Original Paper

CheckPD App Smartphone Motor Time-Series Modeling for Parkinson Disease Case-Control Discrimination: Internal Validation Study

Zar Kyi Win[1], BSc; Teerawat Laosombut[1], MSc; Phattharat Songthung[2,3], MCA; Natthaphat Leabthong[1], BSc; Nurhassneeda Jeh-voh[1], BSc; Varis Ratanasirisawad[1], MD; Piyapon Rattanajun[1], BBA; Chanawat Anan[1], MSc; Saisamorn Phumphid[1], MNS; Kornake Horaruengdecha[1], MSc; Decho Surangsrirat[4], PhD; Jirada Sringean[1], MD, PhD; Peerapon Vateekul[2], PhD; Roongroj Bhidayasiri[1,5], MD, FRCP

[1]Chulalongkorn Centre of Excellence for Parkinson's Disease & Related Disorders, Department of Medicine, Faculty of Medicine, Chulalongkorn University and King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand

[2]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

[3]National Electronics and Computer Technology Centre, National Science and Technology Development Agency, Pathum Thani, Thailand

[4]Digital Healthcare Platform Innovation Group, National Science and Technology Development Agency, Pathum Thani, Thailand

[5]The Academy of Science, The Royal Society of Thailand, Bangkok, Thailand

Corresponding Author:

Roongroj Bhidayasiri, MD, FRCP
Chulalongkorn Centre of Excellence for Parkinson's Disease & Related Disorders, Department of Medicine, Faculty of Medicine
Chulalongkorn University and King Chulalongkorn Memorial Hospital, Thai Red Cross Society
1873 Rama 4 Road
Bangkok 10330
Thailand
Phone: 66 22564000
Email: rbh@chulapd.org

Abstract

Background: Smartphone inertial measurement unit (IMU) and touchscreen tasks can capture Parkinson disease (PD)-related motor information, but deep-learning analyses require participant-level leakage control and cautious clinical interpretation.

Objective: This study aimed to compare bidirectional long short-term memory (BiLSTM) and one-dimensional convolutional BiLSTM (Conv1D-BiLSTM) models applied to CheckPD smartphone motor-task time series and evaluate whether stacked ensembles improve internal case-control discrimination.

Methods: We performed a retrospective case-control modeling analysis of 679 participants: 344 with neurologist-confirmed PD and 335 neurologist-confirmed healthy controls. Five tasks were analyzed: Rest Tremor, Postural Tremor, Balance, GaitWalk, and PinchToSize. Participant-level splitting preceded windowing or touchscreen-session construction. Models were developed using stratified 5-fold cross-validation, with out-of-fold predictions used for ensemble training. Five-fold–trained task-specific models were evaluated separately on the same internal held-out complete-case cohort of 90 participants (45 PD and 45 healthy controls), and individual-task performance metrics were averaged across models. Ensembles were evaluated at the participant level.

Results: Conv1D-BiLSTM achieved a higher receiver operating characteristic area under the curve (ROC-AUC) than BiLSTM for Rest Tremor (0.8854 vs 0.8087), Postural Tremor (0.8750 vs 0.6813), Balance (0.8627 vs 0.6675), and GaitWalk (0.8700 vs 0.7090). PinchToSize showed a slightly higher ROC-AUC with BiLSTM (0.8410 vs 0.8340). The Conv1D-BiLSTM Main Symptoms ensemble (Rest Tremor, GaitWalk, and PinchToSize) achieved the highest observed ROC-AUC (0.9131, 95% CI 0.8439-0.9695) and precision-recall area under the curve (0.9354, 95% CI 0.8901-0.9822). The Conv1D-BiLSTM Easy ensemble (Rest Tremor, Postural Tremor, and Balance) achieved the highest macro-F_1-score (0.8534) and accuracy (77/90, 86%) at the evaluation classification rule.

Conclusions: Deep sequence modeling of CheckPD smartphone motor-task time series showed promising internal case-control discrimination. The findings support further evaluation of multidomain smartphone motor assessment as a potential adjunct

to prespecialist clinical assessment. They do not establish diagnostic validity, calibrated clinical risk prediction, or readiness for clinical use. Prospective external validation, differential diagnosis testing, calibration, and workflow evaluation are required.

JMIR Neurotech 2026;5:e109459; doi: 10.2196/109459

Keywords: Parkinson's disease; CheckPD App; smartphone; inertial measurement unit; touchscreen; digital motor measures; deep learning; time-series modeling; Conv1D-BiLSTM; internal validation; one-dimensional convolutional bidirectional long short-term memory

Introduction

Background

Parkinson disease (PD) is a progressive neurodegenerative disorder with motor manifestations that include bradykinesia, tremor, rigidity, gait disturbance, and balance impairment [1]. It contributes substantially to the global burden of neurological disease, and its rising prevalence has been framed as a "Parkinson pandemic" requiring scalable public health responses [2,3]. Despite advances in digital assessment and biomarker research, PD diagnosis remains primarily clinical, relying on specialist neurological examination guided by criteria such as the Movement Disorder Society Clinical Diagnostic Criteria for Parkinson disease [4]. However, clinicopathological evidence indicates that diagnostic accuracy varies according to clinical setting, examiner expertise, and duration of follow-up [5]. In community and prereferral settings, subtle or intermittent motor changes may be overlooked or misattributed to aging, musculoskeletal problems, essential tremor, or other common conditions [6,7]. This creates a measurement gap between the need for scalable early motor assessment and the limited availability of specialist neurological expertise.

Smartphone-based sensing offers a scalable approach for collecting objective motor data beyond specialist clinical settings. Embedded inertial measurement units (IMUs) can quantify acceleration and angular velocity during tremor, standing balance, and gait tasks, and prior wearable- and smartphone-based studies support the relevance of these motion signals for PD classification and early motor assessment [8,9]. Recent reviews of machine learning and deep learning studies in PD further highlight the range of available datasets, algorithms, sensing modalities, and generalization challenges in digital motor assessment research [10-13]. Annotated smartwatch datasets also illustrate the importance of clinically structured and well-labeled sensor data when developing machine learning models for movement disorders [14]. In addition to inertial sensing, smartphone touchscreen interactions provide a complementary source of fine-motor information. Touchscreen typing, touch-gesture, smartphone finger-drawing, and handwriting/dexterity studies suggest that movement timing, scaling accuracy, drawing trajectories, and fine-motor traces can contain PD-relevant motor information, including signals related to bradykinesia, dexterity impairment, and early motor dysfunction [15-18].

From a clinical perspective, motor impairment in PD is multidimensional and cannot be adequately captured by a single sign or task [19]. Rest tremor may be intermittent or absent, postural tremor may overlap with other tremor syndromes, gait impairment may reflect bradykinesia, reduced arm swing, freezing tendency, balance dysfunction, or nonneurological comorbidity, and fine-motor impairment may be influenced by bradykinesia, rigidity, dystonia, arthritis, or task familiarity [20]. Smartphone-based motor assessment therefore requires interpretation across complementary motor domains rather than reliance on any single digital task. This multidomain structure also creates an analytical challenge: clinically relevant information may be contained not only in summary measures, but also in the temporal organization of tremor, balance, gait, and touchscreen interaction signals. Deep sequence models are well suited to this setting because they can analyze these temporal patterns while preserving more of the original motor time-series structure than handcrafted features alone.

These clinical and signal-level complexities require careful machine learning design to avoid overinterpretation. Recurrent neural network architectures, including long short-term memory (LSTM)– and gated recurrent unit (GRU)–based models, have previously been explored for PD classification using sequential motor data [21], supporting the relevance of sequence modeling in this field. However, applying deep sequence models to smartphone-derived PD motor measures raises several important considerations. First, PD motor manifestations are multidomain and heterogeneous, so single-task models may miss complementary information from tremor, standing balance, gait, and fine-dexterity tasks. Second, windowed sensor pipelines are vulnerable to data leakage if segments from the same participant appear in both training and evaluation partitions. In repeated-measures digital health datasets, participant-aware validation is essential because windows, trials, or sessions from the same participant are correlated and can leak participant-specific information across folds if split at the sample level [22]. Leakage can inflate biomedical prediction performance and lead to overoptimistic conclusions [23]. Third, deep learning models applied to correlated sensor streams require cautious interpretability analysis because feature-attribution methods describe model reliance rather than causal physiology. Shapley additive explanations (SHAP) provide a framework for prediction attribution [24], but dependence among features can affect Shapley-value approximations and interpretation [25].

Objectives

Therefore, the objective of this study was to evaluate whether recurrent and convolutional-recurrent deep sequence models can learn PD-related motor information directly from brief CheckPD App smartphone IMU and touchscreen

motor-task time series. Specifically, we compared bidirectional long short-term memory (BiLSTM) and one-dimensional convolutional BiLSTM (Conv1D-BiLSTM) architectures across individual tremor, balance, gait, and PinchToSize tasks; evaluated multimodal neural stacking across complementary motor domains; and examined post hoc model-reliance patterns. The inclusion of PinchToSize allowed the analysis to extend beyond inertial sensing to touchscreen-based fine-motor dynamics, including scaling accuracy, timing, and adjustment behavior. This study should be interpreted as a leakage-aware internal validation of deep sequence modeling for smartphone-derived motor time series, rather than as evidence of calibrated clinical risk prediction, operating threshold development, differential diagnosis validation, or workflow implementation.

Methods

Study Design and Participants

This was a retrospective case-control modeling analysis using a curated, labeled subset of smartphone-derived motor assessment data collected through the CheckPD App, a national digital PD screening program in Thailand. The broader CheckPD program has been implemented in collaboration with the Thai Red Cross Society and the National Health Security Office, and has been delivered through multiple real-world pathways, including the PD clinic at Chulalongkorn Center of Excellence for Parkinson Disease and Related Disorders (ChulaPD), field screening campaigns, online training–activated screening sessions, village health volunteer (VHV)-assisted community screening, and self-initiated app downloads [26]. Data for the current analysis were collected between January 2024 and April 2026 and exported from the CheckPD analytics infrastructure after deidentification. Details regarding program registration, ethics approval, informed consent, data protection, and data availability are provided in the Protocol and Registration, Ethical Considerations, and Data Availability sections.

The current analytic cohort comprised 679 labeled participants with sufficient task data for the selected analyses, including 344 participants with neurologist-confirmed PD and 335 neurologist-confirmed healthy controls. PD diagnoses were established by certified neurologists according to the Movement Disorder Society Clinical Diagnostic Criteria for PD [4]. Healthy controls were selected from the available neurologist-confirmed control pool to approximate the PD group by age and gender before model development, thereby reducing, but not eliminating, demographic confounding.

PD participants were recruited through the CheckPD program, including participants from the PD clinic at ChulaPD and participants identified or assessed through CheckPD screening workflows. When available, recruitment source was categorized as clinic-based, field/community screening, online training–activated screening, VHV-assisted screening, or self-initiated app use. Among the 344 participants with PD, 120 were previously diagnosed clinic-based PD participants and 224 were identified or assessed through CheckPD field/community workflows. Because recruitment pathway may be associated with disease severity, task supervision, and task completion, recruitment-source heterogeneity was considered a limitation of this retrospective analysis. Additionally, separate counts for online training–activated screening, VHV-assisted screening, and self-initiated app use were not consistently available in the deidentified analytic dataset and therefore could not be reported. Healthy controls were independent field/community participants or relatives recruited through the CheckPD screening workflow. Participants were excluded from the healthy control pool if they had a diagnosed movement disorder or if neurologist assessment identified PD or clinically evident parkinsonism. The comparison group should therefore be interpreted as neurologist-confirmed healthy controls rather than as an unverified non-PD group. However, controls were not selected to represent a structured symptomatic differential diagnosis cohort, and the dataset did not include separate confirmed subgroups for essential tremor, atypical parkinsonism, drug-induced parkinsonism, neuropathy, frailty, arthritis, or musculoskeletal gait disorders.

Eligibility Criteria, Exclusions, and Clinical Characterization

Eligibility for the source CheckPD program and inclusion in the curated modeling dataset required all of the following: age 40 years or older at screening, digital informed consent within the CheckPD application, an available diagnostic label for the curated modeling dataset, and evaluable data for at least one of the 5 analyzed motor tasks. For task-specific analyses, records were excluded if consent was incomplete or withdrawn, if technical failure prevented capture of evaluable data, or if the corresponding task recording did not pass quality control.

Clinical characterization of the PD cohort was based on neurologist-confirmed diagnosis, disease duration, Movement Disorder Society–sponsored version of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III motor score, and cohort-level motor phenotype. Mean disease duration was 7.25 (SD 7.05) years, and mean MDS-UPDRS Part III motor score was 11.84 (SD 12.61) [25]. Among participants with available motor phenotype data, the reported presentations included bradykinetic-rigid/gait-dominant and tremor-dominant presentations. Motor phenotype and MDS-UPDRS Part III data were available for only 21.8% of participants with PD. The available information described mixed symptoms and did not allow calculation of percentages for separate motor subtypes. Exact dopaminergic medication ON/OFF state at the time of each smartphone task was not consistently recorded, and Hoehn and Yahr (HY) stage was not available. Therefore, this analysis could not perform medication state–specific or formal stage–specific subgroup modeling. Demographic and clinical characteristics are summarized in Table 1.

Table 1. Demographic and clinical characteristics of the analytic cohort[a].

Characteristics and category	PD[b] group (n=344)	Healthy controls (n=335)
Gender, n		
Women	157	162
Men	176	139
Other	11	34
Age (years), mean (SD)	63.74 (12.45)	61.75 (13.58)
Diagnostic status confirmation	Neurologist-confirmed PD	Neurologist-confirmed healthy control
Disease duration (years), mean (SD)	7.25 (7.05)	N/A[c]
MDS-UPDRS III[d], mean (SD)	11.84 (12.61)	N/A
Dominant motor and predominant phenotype	Bradykinetic-rigid/gait-dominant, tremor-dominant	N/A

[a]Only 21.8% of PD participants have predominant phenotype and MDS-UPDRS III information.
[b]PD: Parkinson disease.
[c]Clinical variables listed as N/A for healthy controls were available only in the PD cohort or were not applicable to healthy control participants.
[d]MDS-UPDRS: Movement Disorder Society-sponsored version of the Unified Parkinson disease Rating Scale.

Task availability and missing modality patterns were summarized descriptively. We did not formally model whether missingness differed by PD status, disease severity, task usability, device factors, or other clinical and technical variables. No inverse probability weighting, missingness prediction model, or formal usability-adjusted analysis was performed. Therefore, task availability and complete-case ensemble results may have been affected by nonrandom task completion patterns.

Because this study evaluates smartphone-derived sensor and touchscreen motor measures and develops machine learning prediction models, reporting was guided by relevant EVIDENCE (Evaluating Connected Sensor Technologies) items for connected sensor technologies and TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI) items for clinical prediction models using AI or machine learning methods [27,28]. Completed EVIDENCE and TRIPOD+AI checklist tables are provided in Checklist 1.

Protocol and Registration

The broader CheckPD/Parkinson disease risk survey program was registered in the Thai Clinical Trials Registry (TCTR20231025005) and on ClinicalTrials.gov as an observational patient registry (NCT06609681). This study is a retrospective model development and internal validation analysis of deidentified CheckPD App data conducted under the ethics-approved broader CheckPD program. It was not registered as a separate clinical trial, and no separate publicly registered prospective analysis protocol was prepared for this retrospective modeling analysis.

Ethical Considerations

The broader CheckPD program and the current retrospective deidentified modeling analysis were conducted under ethics approval from the Institutional Review Board (IRB) of the Faculty of Medicine, Chulalongkorn University (IRB number 0804/65), and the Central Research Ethics Committee (CREC) under the approved protocol entitled “Flipping the Paradigm of Parkinson’s Disease: A Model of National ‘Eat, Move, Sleep’ Digital Interventions to Prevent or Slow the Rise of Non-Communicable Diseases in Thailand” (CREC number CREC023/68BR-MED06; certificate number COA-CREC105/2025). Digital informed consent was obtained from all participants through the CheckPD App before screening or assessment. This analysis used deidentified data exported from the CheckPD analytics infrastructure. Data handling was conducted in accordance with institutional governance requirements and Thailand’s Personal Data Protection Act. No compensation was provided to participants. No individual participants or users are identifiable in any images included in the manuscript or multimedia appendices.

Smartphone Assessment Protocol

Data were collected through the CheckPD mobile app using either participants’ personal smartphones or program-provided smartphones as part of the broader national screening workflow. However, device ownership status was not consistently retained in the deidentified analytic dataset; therefore, the proportions of personal versus program-provided devices could not be determined. Operating system, device manufacturer, and specific smartphone model were also not systematically available at the participant or recording level. The full CheckPD workflow includes app access, user registration, digital informed consent, guided digital motor and voice assessments, structured symptom and risk-factor questionnaires, and automated processing of screening results [29]. Within this workflow, the app provides in-app instructions for performance tasks, including tremor, balance, gait, voice, and finger-dexterity or touchscreen-based assessments.

The current modeling analysis focused on 5 smartphone motor-task datasets relevant to digital motor assessment: Rest Tremor, Postural Tremor, Balance, GaitWalk, and PinchToSize. These tasks were selected because they capture complementary motor domains relevant to PD-related motor impairment, including tremor, standing balance, walking-related movement, and touchscreen fine-motor interaction. Participants in the broader implementation workflow could skip individual tasks or discontinue specific components while continuing the assessment; therefore, task-level

missingness was expected in this real-world app-derived dataset [29]. Table S1 in Multimedia Appendix 1 summarizes the acquisition method for each analyzed task.

Signal Processing and Windowing

Overview

Structured CheckPD application data were stored as JSON records within the platform infrastructure, with curated diagnostic labels managed in the analytic data warehouse as described in the national implementation program [29]. The current analysis used deidentified task-level exports from this infrastructure.

Mobile Data Collection Libraries and IMU Data Schema

Smartphone IMU data were collected within the CheckPD mobile app using the Flutter sensors_plus package, version 7.0.0, on Android and iOS devices [30]. This package provides a common app-level Dart interface for accelerometer and gyroscope streams across platforms. Although the underlying native sensor hardware and operating system implementations may differ across devices, all IMU recordings were stored using a common 6-channel schema consisting of triaxial accelerometer and triaxial gyroscope signals, consistent with prior smartphone inertial-sensing work [31]. Smartphone accelerometer studies also support smartphone-based inertial sensing as a practical source of movement time-series data [32], and public inertial-walking datasets in PD further demonstrate the relevance of structured gait sensor recordings for reproducible model development [33]. Accelerometer values were stored in m/s^2, consistent with the sensors_plus accelerometer event definition [34], and gyroscope values were stored as angular velocity in rad/s, consistent with the sensors_plus gyroscope event definition [35].

Because smartphone sensor streams may show timing jitter and variable sampling intervals, time stamps were retained for each sample. The effective sampling rate of each recording was estimated from the median intersample interval, and signals were aligned to a fixed 40-Hz analysis grid using linear interpolation when required. This time stamp–based approach was used because the requested sampling interval may not be achieved exactly on all devices, particularly given variation across Android versions, hardware models, and vendor operating system customizations described in the sensors_plus documentation [30]. Prior smartphone inertial sensor PD work further illustrates the importance of treating smartphone-based walking recordings as device-acquired time series rather than fixed laboratory sensor streams [36,37].

IMU Preprocessing

The IMU preprocessing pipeline included low-pass filtering, accelerometer gravity component removal, orientation consistency checking, and channel-wise normalization. First, a fourth-order Butterworth low-pass filter with a maximum cutoff frequency of 20 Hz was applied to reduce high-frequency sensor noise while preserving movement-related signal components. The cutoff frequency was constrained to remain below the Nyquist frequency when required. Second, slow gravity–related components were removed only from accelerometer channels using a fourth-order Butterworth high-pass filter with a cutoff frequency of 0.5 Hz. This step was applied because raw smartphone accelerometer signals include both body movement and gravitational acceleration, a common consideration in smartphone human-activity recognition [31]. Gyroscope channels measure angular velocity and were therefore not subjected to gravity removal.

Third, accepted IMU recordings were acquired according to task-specific CheckPD instructions that specified a standardized phone orientation and device axis convention. Before analysis, recordings were checked for orientation consistency according to the expected task-specific device-frame axis pattern. Recordings with inconsistent or nonstandard orientation were corrected when a deterministic axis transformation was possible; otherwise, they were excluded during quality control. Therefore, IMU channels were analyzed in a standardized device-centric coordinate frame. Because axes were not transformed into anatomical or global coordinates, axis-level findings should be interpreted as task-specific device-frame signal patterns rather than anatomical axis biomarkers. Fourth, each channel was standardized using z-score normalization to reduce interdevice and interchannel scale differences and improve model training stability, consistent with common inertial sensor activity recognition preprocessing practice [38]. Smartphone activity recognition surveys also emphasize variability across devices and sensing contexts [39].

Touchscreen Data Processing

Touchscreen data for the PinchToSize task were collected using the core Flutter pointer and gesture APIs rather than the IMU sensor library. Flutter pointer events provide touch positions in logical pixels and time stamps relative to an arbitrary event-dispatch timeline [40-42]. Therefore, IMU-specific filtering and gravity removal were not applied to PinchToSize data. Instead, 7 touchscreen-derived time-series features were standardized before model input: timestamp, diameter, dist_to_target, dist_to_target_pct, expansion_rate, diameter_diff, and time_diff. Here, “timestamp” denotes the event time of the touch sample, whereas “time_diff” denotes the elapsed time between consecutive samples. “diameter” represents the current pinch diameter, “diameter_diff” represents the difference between the current pinch diameter and the target diameter, and “dist_to_target” and “dist_to_target_pct” represent the absolute and percentage distance from the target size.

PinchToSize was analyzed as a touchscreen time-series task rather than as a direct clinical bradykinesia assessment. These variables were selected to quantify touchscreen scaling accuracy, interevent timing, and adjustment rate. The raw time stamp feature was interpreted cautiously, because it may also encode task duration, event ordering, platform behavior, or session-level acquisition effects. Prior smartphone touch-gesture work has shown that common touchscreen

interactions can capture motor impairment patterns in early PD [15]. Quantitative finger-dexterity time-series analysis has also shown sensitivity to impaired finger dexterity in PD [17]. Because clinician-rated hand motor scores were not available in the current dataset, PinchToSize variables were interpreted as fine-motor and touchscreen interaction features rather than direct measures of bradykinesia [43,44].

Device Effects and Quality Control

Although using Flutter-based app-level sensor and touch-event interfaces reduced platform-level API inconsistency between Android and iOS, residual device model effects, sensor calibration differences, sampling irregularity, and phone orientation effects were not fully eliminated. These effects were mitigated through time stamp–based alignment, common unit handling, orientation consistency checking, channel-wise normalization, and participant-level evaluation, but they remain limitations requiring external validation across device models. No dedicated stratified analysis by operating system, device manufacturer, or sensor model was performed in this analysis. All preprocessing parameters were estimated within training folds only to prevent information leakage.

Quality control procedures excluded trials with sensor dropout or duration below 90% of the nominal protocol length for fixed-duration 20-second IMU tasks, corresponding to less than 18 seconds of usable signal for Rest Tremor, Postural Tremor, or Balance. The nominal duration reported in the acquisition table and the quality control threshold refer to the same raw task protocol; no task-specific exception was applied to Balance. For the PinchToSize task, each recording comprised 5 sessions targeting circle diameters of 200, 250, 300, 200, and 270 pixels.

Model Development and Ensembles

Architecture Scope and Rationale

BiLSTM and Conv1D-BiLSTM were selected because the CheckPD signals may contain information at both local and longer temporal scales. BiLSTM was used as a recurrent-only baseline to model temporal dependencies across IMU windows and touchscreen-event sequences, whereas Conv1D-BiLSTM was used to test whether a convolutional front end could additionally extract short local patterns, such as tremor oscillations, postural-sway fluctuations, gait-related rhythms, or touchscreen adjustment dynamics, before recurrent aggregation. This controlled comparison therefore assessed the incremental value of convolutional feature extraction while retaining a comparable recurrent component. The novelty of this study does not lie in proposing a new neural network architecture, but in applying and comparing these models across 5 complementary CheckPD motor tasks using participant-level splitting before windowing, fold-contained preprocessing, out-of-fold (OOF) stacked ensemble fusion, and evaluation on a common internal held-out cohort. This combination distinguishes the study from prior work that evaluated related architectures or individual digital motor modalities without the same CheckPD multidomain and leakage-aware modeling framework.

BiLSTM Architecture

BiLSTM was used as the recurrent sequence classification model for task-specific smartphone motor time series. For IMU tasks, the input consisted of 2.5-second windows sampled on the 40-Hz analysis grid with 6 channels: ax, ay, az, gx, gy, and gz. For PinchToSize, the model used the 7 standardized touchscreen time-series features described above.

Input sequences were processed by stacked bidirectional LSTM layers. The number of recurrent layers, hidden size, dropout rate, and optional attention module were selected through random-search hyperparameter optimization. The resulting temporal representation was passed to a multilayer classifier to generate class probabilities for PD and healthy control labels. A schematic of the BiLSTM architecture is provided in Figure S1 in Multimedia Appendix 1.

Conv1D-BiLSTM Architecture

In this study, Conv1D-BiLSTM denotes a one-dimensional convolutional temporal feature extraction front end followed by bidirectional LSTM temporal modeling. It is not a canonical ConvLSTM recurrent cell, in which convolutional operations are incorporated into recurrent gates for spatiotemporal sequence modeling [45]. The term Conv1D-BiLSTM is used here to describe the implemented pipeline architecture: local temporal motif extraction by Conv1D layers followed by recurrent temporal aggregation using stacked BiLSTM layers.

The Conv1D-BiLSTM model accepted the same task-specific time-series inputs as the BiLSTM model, with 6 channels for IMU tasks and 7 standardized touchscreen features for PinchToSize. A stack of Conv1D layers, followed by group normalization, rectified linear unit (ReLU) activation, and dropout, was used to extract local temporal features. These features were then processed by stacked BiLSTM layers, with an optional multihead self-attention mechanism applied according to the selected hyperparameters. Finally, the learned representation was aggregated and passed to fully connected layers to produce class probabilities for the PD and healthy control labels. A schematic of the Conv1D-BiLSTM architecture is provided in Figure S2 in Multimedia Appendix 1.

Ensemble Modeling

Inputs to the Ensemble

To evaluate whether different smartphone motor-task domains provided complementary information, we constructed stacked ensemble models using OOF-predicted probabilities from the task-specific base models. This followed the stacked generalization principle, in which a second-stage learner is trained on predictions generated by first-stage models [46,47]. The use of OOF predictions ensured that the meta-learner was trained using predictions from participants who were not seen by the corresponding base model during that fold.

Each task-specific base classifier produced 2 softmax probabilities: one for the healthy control class and one for the PD class. Let "s" index participants, "r" index recordings or files within participant "s, j" index windows or touchscreen sessions within recording "r", "m" index modality, and "c ∈ {healthy control, PD}" index class. For each task-specific base model, the softmax probability for window or session "j" was denoted as $p_{srjm}(c)$.

Recording-level probabilities were first obtained by averaging the available window-level or session-level probabilities within each file for each class. Here, n_{srm} denotes the number of windows or sessions available for recording "r" from participant "s" in modality "m." For participants with more than one accepted recording within the same modality, recording-level probabilities were then averaged across accepted recordings to obtain a participant-level modality probability. Here, N_{sm} denotes the number of accepted recordings for participant "s" in modality "m."

Ensemble Configurations

The ensemble analysis was designed to compare task sets with different levels of participant burden and motor domain coverage. The practical difficulty of each task was classified based on an internal survey of CheckPD data collection staff and members of the clinical and research team at the ChulaPD. Rest Tremor, Postural Tremor, and Balance were classified as easy, GaitWalk as moderate, and PinchToSize as difficult, based on the practical effort required to explain, administer, and complete each task. These categories reflected task burden and feasibility rather than model performance. Four configurations were defined:

- Easy: Rest Tremor, Postural Tremor, and Balance.
- Easy + Moderate: Rest Tremor, Postural Tremor, Balance, and GaitWalk.
- All-task: Rest Tremor, Postural Tremor, Balance, GaitWalk, and PinchToSize.
- Main Symptoms: Rest Tremor, GaitWalk, and PinchToSize, representing tremor, gait, and fine-motor/touchscreen domains.

For each configuration, participant-level OOF probabilities from the included task-specific classifiers were concatenated and used as inputs to the stacked meta learner. This design allowed comparison of low-burden, expanded, all-task, and symptom-focused fusion strategies.

Modality Intersection and Feature Construction

For each ensemble configuration, only participants with complete OOF probability data for every included modality were retained for meta-model training. In the Main Symptoms configuration, for example, the complete-case intersection required Rest Tremor, GaitWalk, and PinchToSize probabilities from the same participant. These modality-intersection counts refer to the development-set OOF tables used for ensemble meta-model training and should not be interpreted as held-out test denominators. They are summarized in Table S2 in Multimedia Appendix 1.

For each retained complete-case participant, the ensemble feature vector was constructed by concatenating the healthy control and PD probabilities for each included modality. Because the base models used 2-class softmax outputs, the ensemble input dimension was d_in=2×M, where M denotes the number of modalities included in the ensemble configuration. For example, an ensemble containing 3 modalities had 6 input features.

Because ensemble training required complete modality overlap within each configuration, all ensemble analyses were complete-case analyses. Participants who were able to complete all required tasks may differ systematically from those with missing tasks in motor severity, dexterity, fatigue, cognition, or smartphone usability. Therefore, ensemble performance may be biased toward participants who were better able to complete the required tasks. All 90 held-out participants had complete data for all 5 analyzed task types; therefore, complete-case modality intersection did not reduce the held-out ensemble evaluation denominator, and every ensemble configuration was evaluated on the same 45 PD and 45 healthy control held-out participants.

Meta-Model Architecture and Training

A small feed-forward neural network was trained as the ensemble meta learner. The network consisted of 2 hidden layers with batch normalization, ReLU activation, and dropout, followed by a sigmoid output layer for binary classification. The implemented hidden-layer widths were 128 and 64, with dropout probability 0.1. The meta learner produced a single output probability for the PD class.

The meta-model was optimized using binary cross-entropy loss with the Adam optimizer, a learning rate of 10^{-4}, and weight decay of 10^{-2}. Training was performed for up to 200 epochs with early stopping using a patience of 20 epochs. Within each OOF-derived ensemble training table, a stratified 90/10 split was used for validation and early stopping.

Training, Validation, and Statistical Analysis

The experimental pipeline began by grouping all raw task recordings by participant identifier. All training, validation, and test partitions were created at the participant level before preprocessing parameter estimation, normalization, windowing or session construction, augmentation, and model fitting. This participant-level partitioning was used to prevent within-participant leakage, ensuring that windows, touchscreen sessions, or repeated recordings from the same participant did not appear in more than one data partition.

Within each training fold, preprocessing parameters, normalization statistics, windowing or session construction, data augmentation, and model fitting were performed using training participants only. An internal validation split within the training fold was used for early stopping and random-search hyperparameter selection. Hyperparameters were selected through 25 iterations of random search over predefined ranges, using validation receiver operating characteristic area under the curve (ROC-AUC) as

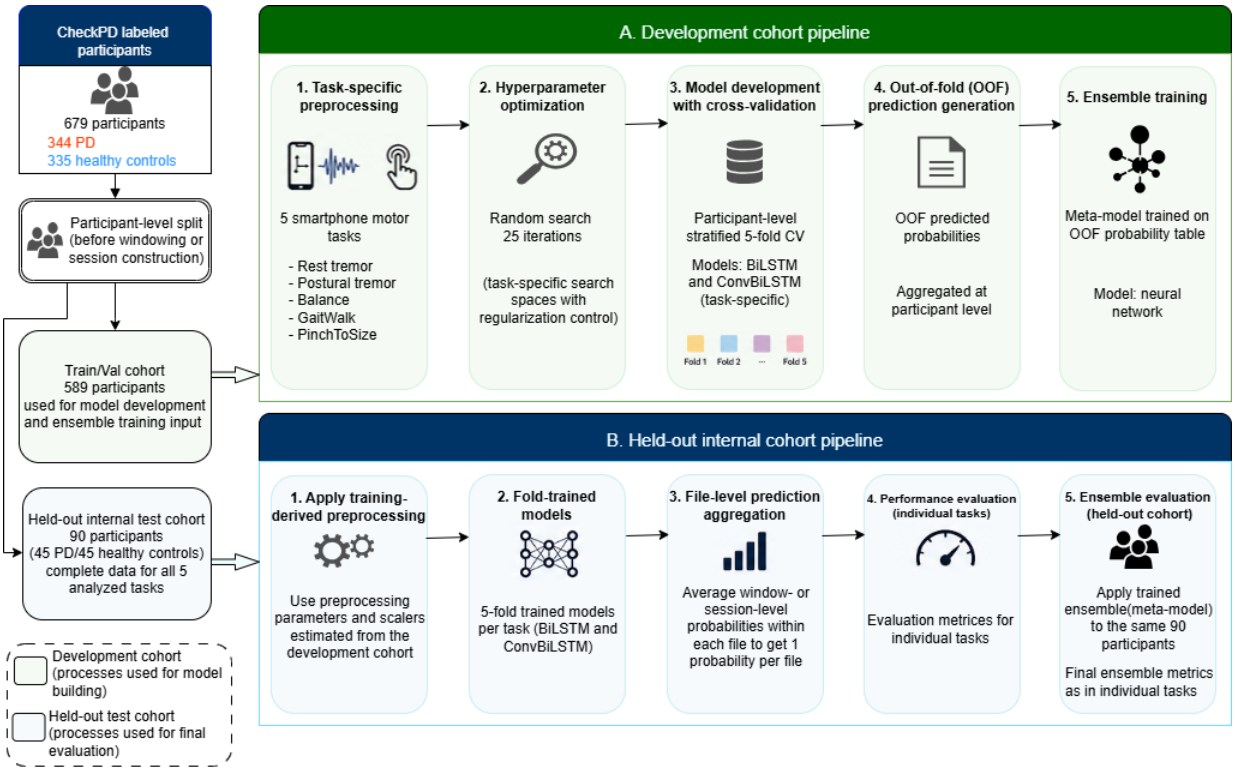
the optimization objective. The full hyperparameter search ranges are provided in Table S3 in Multimedia Appendix 1, final selected hyperparameter values for the BiLSTM and Conv1D-BiLSTM task-specific models are reported in Table S4 in Multimedia Appendix 1, and development-set validation performance summaries are provided in Table S5 in Multimedia Appendix 1.

Class-balancing augmentation was applied only within the development-training partition after participant-level splitting and window or session construction. For IMU tasks, augmentation included Gaussian noise, amplitude scaling, small 3D rotation, temporal segment permutation, and time warping. For PinchToSize, augmentation used jittering, scaling, and small temporal and feature perturbations of the touchscreen time-series features. Augmented samples were used only for model fitting and were not used for validation, OOF prediction generation, SHAP explanation, or held-out internal test evaluation.

After model selection, each of the 5-fold–trained task-specific models was evaluated on the same randomly selected internal held-out complete-case cohort of 90 unique participants, comprising 45 participants with PD and 45 healthy controls with complete data for all 5 analyzed task types. Task-specific metrics were calculated at the file level after within-file aggregation of window-level or session-level predictions. For PinchToSize, the file-level denominator exceeded the participant denominator because some held-out participants contributed more than one accepted file.

For multimodal fusion, OOF base-model probabilities were first aggregated to participant-level modality summaries. Participants were then intersected across the modalities required for each ensemble configuration, and the resulting modality-specific probabilities were concatenated as inputs to the ensemble meta learner. Ensemble models were evaluated on the same 90-participant held-out cohort used for the individual task models. Figure 1 summarizes the participant-level modeling pipeline, including preprocessing, random search hyperparameter optimization, task-specific model training, OOF prediction generation, ensemble training, and held-out internal evaluation.

Figure 1. Overall experimental pipeline showing participant-level splitting, preprocessing, random-search hyperparameter optimization, task-specific model training, out-of-fold prediction generation for ensemble training, and held-out internal test evaluation. BiLSTM: bidirectional long short-term memory; Conv1D-BiLSTM: one-dimensional convolutional bidirectional long short-term memory; CV: cross-validation; OOF: out-of-fold; PD: Parkinson disease.



Validation and Performance Analysis

Model development used stratified 5-fold cross-validation at the participant level to preserve class balance and prevent within-participant leakage. In each fold, approximately 80% of participants were used for model development, including an internal validation split for early stopping and hyperparameter selection, and approximately 20% were held out for OOF prediction generation. All preprocessing parameters, normalization statistics, augmentation procedures, and model weights were estimated or applied using the development training partition only. Validation data, OOF prediction generation, SHAP explanation samples, and held-out internal test data were not augmented.

OOF predictions from the fold-specific development-set evaluation partitions were used to construct ensemble training inputs following the stacked generalization principle [46,47].

Participant-level splitting was applied throughout because repeated windows, touchscreen sessions, or recordings from the same participant are correlated and can introduce subject-level data leakage if they appear in both training and evaluation partitions. Such leakage can inflate biomedical prediction performance and lead to overoptimistic conclusions [23].

After model selection, each of the 5-fold–trained task-specific models was evaluated separately on the same internal held-out cohort of 90 participants, comprising 45 participants with PD and 45 healthy controls. For the individual task results in Table 2, each performance metric was calculated separately for each model, and the 5 resulting metric values were averaged with equal weights. Prediction probabilities were not averaged across the 5 models before metric calculation. This approach provides a descriptive summary across the retained cross-validation fits rather than the performance of a probability averaging ensemble or a single model refitted on all development data. Table 3 is distinct and reports participant-level performance of the stacked meta learners.

Table 2. File-level fold-averaged held-out point estimates for individual BiLSTM[a] and Conv1D-BiLSTM[b] models[c,d].

Models and metric	Rest tremor	Postural tremor	Balance	GaitWalk	PinchToSize
Held-out participants (PD[e]/healthy control), n/N	45/45	45/45	45/45	45/45	45/45
BiLSTM					
ROC-AUC[f]	0.8087	0.6813	0.6675	0.7090	0.8410[g]
Macro F_1-score	0.7017	0.4914	0.5601	0.6357	0.7226[g]
Accuracy (%)	72.88	62.44	62.88	64.67	72.70[g]
Specificity (%)	82.66	82.22	74.66	82.66	77.33
Sensitivity (%)	63.11	42.66	51.11[g]	46.66	69.71[g]
PR-AUC[h]	0.8536	0.7131	0.7230	0.7127	0.8839
Brier score (↓)[i]	0.1814	0.2276	0.2235	0.2178	0.1903[g]
Conv1D-BiLSTM					
ROC-AUC	0.8854[g]	0.8750[g]	0.8627[g]	0.8700[g]	0.8340
Macro F_1-score	0.8416[g]	0.8057[g]	0.7293[g]	0.8425[g]	0.6980
Accuracy (%)	84.44[g]	81.11[g]	74.44[g]	84.44[g]	69.91
Specificity (%)	97.78[g]	97.77[g]	97.77[g]	95.55[g]	83.56[g]
Sensitivity (%)	71.11[g]	64.44[g]	51.11[g]	73.33[g]	61.14
PR-AUC	0.9222[g]	0.9055[g]	0.8904[g]	0.9048[g]	0.8894[g]
Brier score (↓)	0.1496[g]	0.1448[g]	0.1760[g]	0.1359[g]	0.2063

[a]BiLSTM: bidirectional long short-term memory.
[b]Conv1D-BiLSTM: one-dimensional convolutional bidirectional long short-term memory.
[c]ROC-AUC, PR-AUC, and Brier score use continuous model scores; the remaining metrics use binary predictions according to the evaluation rule defined in Methods. Macro F_1 is the unweighted mean of PD- and healthy-control-class F_1-scores. The PinchToSize file-level denominator exceeded the participant denominator because some participants contributed more than one accepted file. 95% CIs are reported in Table S6 in Multimedia Appendix 1.
[d]Metrics were computed on the balanced held-out cohort of 90 participants after within-file aggregation.
[e]PD: Parkinson disease.
[f]ROC-AUC: receiver operating characteristic area under the curve.
[g]Best-performing result within each task–metric pair.
[h]PR-AUC: precision-recall area under the curve.
[i]↓ indicates that a lower Brier score is better.

Table 3. Held-out point estimates for stacked BiLSTM[a] and Conv1D-BiLSTM[b] ensembles across 4 feature-set configurations[c,d].

Model and metric[e]	Easy[f]	Easy+Moderate[g]	All ensemble	Main symptoms
Held-out participants (PD[h]/healthy control), (n/N)	45/45	45/45	45/45	45/45
BiLSTM ensemble				
ROC-AUC	0.8514	0.8538	0.8884	0.9077
Macro F_1-score	0.7935	0.7935	0.7685	0.7558
Accuracy, n/N (%)[i]	72/90 (80)	72/90 (80)	70/90 (78)	69/90 (77)
Specificity, n/N (%)	44/45 (98)[j]	44/45 (98)[j]	44/45 (98)[j]	44/45 (98)[j]
Sensitivity, n/N (%)	28/45 (62)	28/45 (62)	26/45 (58)	25/45 (56)
PR-AUC[k]	0.8960	0.9002	0.9146	0.9233

Model and metric[e]	Easy[f]	Easy+Moderate[g]	All ensemble	Main symptoms
Brier score (↓)[l]	0.1658	0.1681	0.1554	0.1591
Conv1D-BiLSTM ensemble				
ROC-AUC[m]	0.8968[j]	0.9042[j]	0.9012[j]	0.9131[j]
Macro F_1-score	0.8534[j]	0.8178[j]	0.8298[j]	0.8298[j]
Accuracy, n/N (%)	77/90 (86)[j]	74/90 (82)[j]	75/90 (83)[j]	75/90 (83)[j]
Specificity, n/N (%)	44/45 (98)[j]	44/45 (98)[j]	44/45 (98)[j]	44/45 (98)[j]
Sensitivity, n/N (%)	33/45 (73)[j]	30/45 (67)[j]	31/45 (69)[j]	31/45 (69)[j]
PR-AUC	0.9280[j]	0.9337[j]	0.9241[j]	0.9354[j]
Brier score (↓)	0.1346[j]	0.1471[j]	0.1515[j]	0.1432[j]

[a]BiLSTM: bidirectional long short-term memory.
[b]Conv1D-BiLSTM: one-dimensional convolutional bidirectional long short-term memory.
[c]ROC-AUC, PR-AUC, and Brier score use continuous model scores; the remaining metrics use binary predictions according to the evaluation rule defined in Methods. Macro F_1 is the unweighted mean of PD- and healthy-control-class F_1-scores. Available 95% CIs are reported in Table S8 in Multimedia Appendix 1.
[d]The labels "easy" and "moderate" refer to the practical effort needed to explain, administer, and complete the tasks, not to classification difficulty or model performance.
[e]Metrics were computed on the same complete-case held-out cohort of 90 participants.
[f]The easy configuration includes rest tremor, postural tremor, and balance.
[g]The easy+moderate configuration adds GaitWalk.
[h]PD: Parkinson disease.
[i]Accuracy, specificity, and sensitivity are reported as n/N (%), where n is the number correctly classified and N is the total in the relevant group. Percentages are rounded to whole numbers.
[j]Best-performing result within each task–metric pair.
[k]PR-AUC: precision-recall area under the curve.
[l]↓ indicates that a lower Brier score is better.
[m]ROC-AUC: receiver operating characteristic area under the curve.

Validation ROC-AUC was used for hyperparameter selection, and held-out ROC-AUC was designated as the primary evaluation metric for comparing individual models and ranking ensemble configurations because it is threshold-independent, and the development and held-out cohorts were approximately balanced. Precision-recall area under the curve (PR-AUC), macro-F_1-score, accuracy, sensitivity, specificity, and Brier score were treated as complementary secondary measures. Accordingly, conclusions regarding the best-performing model or ensemble configuration were based primarily on ROC-AUC. ROC-AUC, PR-AUC, and Brier score were calculated from continuous model scores after within-file aggregation of window-level or session-level predictions. Accuracy, macro-F_1-score, sensitivity, and specificity were calculated from binary file-level predictions exported by the evaluation pipeline, without post hoc optimization on held-out labels. Macro-F_1-score was calculated within each fold as the unweighted mean of the PD-class and healthy control–class F_1-scores and then averaged across the 5-fold–trained evaluations. Sensitivity was defined as the proportion of PD observations classified as PD, and specificity was defined as the proportion of healthy control observations classified as healthy control.

No post hoc probability-calibration model, such as Platt scaling, isotonic regression, temperature scaling, or prevalence recalibration, was fitted. Predicted probabilities should therefore be interpreted as model scores for discrimination and internal classification rather than as calibrated estimates of absolute PD probability, consistent with the distinction between discrimination and calibration in clinical prediction modeling [48]. Calibration and decision-curve analysis under representative screening prevalence were deferred to future prospective validation.

In the held-out discrimination estimates, 95% CIs for ROC-AUC and PR-AUC were calculated using 2000 diagnosis-stratified bootstrap resamples. For the individual-task analyses, the same resampled evaluation observations were applied to all 5-fold–trained models; the metric was calculated separately for each model and the 5 metric values were then averaged within each bootstrap resample. The 2.5th and 97.5th percentiles of the resulting distribution defined the 95% CI. For the stacked ensembles, bootstrap resampling was performed at the participant level on the 90-participant held-out cohort. Because the 5 task-specific models were evaluated on the same held-out observations and were trained on overlapping development subsets, the 5 evaluations were not treated as independent test cohorts. The intervals therefore quantify sampling uncertainty in the reported held-out statistic with the trained models fixed and do not represent uncertainty from repeating the complete model development and hyperparameter selection procedure.

To formally compare individual-task BiLSTM and Conv1D-BiLSTM models, metric differences were calculated as Conv1D-BiLSTM minus BiLSTM for each task-fold evaluation. The formal paired comparison was restricted to ROC-AUC as the primary comparison metric and accuracy and macro-F_1-score as secondary comparison metrics. Because the number of paired task-fold evaluations was limited and normality of paired differences was not assumed, 2-sided exact sign tests were used. Ensemble results were

summarized descriptively because they represented multimodal fusion configurations rather than the primary individual-model architecture comparison.

An exploratory demographic sensitivity analysis was performed to assess whether residual age or gender signal could explain model discrimination. Age and gender were summarized descriptively, and a complete-case logistic regression model using age and gender as predictors was evaluated among participants with valid metadata. Performance was summarized using ROC-AUC, PR-AUC, Brier score, accuracy, sensitivity, and specificity. This analysis was interpreted as a demographic-confounding sensitivity analysis and was not used for model selection.

Because the final comparison used fold-averaged descriptive performance on a fixed 90-participant held-out internal cohort, fold-based Wilcoxon signed-rank testing was not used as a primary inferential test. Small-sample cross-validation can yield unstable performance estimates with wide uncertainty, so held-out model comparisons were interpreted descriptively [49].

Model-Reliance Analysis

Model interpretability was assessed using SHAP as a post hoc model inspection analysis [24]. This analysis was separate from the held-out internal evaluation used to report the main case-control discrimination results. SHAP was not used for model selection, hyperparameter optimization, threshold selection, or final performance estimation. DeepSHAP was applied to trained Conv1D-BiLSTM models to examine which input channels contributed most strongly to the model output for the PD class. DeepSHAP approximates SHAP values for deep neural networks by estimating how each input feature contributes to the model output relative to a selected background dataset [24]. For each task-specific model, the background set and explanation set each included up to 3000 training windows or touchscreen sessions from the corresponding model development partition. Held-out test windows or sessions were not used for SHAP computation. For each feature channel and time point, SHAP values represented the contribution of that input to the deviation from the expected model output under the selected training window or training session background distribution. Because inputs were time-series windows or touchscreen-event sequences, SHAP values were aggregated across the temporal dimension to obtain channel-level attributions. Global feature importance was summarized as the mean absolute SHAP value across explained training windows or sessions.

SHAP values were interpreted as model-dependent feature attributions rather than causal evidence of disease mechanisms. Smartphone IMU and touchscreen channels are temporally correlated, and feature dependence can affect SHAP value approximations and the distribution of attribution mass across correlated inputs [25]. Therefore, SHAP findings were treated as internal model-reliance patterns. Task-specific attribution patterns were interpreted cautiously in relation to expected task structure and established motor physiology, including gyroscope contributions during tremor tasks, accelerometer and gyroscope contributions during balance and gait tasks, and touchscreen targeting, scaling, and timing features during PinchToSize. Because IMU axes remained in a device-centric coordinate frame, axis-level SHAP findings should not be interpreted as anatomical axis biomarkers.

Public Dataset Task Domain Benchmarking

Public dataset analyses were conducted to provide task domain benchmarks for the Conv1D-BiLSTM modeling workflow using PADS (Parkinson's Disease Smartwatch) [14] and WearGait-PD (Wearables Dataset for Gait in Parkinson's Disease and Age-Matched Controls) [33]. Only participants with PD and healthy controls were included. The PADS analytic sample comprised 355 participants, including 276 participants with PD and 79 healthy controls. The Relaxed and Stretch tasks were mapped to the CheckPD Rest Tremor and Postural Tremor domains, respectively. The WearGait-PD analytic sample comprised 185 participants, including 100 participants with PD and 85 healthy controls. The Balance and SelfPace tasks were mapped to the CheckPD Balance and GaitWalk domains, respectively. These mappings provided comparisons across related motor task domains.

Conv1D-BiLSTM models were retrained separately within each public dataset using the same overall task-specific modeling workflow. Participant-level splitting was performed before windowing. Preprocessing and channel-wise z-score normalization parameters were estimated using training participants only and applied unchanged to validation or test participants. No CheckPD-derived normalization, public dataset augmentation, probability recalibration, or post hoc decision-rule optimization was applied. Accordingly, these analyses assessed the modeling workflow within each public dataset rather than externally validating a frozen CheckPD-trained model.

Benchmark performance was summarized using the available aggregate ROC-AUC, macro-F_1-score, accuracy, and Brier score. PR-AUC, sensitivity, specificity, and balanced accuracy were not inferred in the absence of participant-level prediction exports. Corresponding CheckPD internal held-out results were used as descriptive reference values. Brier scores were interpreted within each dataset rather than directly compared across datasets because class prevalence differed and probabilities were not recalibrated. Task mappings and benchmarking results are presented in Tables S9 and S10 in Multimedia Appendix 1, respectively.

Results

Participant and Task Availability

The analysis included 679 participants, comprising 344 neurologist-confirmed PD participants and 335 neurologist-confirmed healthy controls. Task-level data availability is summarized in Figure S3 in Multimedia Appendix 1. Because this was a retrospective case-control modeling analysis

using a curated labeled dataset, model performance should be interpreted as internal case-control discrimination from smartphone motor time series, rather than as population-screening effectiveness, calibrated risk prediction, or evidence of clinical deployment readiness.

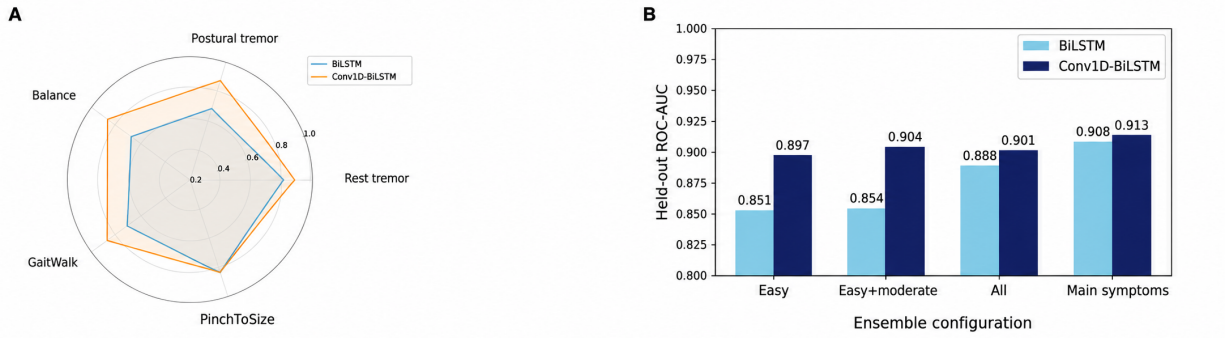
The exploratory demographic sensitivity analysis using age and gender alone showed limited discrimination, with ROC-AUC 0.5784, PR-AUC 0.6170, Brier score 0.2422, accuracy 59.68%, sensitivity 41.29%, and specificity 75.66 %. These results suggest that age and gender alone did not account for the stronger discrimination observed in the smartphone motor time-series models, although residual demographic confounding from measured or unmeasured factors cannot be fully excluded.

Individual-Task Deep Sequence Model Performance

Individual-task model performance was evaluated on the internal held-out cohort described in the Methods section. For each architecture and task, Table 2 reports the mean held-out performance across 5-fold–trained models applied to the same held-out evaluation cohort. Window-level or session-level predictions were aggregated within each task file before metric calculation. For PinchToSize, the file-level denominator exceeded the participant denominator because some held-out participants contributed more than one accepted file. The uncertainty around these individual-task estimates is reported in Table S6 in Multimedia Appendix 1. For example, the Conv1D-BiLSTM GaitWalk model had an observed ROC-AUC of 0.8700 (95% CI 0.7992-0.9551) and PR-AUC of 0.9048 (95% CI 0.8139-0.9674), illustrating the sampling uncertainty around the held-out point estimates.

Across individual tasks, Conv1D-BiLSTM models showed stronger performance than BiLSTM models for the 4 IMU-based tasks. Conv1D-BiLSTM ROC-AUC values were 0.8854 for Rest Tremor, 0.8750 for Postural Tremor, 0.8627 for Balance, and 0.8700 for GaitWalk, compared with corresponding BiLSTM ROC-AUC values of 0.8087, 0.6813, 0.6675, and 0.7090. PinchToSize showed a different pattern, with slightly higher ROC-AUC for BiLSTM than Conv1D-BiLSTM (0.8410 vs 0.8340), suggesting that the touchscreen task may differ from IMU tasks in its optimal sequence-model architecture. Figure 2A summarizes individual-task ROC-AUC results, and full individual-model held-out ROC curves are provided in Figures S4 and S5 in Multimedia Appendix 1.

Figure 2. Held-out internal receiver operating characteristic area under the curve (ROC-AUC) results. (A) File-level fold-averaged ROC-AUC across individual smartphone motor tasks for bidirectional long short-term memory and one-dimensional convolutional bidirectional long short-term memory models. (B) ROC-AUC across complete-case stacked ensemble configurations. Full numerical metrics are reported in Tables 2 and 3. Extended ROC curves are provided in Multimedia Appendix 1. BiLSTM: bidirectional long short-term memory; Conv1D-BiLSTM: one-dimensional convolutional bidirectional long short-term memory; ROC-AUC: receiver operating characteristic area under the curve.



BiLSTM Versus Conv1D-BiLSTM Architecture Comparison

In paired task-fold comparisons across the 5 individual motor tasks, ROC-AUC favored Conv1D-BiLSTM over BiLSTM in 22 of 25 evaluations (2-sided exact sign test, P<.001). Accuracy and macro-F_1-score also favored Conv1D-BiLSTM in 22 of 25 evaluations, with P<.001 for both metrics. These findings indicate a consistent internal performance advantage of the convolutional-recurrent architecture over the recurrent-only architecture for task-specific classification.

This advantage was most evident for the 4 IMU-based tasks. Conv1D-BiLSTM models achieved higher ROC-AUC, macro-F_1-score, and accuracy than the corresponding BiLSTM models for Rest Tremor, Postural Tremor, Balance, and GaitWalk. The largest ROC-AUC improvements were observed for Balance, increasing from 0.6675 to 0.8627, Postural Tremor, from 0.6813 to 0.8750, and GaitWalk, from 0.7090 to 0.8700. These results suggest that the Conv1D front end helped extract local temporal motifs from short multichannel IMU windows before recurrent temporal aggregation. In contrast, PinchToSize showed mixed architecture-dependent results, suggesting that event-driven touchscreen sequences may differ from fixed-duration IMU signals in their optimal modeling architecture.

PinchToSize-Specific Findings

PinchToSize showed a different architecture-dependent pattern from the IMU-based tasks. The BiLSTM model achieved slightly higher ROC-AUC (0.8410 vs 0.8340), macro-F_1-score (0.7226 vs 0.6980), accuracy (72.70% vs 69.91%), sensitivity (69.71% vs 61.14%), and lower Brier

score (0.1903 vs 0.2063) than the Conv1D-BiLSTM model. In contrast, the Conv1D-BiLSTM model achieved higher specificity (83.56% vs 77.33%) and PR-AUC (0.8894 vs 0.8839). These findings suggest that PinchToSize touchscreen time series contain task-relevant case-control information, but their optimal architecture and classification behavior may differ from fixed-duration IMU signals. This is plausible because PinchToSize is an event-driven touchscreen interaction task rather than a fixed-rate inertial recording, and its features reflect scaling accuracy, interevent timing, and adjustment dynamics rather than continuous tremor, balance, or gait movement.

Neural Ensemble Performance

Multimodal stacked ensembles were trained using development-cohort out-of-fold predicted probabilities after participant-level aggregation and complete-case modality intersection. The neural network meta learner was then evaluated on the corresponding held-out complete-case ensemble inputs. As described in the Methods section, all ensemble configurations were evaluated on the same internal held-out cohort of 90 participants with complete data for all 5 analyzed task types.

Table 3 summarizes the held-out performance of BiLSTM and Conv1D-BiLSTM ensembles across four task-set configurations. Among all ensemble models, the Conv1D-BiLSTM Main Symptoms configuration, combining Rest Tremor, GaitWalk, and PinchToSize, achieved the highest observed ranking performance, with ROC-AUC 0.9131 (95% CI 0.8439-0.9695) and PR-AUC 0.9354 (95% CI 0.8901-0.9822). The Conv1D-BiLSTM Easy configuration, combining Rest Tremor, Postural Tremor, and Balance, achieved the strongest binary classification performance at the exported evaluation rule, with macro-F_1-score 0.8534, accuracy 85.56%, sensitivity 73.33%, and specificity 97.78 %. This configuration also had the lowest Brier score among the ensembles, at 0.1346.

Relative to the best single-task value for each metric, all 4 Conv1D-BiLSTM ensembles improved ROC-AUC by 0.0114 to 0.0277. Macro-F_1-score differences were smaller and less consistent, ranging from –0.0247 to +0.0109 (Table S7 in Multimedia Appendix 1). These results suggest complementary information across tremor, balance, gait, and touchscreen fine-motor domains, while also showing that the preferred ensemble configuration depends on the selected performance metric. Development-set OOF complete-case counts are reported in Table S2 in Multimedia Appendix 1, full ensemble held-out ROC curves are provided in Figures S6 and S7 in Multimedia Appendix 1, and Figure 2B summarizes ensemble ROC-AUC results.

SHAP Model-Reliance Findings

SHAP analyses were interpreted as post hoc model-reliance patterns rather than causal explanations or independently validated feature importance findings. Across task-specific Conv1D-BiLSTM models, attribution patterns were broadly consistent with the structure of the corresponding motor tasks (Table 4). The Rest Tremor model showed a gyroscope-dominant reliance pattern, particularly involving gyro_x and gyro_y, with smaller accelerometer contributions. The Postural Tremor model also relied prominently on gyroscope channels, with additional accelerometer contribution, especially acc_x and acc_z. The Balance model showed multiaxial IMU reliance, with stronger accelerometer contribution and secondary gyroscope involvement. The GaitWalk model relied on both accelerometer and gyroscope channels, suggesting use of linear acceleration and angular velocity information during walking. For PinchToSize, the most influential inputs were touchscreen targeting, scaling, and timing features, including dist_to_target, diameter, dist_to_target_pct, timestamp, and expansion_rate.

Table 4. Post hoc SHAP[a] model-reliance summary for task-specific Conv1D-BiLSTM[b] models[c].

Task	Post hoc SHAP model-reliance summary
Rest tremor	Gyroscope-dominant reliance pattern, especially gyro_x and gyro_y, with smaller accelerometer contributions.
Postural tremor	Gyroscope channels remained prominent, with additional accelerometer involvement, especially acc_x and acc_z.
Balance	Multiaxial IMU[d] reliance was observed, with stronger accelerometer contribution and secondary gyroscope involvement.
GaitWalk	Combined accelerometer and gyroscope reliance, suggesting use of both linear-acceleration and angular-velocity information during walking.
PinchToSize	Touchscreen targeting, scaling, and timing features contributed most, including dist_to_target, diameter, dist_to_target_pct, timestamp, and expansion_rate.

[a]SHAP: Shapley additive explanations.
[b]Conv1D-BiLSTM: one-dimensional convolutional bidirectional long short-term memory.
[c]SHAP values were computed on development samples and aggregated across time points or touch-event samples.
[d]IMU: inertial measurement unit.

Overall, the task-specific SHAP patterns support the face validity of the learned representations by showing that each model relied mainly on inputs relevant to its corresponding motor task. These distinct reliance patterns may also help explain why combining different motor domains provided complementary information in the ensemble models. SHAP values were computed on model development samples, the input channels were temporally correlated, and IMU axes remained in a device-centric coordinate frame. Therefore, these attribution patterns indicate internal model reliance

rather than causal physiology, anatomical axis biomarkers, or independently validated digital motor biomarkers. Full post hoc SHAP beeswarm plots are provided in Figure S8 in Multimedia Appendix 1.

Public Dataset Task Domain Benchmarking

Public dataset ROC-AUC estimates ranged from 0.7333 to 0.9000, indicating that the task-specific Conv1D-BiLSTM workflow produced discrimination across related public sensor-task domains. PADS tremor domain results were close to the corresponding CheckPD internal reference results, with ROC-AUC 0.8417 vs 0.8854 for Rest Tremor and 0.8384 vs 0.8750 for Postural Tremor. WearGait-PD GaitWalk achieved ROC-AUC 0.9000, slightly higher than the CheckPD GaitWalk reference value of 0.8700. Balance showed greater between-dataset variation, with ROC-AUC 0.7333 in WearGait-PD compared with 0.8627 in CheckPD.

Macro-F_1-score and accuracy varied more across datasets than ROC-AUC, consistent with the greater sensitivity of classification metrics to class distribution, decision rules, and cohort composition. In PADS, the majority-class accuracy baseline was 77.75%, compared with observed accuracies of 81.52% for Rest Tremor and 77.47% for Postural Tremor. Because class prevalence differed across datasets and probabilities were not recalibrated, Brier scores were interpreted within each dataset rather than compared directly across datasets.

Overall, these retrained public dataset analyses provide complementary task-domain benchmarking evidence for the modeling workflow, but they do not constitute external validation of the CheckPD model. Frozen-model external validation in prospectively collected independent cohorts remains necessary before clinical interpretation or patient-level use.

Discussion

Principal Findings

This study evaluated recurrent, convolutional-recurrent, and stacked-ensemble modeling of CheckPD App smartphone IMU and touchscreen motor-task time series. Conv1D-BiLSTM models outperformed BiLSTM baselines across the 4 IMU-based tasks, whereas PinchToSize showed mixed architecture-dependent results. Multimodal stacking suggested complementarity across tremor, balance, gait, and touchscreen fine-motor domains: the Main Symptoms ensemble achieved the highest-ranking performance, while the lower-burden Easy ensemble achieved the strongest binary classification performance. Together, these findings indicate that brief smartphone motor tasks contain PD-related information suitable for further technical evaluation, while remaining limited to internal case-control discrimination in this study.

The public dataset benchmarks provided additional task domain context. ROC-AUC values ranged from 0.7333 to 0.9000 across retrained PADS and WearGait-PD analyses, with tremor domain results broadly aligned with CheckPD internal reference values and strong GaitWalk discrimination in WearGait-PD. The Balance domain showed the largest between-dataset difference among the benchmarked tasks, with an ROC-AUC of 0.8627 (95% CI 0.7717-0.9415) in CheckPD and 0.7333 (95% CI 0.3996-1.0000) in WearGait-PD. The 95% CIs overlapped. Differences in acquisition protocol, sensor setup, device characteristics, and cohort composition may have contributed to this variation; however, the current benchmarking design cannot isolate the contribution of any individual factor. Accordingly, this difference should be interpreted descriptively rather than as evidence of a statistically established difference in model performance. Because models were retrained separately within each public dataset, these analyses should be interpreted as task domain benchmarking rather than frozen-model external validation.

Interpretation and Comparison With Prior Work

The findings are directionally consistent with prior smartphone and wearable-sensor studies showing that motion, gait, finger-motor, and touchscreen-derived signals can support PD classification or early motor impairment detection. Reviews of automated and AI-enabled PD detection highlight the diversity of modalities, datasets, algorithms, and generalization challenges in this field [11-13]. Related studies have evaluated smartphone walking recordings, hand-drawing traces, LSTM and GRU architectures, quantitative finger-dexterity time-series analysis, smartphone touch gestures, mobile-based PD assessments, gait time-series augmentation, smartwatch movement datasets, and multidomain smartphone-derived features [8,9,14,17,21,36,37,50].

This study extends this literature by evaluating CheckPD App motor tasks collected within a national digital screening context, combining IMU-derived tremor, balance, and gait tasks with a touchscreen fine-motor interaction task. However, direct numerical comparison with previous studies is inappropriate because studies differ substantially in recruitment setting, disease stage, control-group definition, sensor placement, task protocol, validation design, and handling of repeated windows or sessions from the same participant. Table S11 in Multimedia Appendix 1 provides a descriptive comparison with selected prior digital PD studies.

The architecture comparison suggests that the Conv1D front end was useful for short multichannel IMU windows, where local temporal motifs may reflect tremor oscillations, postural sway, or rhythmic gait-related movement before recurrent temporal aggregation. The mixed PinchToSize results are also clinically and technically plausible, because PinchToSize is an event-driven touchscreen interaction task rather than a fixed-rate inertial recording. Its features represent touchscreen scaling accuracy, interevent timing, and adjustment dynamics, which may not benefit from convolutional temporal feature extraction in the same way as IMU signals.

The ensemble results reinforce the multidomain nature of PD motor expression. No single smartphone task is

expected to capture all clinically relevant manifestations. The Main Symptoms ensemble, combining Rest Tremor, GaitWalk, and PinchToSize, achieved the highest ROC-AUC and PR-AUC, whereas the Easy ensemble, combining Rest Tremor, Postural Tremor, and Balance, achieved the strongest macro-F_1-score and accuracy. This distinction is important for future evaluation because the optimal task set may depend on whether the intended priority is ranking discrimination, binary classification, participant burden, task feasibility, or robustness to missingness.

The PD cohort had a mean MDS-UPDRS Part III score of 11.84 (SD 12.61), but these summary values alone do not establish the severity distribution. The overall ROC-AUC values do not show whether discrimination was driven mainly by participants with greater motor impairment or would be similar in a mild PD group. Future validation should assess performance by motor severity, with consistent recording of medication state and disease stage.

The SHAP analysis provided an internal post hoc view of model reliance patterns. Gyroscope-dominant reliance in tremor tasks, multiaxial IMU reliance in balance and gait, and touchscreen targeting, scaling, and timing reliance in PinchToSize were broadly consistent with the corresponding task structures. The PinchToSize task is clinically relevant because impaired dexterity and slowed motor adjustment are common in PD, although its outputs remain touchscreen interaction features rather than clinician-rated bradykinesia measures [17,43,44]. These attribution patterns support the face validity of the learned representations, but they should not be interpreted as causal physiology, anatomical axis biomarkers, or independently validated digital motor biomarkers. Because SHAP explanations were computed on model development samples and the input channels were temporally correlated, future validation should assess whether similar model reliance patterns are reproduced across devices, protocols, medication states, and clinically distinct motor phenotypes.

Clinically, the findings suggest that brief smartphone motor tasks can capture multiple PD-relevant motor domains. Tremor tasks may capture rhythmic oscillatory movement, balance tasks may reflect postural-sway control, gait tasks may capture rhythmicity and dynamic motor control, and PinchToSize may provide a touchscreen-based fine-motor interaction domain. The main strengths of the study include neurologist-confirmed PD and healthy control labels, multiple smartphone motor domains, participant-level partitioning before window or session construction, fold-contained preprocessing and model development, OOF ensemble training, complete-case multimodal fusion, and held-out internal evaluation on a balanced 90-participant cohort.

Limitations

These strengths should be interpreted alongside important limitations. This was a retrospective case-control model development study, not a diagnostic validation study. The control group consisted of neurologist-confirmed healthy controls rather than symptomatic differential diagnosis controls; therefore, performance against essential tremor, atypical parkinsonism, drug-induced parkinsonism, vascular parkinsonism, neuropathy, frailty, arthritis, vestibular disorders, or musculoskeletal gait disorders remains unknown. The held-out cohort was internal, balanced, and complete case, which may not reflect real-world screening prevalence or task-completion patterns. The held-out cohort size was determined pragmatically from complete-case availability rather than through a formal sample size calculation. Although the balanced 90-participant design enabled consistent comparison of all task and ensemble configurations, the resulting performance estimates remain subject to sampling uncertainty; for participant-level classification, one observation corresponds to approximately 1.1 percentage points of accuracy. Larger independent validation cohorts will therefore be required for more precise performance estimation. Development-set task missingness and complete-case ensemble training may have favored participants who were better able to complete the required tasks. Additional limitations include unmeasured device effects, incomplete operating system or device model auditing, lack of probability calibration, and retrained public dataset benchmarking rather than frozen external validation. Exact dopaminergic medication ON/OFF state at task recording was not consistently available, and formal HY stage was unavailable. Therefore, performance could not be evaluated by medication state or formal disease stage, and the models should not be interpreted as medication-state-specific motor biomarkers. A high model score should be viewed only as a signal that further clinical or neurological assessment may be warranted, whereas a low score should not be used to reassure symptomatic individuals or exclude them from appropriate clinical evaluation. In digital measure terminology, this study should be viewed as an internal model development and validation step rather than as full fit-for-purpose validation of a deployable digital measure; future work should address verification across device types, analytical validation across acquisition contexts, and clinical validation for the intended prespecialist assessment use case [51].

Future Work

Future work should prioritize prospective multisite external validation across diverse smartphones, operating systems, clinical settings, and representative screening populations. Such studies should include symptomatic non-PD controls, structured medication-state capture, device and protocol auditing, prespecified subgroup analyses, usability assessment, and linkage to clinician-rated severity outcomes [48,52]. Prospective validation should also capture assistance required, device type, operating system, task environment, task failure reasons, and participant burden, so that missingness and performance can be interpreted alongside real-world feasibility. Calibration under realistic screening prevalence will be essential before any operating threshold can be proposed for patient-level use. Clinical utility should also be evaluated using net-benefit approaches, such as decision-curve analysis, to determine whether model-informed assessment would improve referral decisions compared with usual practice across plausible risk thresholds [53]. For PinchToSize specifically, future studies should

examine its relationship with MDS-UPDRS finger tapping, hand movements, pronation-supination, functional dexterity measures, and Parkinson's Disease Questionnaire-8 or related quality-of-life outcomes [17,44,54]. These steps are necessary before the approach can be considered for future screening support evaluation in real-world settings.

Conclusions

In this retrospective internal validation study of CheckPD app data, deep sequence modeling of smartphone IMU and touchscreen motor-task time series showed promising internal case-control discrimination between neurologist-confirmed participants with PD and neurologist-confirmed healthy controls. Conv1D-BiLSTM models improved performance across most IMU tasks compared with BiLSTM models, and multimodal neural stacking suggested complementarity across tremor, balance, gait, and touchscreen fine-motor interaction domains. These findings support further evaluation of recurrent and convolutional-recurrent models as a technical approach for extracting PD-related information from smartphone motor time series. They do not establish diagnostic validity, calibrated clinical risk prediction, differential-diagnosis performance, or readiness for patient-level use.

Acknowledgments

This national Parkinson disease screening initiative was conducted in honor of Her Royal Highness Princess Maha Chakri Sirindhorn, in recognition of her longstanding commitment to public health, medical education, and the well-being of the Thai population. The authors gratefully acknowledge the collaboration with the National Health Security Office, the Thai Red Cross Society, participating provincial Red Cross chapters, provincial public health offices, hospitals, local coordinators, health volunteers, administrative personnel, and all participants who contributed to the CheckPD program. A full list of participating provincial collaborators is provided in Multimedia Appendix 1.

Patients and members of the public were not formally involved in the design, analysis, interpretation, or writing of this retrospective modeling study. The broader CheckPD program was implemented with support from public health partners and participating communities, and participants contributed data through program activities. Future prospective validation should include structured patient and public involvement.

Generative AI Use

Generative AI tools (ChatGPT) were used to assist with grammar, vocabulary, language editing, document formatting, and consistency checks during manuscript revision.

Funding

This research was funded by the following grants: the Thailand Science Research and Innovation Fund (Program Management Unit for Competitiveness, C01F670185), the National Economic and Social Development Council, Technology and Innovation in Life Sciences National Agency (TN 3/2569, TC (ERP) 31/2568), National Research Council of Thailand (N42A680591, N35E680087), Ratchadapiseksompotch Fund (RA-MF-22/69) and the Center of Excellence grant (CE68_028_3000_004), both awarded by Chulalongkorn University, and the Thai Red Cross Education and Research Committee Fund.

Data Availability

The participant-level CheckPD data generated or analyzed during this study are not publicly available because they contain sensitive health-related information and are subject to participant consent, privacy, ethical approval, institutional governance, and Thai data-protection requirements. Deidentified aggregate results and analysis details are provided in the manuscript and Multimedia Appendix 1. Public benchmark datasets used for contextual comparison are available from their cited public repositories or publications. Requests for additional deidentified aggregate information, or for controlled access to deidentified data for legitimate research purposes, may be directed to the corresponding author and will be considered on a case-by-case basis, subject to approval by the relevant institutional committees, applicable data-sharing agreements, and Thai data-protection regulations.

The full study-specific preprocessing, modeling, evaluation code, trained model weights, and derived prediction files are not publicly available because they include project-specific data ingestion, preprocessing, and platform-integration components linked to governed CheckPD data. The manuscript and Multimedia Appendix 1 provide model architecture, preprocessing steps, hyperparameter ranges, validation procedures, and evaluation metrics to support independent reimplementation. Nonsensitive implementation details or pseudocode may be made available by the corresponding author on reasonable request, subject to institutional approval.

Authors' Contributions

Conceptualization: RB
Data curation: TL
Formal analysis: ZKW, TL
Funding acquisition: RB
Investigation: PS, NL, NJ, VR, PR, CA, SP, KH, DS
Methodology: ZKW, PV, RB
Project administration: PS, NL, NJ, VR, PR, CA, SP, KH, DS, JS, RB

Software: ZKW, TL
Supervision: PV, RB
Validation: PS, NL, NJ, VR, PR, CA, SP, KH, DS, JS, PV, RB
Visualization: ZKW
Writing – original draft: ZKW
Writing – review & editing: JS, RB

Conflicts of Interest

Several authors were involved in the development, implementation, data management, analysis, or clinical supervision of the CheckPD program as part of their academic, institutional, or public-health roles. The authors declare that the present study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1

Model architecture schematics, individual-task and ensemble ROC curves, post hoc SHAP beeswarm plots, hyperparameter-search ranges, selected hyperparameter values, development-set validation performance, held-out CIs, out-of-fold ensemble complete-case counts, incremental-value comparisons, public-dataset task-domain benchmarking details, contextual comparison with selected prior digital Parkinson disease studies.
[DOCX File (Microsoft Word File), 6206 KB-Multimedia Appendix 1]

Checklist 1

EVIDENCE and TRIPOD+AI reporting checklists.
[DOCX File (Microsoft Word File), 25 KB-Checklist 1]

References

1. Tanner CM, Ostrem JL. Parkinson's disease. N Engl J Med. Aug 1, 2024;391(5):442-452. [doi: 10.1056/NEJMra2401857] [Medline: 39083773]
2. Dorsey ER, Bloem BR. The Parkinson pandemic-a call to action. JAMA Neurol. Jan 1, 2018;75(1):9-10. [doi: 10.1001/jamaneurol.2017.3299] [Medline: 29131880]
3. GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol. Apr 2024;23(4):344-381. [doi: 10.1016/S1474-4422(24)00038-3] [Medline: 38493795]
4. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord. Oct 2015;30(12):1591-1601. [doi: 10.1002/mds.26424] [Medline: 26474316]
5. Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis. Neurology. Feb 9, 2016;86(6):566-576. [doi: 10.1212/WNL.0000000000002350] [Medline: 26764028]
6. Bhidayasiri R. Old problems, new solutions: harnessing technology and innovation in Parkinson's disease-evidence and experiences from Thailand. J Neural Transm (Vienna). Jun 2024;131(6):721-738. [doi: 10.1007/s00702-023-02727-1] [Medline: 38189972]
7. Joutsa J, Gardberg M, Röyttä M, Kaasinen V. Diagnostic accuracy of parkinsonism syndromes by general neurologists. Parkinsonism Relat Disord. Aug 2014;20(8):840-844. [doi: 10.1016/j.parkreldis.2014.04.019] [Medline: 24816002]
8. Lim WS, Fan SP, Chiu SI, et al. Smartphone-derived multidomain features including voice, finger-tapping movement and gait aid early identification of Parkinson's disease. NPJ Parkinsons Dis. May 5, 2025;11(1):111. [doi: 10.1038/s41531-025-00953-w] [Medline: 40325040]
9. Zhang Y, Zeng Z, Mirian MS, et al. Investigating the efficacy and importance of mobile-based assessments for Parkinson's disease: uncovering the potential of novel digital tests. Sci Rep. Mar 4, 2024;14(1):5307. [doi: 10.1038/s41598-024-55077-7] [Medline: 38438438]
10. Shokrpour S, MoghadamFarid A, Bazzaz Abkenar S, Haghi Kashani M, Akbari M, Sarvizadeh M. Machine learning for Parkinson's disease: a comprehensive review of datasets, algorithms, and challenges. NPJ Parkinsons Dis. Jul 1, 2025;11(1):187. [doi: 10.1038/s41531-025-01025-9] [Medline: 40595773]
11. Dixit S, Bohre K, Singh Y, et al. A comprehensive review on AI-enabled models for Parkinson's disease diagnosis. Electronics (Basel). 2023;12(4):783. [doi: 10.3390/electronics12040783]
12. Rabie H, Akhloufi MA. A review of machine learning and deep learning for Parkinson's disease detection. Discov Artif Intell. 2025;5(1):24. [doi: 10.1007/s44163-025-00241-9] [Medline: 40092968]
13. Loh HW, Hong W, Ooi CP, et al. Application of deep learning models for automated identification of Parkinson's disease: a review (2011-2021). Sensors (Basel). Oct 23, 2021;21(21):7034. [doi: 10.3390/s21217034] [Medline: 34770340]

14. Varghese J, Brenner A, Fujarski M, van Alen CM, Plagwitz L, Warnecke T. Machine Learning in the Parkinson's disease smartwatch (PADS) dataset. npj Parkinsons Dis. Jan 5, 2024;10(1):9. [doi: 10.1038/s41531-023-00625-7] [Medline: 38182602]
15. Tian F, Fan X, Fan J, Zhu Y, Gao J, Wang D, et al. What can gestures tell? Detecting motor impairment in early Parkinson's from common touch gestural interactions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2019:1-14. [doi: 10.1145/3290605.3300313]
16. Gardoni A, Sarasso E, Basaia S, et al. Handwriting, touchscreen dexterity and bradykinesia measures in Parkinson's disease: a feature selection study. J Neurol. May 11, 2025;272(6):389. [doi: 10.1007/s00415-025-13121-0] [Medline: 40349278]
17. Panyakaew P, Duangjino K, Kerddonfag A, et al. Exploring the complex phenotypes of impaired finger dexterity in mild-to-moderate stage Parkinson's disease: a time-series analysis. J Parkinsons Dis. 2023;13(6):975-988. [doi: 10.3233/JPD-230029] [Medline: 37574743]
18. Iakovakis D, Chaudhuri KR, Klingelhoefer L, et al. Screening of Parkinsonian subtle fine-motor impairment from touchscreen typing via deep learning. Sci Rep. Jul 28, 2020;10(1):12623. [doi: 10.1038/s41598-020-69369-1] [Medline: 32724210]
19. Bhidayasiri R, Trenkwalder C. Parkinson's disease, diagnosis. In: England JD, editor. Encyclopedia of the Neurological Sciences. Academic Press; 2025:386-397. [doi: 10.1016/B978-0-323-95702-1.00196-2]
20. Bhidayasiri R, Martinez-Martin P. Clinical assessments in Parkinson's disease: scales and monitoring. Int Rev Neurobiol. 2017;132:129-182. [doi: 10.1016/bs.irn.2017.01.001] [Medline: 28554406]
21. Rehman A, Saba T, Mujahid M, Alamri FS, ElHakim N. Parkinson's disease detection using hybrid LSTM-GRU deep learning model. Electronics (Basel). 2023;12(13):2856. [doi: 10.3390/electronics12132856]
22. Karbalaie A, Abtahi F, Häger CK. Participant-aware model validation for repeated-measures data: comparative cross-validation study. JMIR AI. Apr 30, 2026;5:e87728. [doi: 10.2196/87728] [Medline: 42060920]
23. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun. Feb 28, 2024;15(1):1829. [doi: 10.1038/s41467-024-46150-w] [Medline: 38418819]
24. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc; 2017:4768-4777. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [Accessed 2026-10-31]
25. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell. Sep 2021;298:103502. [doi: 10.1016/j.artint.2021.103502]
26. Bhidayasiri R, Phumphid S, Sringean J, et al. National implementation of a digital Parkinson's disease screening programme in Thailand: reach, adoption, and real-world performance of the CheckPD app. BMC Public Health. May 8, 2026;26(1):2008. [doi: 10.1186/s12889-026-27654-0] [Medline: 42104339]
27. Manta C, Mahadevan N, Bakker J, et al. EVIDENCE publication checklist for studies evaluating connected sensor technologies: explanation and elaboration. Digit Biomark. 2021;5(2):127-147. [doi: 10.1159/000515835] [Medline: 34179682]
28. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
29. Bhidayasiri R, Sringean J, Phumphid S, et al. The rise of Parkinson's disease is a global challenge, but efforts to tackle this must begin at a national level: a protocol for national digital screening and "eat, move, sleep" lifestyle interventions to prevent or slow the rise of non-communicable diseases in Thailand. Front Neurol. 2024;15:1386608. [doi: 10.3389/fneur.2024.1386608] [Medline: 38803644]
30. Fluttercommunity.dev. Sensors_plus 7.1.0. Pub.dev. 2025. URL: https://pub.dev/packages/sensors_plus/versions/7.1.0 [Accessed 2026-11-11]
31. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. A public domain dataset for human activity recognition using smartphones. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). i6doc.com; 2013:437-442. URL: https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf [Accessed 2026-10-31]
32. Bayat A, Pomplun M, Tran DA. A study on human activity recognition using accelerometer data from smartphones. Procedia Comput Sci. 2014;34:450-457. [doi: 10.1016/j.procs.2014.07.009]
33. Anderson AJ, Eguren D, Gonzalez MA, et al. WearGait-PD: an open-access wearables dataset for gait in Parkinson's disease and age-matched controls. Sci Data. Feb 12, 2026;13(1):440. [doi: 10.1038/s41597-026-06806-2] [Medline: 41680227]

34. FlutterCommunity. AccelerometerEvent class. Sensors_plus library. 2025. URL: https://pub.dev/documentation/sensors_plus/7.1.0/sensors_plus/AccelerometerEvent-class.html [Accessed 2026-10-08]
35. FlutterCommunity. GyroscopeEvent class. Sensors_plus library. 2025. URL: https://pub.dev/documentation/sensors_plus/7.1.0/sensors_plus/GyroscopeEvent-class.html [Accessed 2026-10-08]
36. He T, Chen J, Xu X, Fortino G, Wang W. Early detection of Parkinson's disease using deep NeuroEnhanceNet with smartphone walking recordings. IEEE Trans Neural Syst Rehabil Eng. 2024;32:3603-3614. [doi: 10.1109/TNSRE.2024.3462392] [Medline: 39288062]
37. Juutinen M, Wang C, Zhu J, et al. Parkinson's disease detection from 20-step walking tests using inertial sensors of a smartphone: machine learning approach based on an observational case-control study. PLoS One. 2020;15(7):e0236258. [doi: 10.1371/journal.pone.0236258] [Medline: 32701955]
38. Demrozi F, Pravadelli G, Bihorac A, Rashidi P. Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. IEEE Access. 2020;8:210816-210836. [doi: 10.1109/access.2020.3037715] [Medline: 33344100]
39. Ferrari A, Micucci D, Mobilio M, Napoletano P. Trends in human activity recognition using smartphones. J Reliable Intell Environ. Sep 2021;7(3):189-213. [doi: 10.1007/s40860-021-00147-0]
40. PointerEvent class. Flutter. 2025. URL: https://api.flutter.dev/flutter/gestures/PointerEvent-class.html [Accessed 2026-10-08]
41. PointerEvent.position property. Flutter. 2025. URL: https://api.flutter.dev/flutter/gestures/PointerEvent/position.html [Accessed 2026-10-08]
42. PointerEvent.timestamp property. Flutter. 2025. URL: https://api.flutter.dev/flutter/gestures/PointerEvent/timeStamp.html [Accessed 2026-10-08]
43. Gebhardt A, Vanbellingen T, Baronti F, Kersten B, Bohlhalter S. Poor dopaminergic response of impaired dexterity in Parkinson's disease: Bradykinesia or limb kinetic apraxia? Mov Disord. Sep 15, 2008;23(12):1701-1706. [doi: 10.1002/mds.22199] [Medline: 18649388]
44. Růžička E, Krupička R, Zárubová K, Rusz J, Jech R, Szabó Z. Tests of manual dexterity and speed in Parkinson's disease: not all measure the same. Parkinsonism Relat Disord. Jul 2016;28:118-123. [doi: 10.1016/j.parkreldis.2016.05.009] [Medline: 27185294]
45. Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems 28 (NeurIPS 2015). Curran Associates, Inc; 2015:802-810. URL: https://proceedings.neurips.cc/paper_files/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html [Accessed 2026-11-11]
46. Wolpert DH. Stacked generalization. Neural Netw. Jan 1992;5(2):241-259. [doi: 10.1016/S0893-6080(05)80023-1]
47. Breiman L. Stacked regressions. Mach Learn. Jul 1996;24(1):49-64. [doi: 10.1007/BF00117832]
48. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. Dec 16, 2019;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
49. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. Oct 15, 2018;180(Pt A):68-77. [doi: 10.1016/j.neuroimage.2017.06.061] [Medline: 28655633]
50. Uchitomi H, Ming X, Zhao C, Ogata T, Miyake Y. Classification of mild Parkinson's disease: data augmentation of time-series gait data obtained via inertial measurement units. Sci Rep. Aug 3, 2023;13(1):12638. [doi: 10.1038/s41598-023-39862-4] [Medline: 37537260]
51. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med. 2020;3:55. [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
52. Riley RD, Snell KIE, Archer L, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. Jan 22, 2024;384:e074821. [doi: 10.1136/bmj-2023-074821] [Medline: 38253388]
53. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]
54. Tsolakis N, Maga-Nteve C, Meditskos G, Vrochidis S, Kompatsiaris I, editors. Ince-PD model for Parkinson's disease prediction using MDS-UPDRS I & II and PDQ-8 score. In: Artificial Intelligence Applications and Innovations. Springer Nature Switzerland; 2023:267-278. [doi: 10.1007/978-3-031-34111-3_23]

Abbreviations

AUC: area under the curve
BiLSTM: bidirectional long short-term memory

ChulaPD: Chulalongkorn Centre of Excellence for Parkinson's Disease and Related Disorders
Conv1D-BiLSTM: one-dimensional convolutional bidirectional long short-term memory
CREC: Central Research Ethics Committee
EVIDENCE: Evaluating Connected Sensor Technologies
HY: Hoehn and Yahr
IMU: inertial measurement unit
IRB: Institutional Review Board
LSTM: long short-term memory
MDS-UPDRS: Movement Disorder Society-sponsored version of the Unified Parkinson Disease Rating Scale
OOF: out-of-fold
PADS: Parkinson's Disease Smartwatch
PD: Parkinson disease
PR-AUC: precision-recall area under the curve
ReLU: rectified linear unit
ROC-AUC: receiver operating characteristic area under the curve
SHAP: Shapley additive explanations
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI
VHV: village health volunteer
WearGait-PD: Wearables Dataset for Gait in Parkinson's Disease and Age-Matched Controls

Edited by Timothee Aubourg; peer-reviewed by Carlos Tomaz, Ranjan Das; submitted 12.Aug.2026; initial comments to author 23.Sep.2026; final revised version received 11.Oct.2026; accepted 13.Oct.2026; published 27.Nov.2026

Please cite as:
Win ZK, Laosombut T, Songthung P, Leabthong N, Jeh-voh N, Ratanasirisawad V, Rattanajun P, Anan C, Phumphid S, Horaruengdecha K, Surangsrirat D, Sringean J, Vateekul P, Bhidayasiri R
CheckPD App Smartphone Motor Time-Series Modeling for Parkinson Disease Case-Control Discrimination: Internal Validation Study
JMIR Neurotech 2026;5:e109459
URL: https://neuro.jmir.org/2026/1/e109459
doi: 10.2196/109459

© Zar Kyi Win, Teerawat Laosombut, Phattharat Songthung, Natthaphat Leabthong, Nurhassneeda Jeh-voh, Varis Ratanasirisawad, Piyapon Rattanajun, Chanawat Anan, Saisamorn Phumphid, Kornake Horaruengdecha, Decho Surangsrirat, Jirada Sringean, Peerapon Vateekul, Roongroj Bhidayasiri. Originally published in JMIR Neurotechnology (https://neuro.jmir.org), 27.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Neurotechnology, is properly cited. The complete bibliographic information, a link to the original publication on https://neuro.jmir.org, as well as this copyright and license information must be included.