved confidence reliability rather than case ranking: Brier score, expected calibration error (ECE), and log loss improved while AUROC changed little, consistent with post-hoc calibration improving probabilistic reliability for downstream decisions (Guo et al., 2017). Selective referral then converted calibrated confidence into an operational action, reducing error among retained cases at the cost of human review.

This trade-off is intrinsically human-centered. A 20% referral policy means that 78 of 390 primary cases require human review, while class-conditional conformal prediction was substantially more conservative, producing ambiguous sets for 76.7% of participants at 90% nominal coverage. Subgroup results add a second caution: referral-rate gaps were small, yet calibration quality differed appreciably between recorded female and male participants. Equal referral frequency therefore cannot be interpreted as evidence of equal uncertainty quality or clinical equity; the subgroup analysis is descriptive and should motivate rather than settle fairness questions (Sun et al., 2025; Tumpa et al., 2025).

The diagnostic-shift experiment provides the clearest safety warning. Distance from a familiar decision boundary is not a general measure of epistemic uncertainty: an unseen disorder may occupy a feature region that the trained model strongly associates with PD, allowing a confidently wrong prediction to remain after borderline cases are removed. Conventional confidence referral therefore did not consistently protect against false-PD decisions under unseen-subtype shift. In contrast, conformal non-singleton rates increased to 57.1%–91.7%, making the shift more visible, although the finite-sample conformal coverage guarantee depends on exchangeability and does not apply under this deliberate shift (Angelopoulos and Bates, 2023; Kwon and Kim, 2026). A practical interface should therefore combine case-level uncertainty with shift monitoring and permit automation to be suspended when the input population changes.

The retained ablation analysis provides an engineering complement to the referral story. Rotation-only sensing preserved more discrimination than acceleration-only sensing, while the right-wrist-only representation slightly exceeded the all-signal logistic ablation. These observations may inform future device simplification, but they do not establish sensor equivalence or a preferred wrist because the comparisons were exploratory, fold-dependent, and not powered for noninferiority. HoldWeight was the strongest grouped task by permutation importance, but its variability argues against reducing the examination protocol on this dataset alone.

Several limitations constrain interpretation. PADS is a single-source, cross-sectional cohort acquired under a standardized examination protocol, with no independent geographic, temporal, or device-level test cohort. The smallest DD subtypes were multiple sclerosis (n = 11) and atypical

parkinsonism (n = 15), making shift estimates imprecise. Race and ethnicity were unavailable, recorded gender was binary, and handedness was too imbalanced for stable subgroup analysis. The 80% risk-coverage operating point was selected retrospectively and must be fixed or adapted using separate validation data before deployment. Hand-engineered features may discard temporal structure, and no clinician-in-the-loop experiment was performed. Future validation should preregister the operating point in a multicenter DD-enriched cohort and compare unaided review, forced-choice AI, calibrated referral, and conformal support using diagnostic accuracy, review time, escalation appropriateness, confidence calibration, and subgroup workload (Collins et al., 2024).

CONCLUSION

Bilateral smartwatch signals provided useful but incomplete information for PD-versus-DD assessment. Calibration improved probability reliability, and selective referral reduced retained-case error, but subgroup variation and unseen diagnostic subtypes revealed important limitations for trustworthy use. Class-conditional conformal prediction made uncertainty more visible, although at the cost of substantial referral burden, while sensor and task ablation findings remained exploratory. Wearable AI should therefore be evaluated as a referral-and-review system in which discrimination, calibration, human workload, subgroup burden, and behavior under diagnostic shift jointly determine appropriate reliance.

ETHICS AND DATA AVAILABILITY

This work is a secondary analysis of the publicly available, de-identified PADS dataset; no participants were recruited and no re-identification was attempted. Primary ethics and informed-consent procedures are reported by the dataset authors (Varghese et al., 2021; Varghese et al., 2024a). PADS v1.0.0 is available through PhysioNet (Varghese et al., 2024b).

REFERENCES

Angelopoulos, A.N. and Bates, S. (2023) 'Conformal Prediction: A Gentle Introduction', Foundations and Trends in Machine Learning, 16(4), pp. 494–591. https://doi.org/10.1561/2200000101

Feng, J., Sondhi, A., Perry, J. and Simon, N. (2023) 'Selective Prediction-Set Models with Coverage Rate Guarantees', Biometrics, 79(2), pp. 811–825. https://doi.org/10.1111/biom.13612

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) 'On Calibration of Modern Neural Networks', Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, 70, pp. 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Kwon, H. and Kim, D.-J. (2026) 'Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift', Scientific Reports, 16, 10016. https://doi.org/10.1038/s41598-026-40637-w

Shokrpour, S., MoghadamFarid, A., Bazzaz Abkenar, S., Haghi Kashani, M., Akbari, M. and Sarvizadeh, M. (2025) 'Machine learning for Parkinson's disease: A comprehensive review of datasets, algorithms, and challenges', npj Parkinson's Disease, 11, 187. https://doi.org/10.1038/s41531-025-01025-9

Sun, X., Nakashima, M., Nguyen, C., Chen, P.-H., Tang, W.H.W., Kwon, D. and Chen, D. (2025) 'FairICP: Identifying biases and increasing transparency at the point of care in post-implementation clinical decision support using inductive conformal prediction', Journal of the American Medical Informatics Association, 32(8), pp. 1299–1309. https://doi.org/10.1093/jamia/ocaf095

Tumpa, Z.N., Zawad, M.R.S., Sollis, L., Parab, S., Chen, I.Y. and Washington, P. (2025) 'Quantifying device type and handedness biases in a remote Parkinson's disease AI-powered assessment', npj Digital Medicine, 8, 550. https://doi.org/10.1038/s41746-025-01934-2

Varghese, J., van Alen, C.M., Fujarski, M., Schlake, G.S., Sucker, J., Warnecke, T. and Thomas, C. (2021) 'Sensor Validation and Diagnostic Potential of Smartwatches in Movement Disorders', Sensors, 21(9), 3139. https://doi.org/10.3390/s21093139

Varghese, J., Brenner, A., Fujarski, M., van Alen, C.M., Plagwitz, L. and Warnecke, T. (2024a) 'Machine Learning in the Parkinson's disease smartwatch (PADS) dataset', npj Parkinson's Disease, 10, 9. https://doi.org/10.1038/s41531-023-00625-7

Varghese, J., Brenner, A., Plagwitz, L., van Alen, C., Fujarski, M. and Warnecke, T. (2024b) 'PADS - Parkinsons Disease Smartwatch dataset' (version 1.0.0) [Dataset]. PhysioNet. https://doi.org/10.13026/m0w9-zx22t