International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Protocol for a systematic review: Machine learning for multimodal biomarker integration in Parkinson's disease diagnosis — a cross-modality performance comparison and clinical translation readiness assessment

INPLASY202640002
doi: 10.37766/inplasy2026.4.0002
Received: 1 April 2026
Published: 1 April 2026

Pinto Neto, O; Pinho, T.

Corresponding author:
Osmar Pinto Neto

osmar@csusm.edu

Author Affiliation:
California State University San Marcos.

ADMINISTRATIVE INFORMATION

Support - This systematic review is conducted without external funding. No financial support was received for this work.

Review Stage at time of this submission - Preliminary searches.

Conflicts of interest - Osmar Pinto Neto is Guest Editor for the Brain Sciences (MDPI) Special Issue on Degenerative Disease / Parkinson's Disease, to which this review may be submitted. No other financial or non-financial conflicts of interest to declare for any author.

INPLASY registration number: INPLASY202640002

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 1 April 2026 and was last updated on 1 April 2026.

INTRODUCTION

Review question / Objective Primary question: What is the diagnostic classification performance (accuracy, AUC, sensitivity, specificity) of machine learning models applied to different biomarker modalities for Parkinson's disease detection, and how do these modalities compare under a standardized evaluation framework?

Secondary questions: (1) Which ML algorithm families perform best within each biomarker modality? (2) What is the impact of validation strategy (cross-validation vs. holdout vs. external validation) on reported performance? (3) How do hand-crafted features compare with learned representations (deep learning, self-supervised learning) across modalities? (4) What is the clinical translation readiness of each modality considering cost, invasiveness, accessibility, and scalability for population screening?

PICOS: Population: Adults with idiopathic PD (any stage, any medication state) and healthy or neurological controls. Index Test: Any ML/DL algorithm applied to biomarker data (voice, gait, grip force, handwriting, EEG, neuroimaging, blood/CSF, microbiome, or multimodal). Comparator: Healthy controls, other movement disorders, or within-PD severity groups. Outcome: Classification accuracy, AUC, sensitivity, specificity, F1-score. Study Design: Cross-sectional, cohort, case-control original research.

Rationale Parkinson's disease (PD) affects over 10 million people worldwide and remains challenging to diagnose early, as no single definitive biomarker exists. Machine learning (ML) and deep learning (DL) have been increasingly applied across diverse PD biomarker modalities including voice/speech, gait/balance, grip force/handwriting, EEG, neuroimaging, blood/CSF biomarkers, and gut

microbiome, each showing promising diagnostic accuracy.

However, existing reviews focus predominantly on single modalities or single algorithmic families. No comprehensive systematic review has compared ML diagnostic performance across ALL major biomarker modalities using standardized metrics in a single framework. Furthermore, the clinical translation potential of ML-based approaches considering accessibility, invasiveness, cost, and scalability has not been systematically assessed.

Our research group has published original research applying ML to voice analysis (Pinto Neto et al., 2024, Journal of Voice), stabilometric biomarkers (Pinto Neto et al., 2025, Neurological International), and grip force variability (Pinto Neto et al., 2025, NeuroMarkers), providing firsthand expertise across multiple modalities.

This review addresses the critical gap by providing: (1) a cross-modality ML performance comparison with head-to-head metrics, and (2) a clinical translation readiness framework mapping each modality's diagnostic potential against real-world implementation feasibility.

Condition being studied Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by the loss of dopaminergic neurons in the substantia nigra pars compacta. Cardinal motor symptoms include bradykinesia, resting tremor, rigidity, and postural instability. Non-motor symptoms (depression, sleep disorders, cognitive decline, autonomic dysfunction) often precede motor onset by years.

Clinical diagnosis relies on established criteria (UK Brain Bank, MDS Clinical Diagnostic Criteria) but remains subjective, with misdiagnosis rates of 10–25% even among specialists. There is no single gold-standard biomarker for PD. Disease staging uses the Hoehn and Yahr scale (I–V) and the Unified Parkinson's Disease Rating Scale (UPDRS/MDS-UPDRS).

This review encompasses idiopathic PD at all stages, prodromal PD (isolated REM sleep behavior disorder, hyposmia), and differential diagnosis from other parkinsonian syndromes (essential tremor, multiple system atrophy, progressive supranuclear palsy). The focus is on ML-based diagnostic classification, severity staging, and progression prediction using measurable biomarkers across multiple physiological modalities.

METHODS

Search strategy Three concept blocks combined with Boolean AND operators, searched on title/abstract/keywords:

Block A (Disease): "Parkinson*" OR "PD" OR "parkinsonian" OR "parkinsonism"

Block B (Method): "machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network*" OR "support vector machine*" OR "random forest*" OR "gradient boosting" OR "convolutional neural network*" OR "recurrent neural network*" OR "transformer*" OR "ensemble learning" OR "transfer learning" OR "self-supervised learning" OR "classification" OR "pattern recognition" OR "explainable AI" OR "XAI"

Block C (Biomarker): "biomarker*" OR "voice" OR "speech" OR "acoustic" OR "gait" OR "postur*" OR "stabilometr*" OR "balance" OR "grip force" OR "handwriting" OR "motor" OR "EEG" OR "electroencephalogra*" OR "neuroimaging" OR "MRI" OR "SPECT" OR "DaTscan" OR "PET" OR "blood" OR "serum" OR "plasma" OR "CSF" OR "cerebrospinal" OR "alpha-synuclein" OR "metabolom*" OR "proteom*" OR "microbiome" OR "wearable" OR "sensor" OR "EMG" OR "electromyogra*"

Combined: Block A AND Block B AND Block C

Databases: PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science, ACM Digital Library, Google Scholar.

Time frame: January 2019 – March 2026.

Supplementary modality-specific sub-searches will be conducted for: voice/speech, gait/balance, motor/force/handwriting, EEG/electrophysiology, neuroimaging, fluid biomarkers (blood/CSF), gut microbiome, and multimodal fusion, each using dedicated terminology combined with Blocks A and B.

Additional supplementary methods: backward reference list scanning, forward citation tracking via Google Scholar, hand-searching key journals (npj Parkinson's Disease, Movement Disorders, Brain, Neurology, Journal of Neural Engineering, IEEE JBHI, Scientific Reports, Sensors, Frontiers in Neurology), and preprint servers (medRxiv, bioRxiv, arXiv).

Participant or population Adults (≥18 years) diagnosed with idiopathic Parkinson's disease at

any Hoehn and Yahr stage (I–V), in any medication state (ON, OFF, or medication-naive), diagnosed via established clinical criteria (UK Brain Bank Criteria or MDS Clinical Diagnostic Criteria). Comparator groups include healthy age-matched controls, individuals with other movement disorders (essential tremor, atypical parkinsonisms), and prodromal PD cohorts (isolated REM sleep behavior disorder, hyposmia). No restrictions on sex, ethnicity, or geographic location.

Intervention Index test: Any machine learning or deep learning algorithm (including but not limited to SVM, Random Forest, Gradient Boosting, DNN, CNN, LSTM, Transformer, ensemble methods, self-supervised learning models) applied to biomarker data from one or more of the following modalities: voice/speech, gait/balance/wearables, grip force/handwriting, EEG/electrophysiology, neuroimaging (MRI, SPECT, PET), blood/CSF fluid biomarkers, gut microbiome, or multimodal combinations thereof. This is a diagnostic accuracy review; no therapeutic intervention is evaluated.

Comparator Healthy controls (age-matched where available), individuals with other neurological or movement disorders (essential tremor, multiple system atrophy, progressive supranuclear palsy), or within-PD severity comparisons (early vs. advanced stage, different H&Y stages). Studies without a comparator group (e.g., PD severity staging using UPDRS regression) are also eligible and will be analyzed in a separate subgroup.

Study designs to be included Original research: cross-sectional, cohort, case-control studies. Methodological papers with clinical validation on PD patient data. Peer-reviewed full-text conference proceedings (IEEE, ACM). Published January 2019 through March 2026.

Eligibility criteria Additional inclusion criteria: (1) Reports at least one quantitative performance metric (accuracy, AUC, sensitivity, specificity, or F1-score); (2) Sample size ≥20 total participants (PD + controls combined); (3) Published in English or with English abstract available.

Exclusion criteria: (E1) Reviews, meta-analyses, editorials, commentaries, letters, book chapters (used for reference tracking only); (E2) Purely statistical methods without ML (t-tests, ANOVA, unregularized logistic regression); (E3) Studies focused exclusively on treatment response or drug efficacy prediction; (E4) Animal or in vitro studies; (E5) Conference abstracts without full text; (E6) Studies using only simulated/synthetic data; (E7) Duplicate publications with no methodological novelty (most complete version retained).

Information sources Electronic databases: PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science, ACM Digital Library, Google Scholar.

Supplementary sources: (1) Backward reference list scanning of all included studies and relevant prior reviews; (2) Forward citation tracking of seminal papers in each modality via Google Scholar; (3) Hand-searching tables of contents of key journals: npj Parkinson's Disease, Movement Disorders, Brain, Neurology, Journal of Neural Engineering, IEEE Journal of Biomedical and Health Informatics, Scientific Reports, Sensors, Frontiers in Neurology; (4) Preprint servers: medRxiv, bioRxiv, arXiv (flagged separately in quality assessment); (5) Author contact for studies with incomplete reporting of performance metrics.

Main outcome(s) Primary outcomes for cross-modality comparison:
(1) Classification accuracy (%);
(2) Area under the receiver operating characteristic curve (AUC/AUROC), with 95% confidence intervals where reported;
(3) Sensitivity (true positive rate);
(4) Specificity (true negative rate).

These metrics will be extracted per biomarker modality and per ML algorithm family to enable standardized head-to-head performance comparison across all modalities, which constitutes the primary novel contribution of this review. Best-reported and median values per modality will be tabulated.

Additional outcome(s) (1) F1-score, precision, recall;
(2) Regression metrics for severity prediction (RMSE, MAE, R^2 for UPDRS/H&Y score prediction);
(3) External validation and cross-dataset generalization performance;
(4) Feature importance rankings from explainable AI methods (SHAP, LIME, Grad-CAM);
(5) Clinical translation readiness indicators per modality, assessed by the review team on standardized scales: accessibility (1–5), invasiveness (1–5), estimated cost (low/medium/high/very high), time to result, and scalability for population screening.

Data management Records from all database searches will be exported and imported into Zotero reference manager for automated and manual deduplication. Deduplicated records will be

uploaded to Rayyan QCRI (https://rayyan.ai) for blinded independent screening by two reviewers.

Data extraction will use a standardized form in Google Sheets with the following categories: (1) Study characteristics (authors, year, journal, DOI, country, design, funding); (2) Population (sample size per group, PD diagnostic criteria, severity distribution, medication status, demographics); (3) Biomarker modality and feature extraction method; (4) ML methodology (algorithm, validation strategy, hyperparameter tuning, class imbalance handling, pre-trained model use); (5) Performance outcomes; (6) Cross-modality comparison variables (accessibility, invasiveness, cost, scalability). Extraction performed independently by two reviewers with inter-rater reliability assessed via Cohen's kappa.

Quality assessment / Risk of bias analysis Quality assessment will use a modified QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies) framework adapted for ML-based studies, evaluating four domains: (1) Patient Selection — consecutive/random sampling, appropriate case-control design; (2) Index Test (ML Model) — independent validation, appropriate train/test split, absence of data leakage; (3) Reference Standard — PD diagnosis confirmed by established clinical criteria; (4) Flow and Timing — complete inclusion of participants.

Additional ML-specific quality flags: (a) holdout test set used (yes/no); (b) external validation on independent dataset (yes/no); (c) sample size adequacy vs. feature dimensionality; (d) code/data publicly available (yes/partial/no); (e) reporting completeness; (f) baseline method comparison provided.

Assessment performed independently by two reviewers. Inter-rater reliability via Cohen's kappa (κ>0.80 acceptable). Disagreements resolved by consensus or third reviewer.

Strategy of data synthesis Qualitative synthesis: Results organized by biomarker modality (voice/speech, gait/balance, motor/force, EEG, neuroimaging, fluid biomarkers, microbiome, multimodal fusion) with narrative synthesis of predominant ML methods, typical sample sizes, feature extraction approaches, and performance ranges. Temporal trends (2019–2026) described per modality.

Cross-modality comparison (primary contribution): Standardized table presenting best-reported and median performance metrics (AUC, accuracy, sensitivity, specificity) across all modalities. Accompanied by a clinical translation readiness matrix plotting diagnostic performance against accessibility, invasiveness, cost, and scalability.

Quantitative synthesis (if feasible): For modalities with sufficient comparable studies, random-effects meta-analyses using bivariate models for diagnostic accuracy (sensitivity/specificity pairs). Heterogeneity via I^2 statistic and prediction intervals. Summary HSROC curves where appropriate. Reporting follows PRISMA 2020 and PRISMA-DTA guidelines.

Subgroup analysis Planned subgroup analyses:
(1) By ML method family: classical ML (SVM, RF, kNN, logistic regression) vs. deep learning (CNN, LSTM, Transformer) vs. ensemble methods (XGBoost, LightGBM, stacking);
(2) By feature type: hand-crafted features (e.g., jitter, shimmer, MFCCs, center-of-pressure) vs. learned representations (SSL embeddings, CNN features, transfer learning);
(3) By validation strategy: cross-validation only vs. holdout test set vs. external dataset validation;
(4) By disease stage: early/de novo PD (H&Y I–II) vs. established PD (H&Y III–V);
(5) By dataset type: public benchmark datasets vs. proprietary clinical data;
(6) Within each biomarker modality: comparison of algorithm performance.

Sensitivity analysis Planned sensitivity analyses to assess robustness of findings:
(1) Excluding studies rated high risk of bias on QUADAS-2 assessment;
(2) Excluding studies without a proper holdout test set (cross-validation only);
(3) Excluding studies with total sample size below 50 participants;
(4) Excluding preprints not yet peer-reviewed;
(5) Excluding studies using only single widely-used benchmark datasets (e.g., UCI Parkinson's dataset with 31 subjects) to assess whether performance estimates are inflated by dataset-specific overfitting.

Language restriction English language publications only. Non-English publications with an English abstract will be considered if sufficient methodological detail is available.

Country(ies) involved United States, Brazil.

Other relevant information This review is intended for submission to the Brain Sciences (MDPI) Special Issue on Degenerative Disease / Parkinson's Disease. The lead author (OPN) has

published original research applying ML to multiple PD biomarker modalities, including voice analysis (Journal of Voice, 2024; DOI: 10.1016/j.jvoice.2024.04.020), stabilometric biomarkers (Neurological International, 2025; DOI: 10.3390/neurolint17090133), and grip force variability (NeuroMarkers, 2025; DOI: 10.1016/j.neumar.2025.100106). This firsthand expertise across multiple modalities uniquely positions the review team to evaluate cross-modality strengths and limitations.

The review follows PRISMA 2020 guidelines and PRISMA-DTA (Diagnostic Test Accuracy) extension. All data extraction tables, search strings per database, and QUADAS-2 assessments will be provided as supplementary materials with the published manuscript.

Keywords Parkinson's disease; machine learning; deep learning; biomarkers; diagnostic accuracy; voice; gait; neuroimaging; EEG; multimodal; systematic review.

Dissemination plans Results will be submitted for peer-reviewed publication in the Brain Sciences (MDPI) Special Issue on Degenerative Disease / Parkinson's Disease. Findings may also be presented at international conferences including the International Congress of Parkinson's Disease and Movement Disorders and IEEE Engineering in Medicine and Biology Society Conference. The PRISMA checklist, complete search strings, data extraction tables, and quality assessment results will be made publicly available as supplementary materials.

Contributions of each author

Author 1 - Osmar Pinto Neto - Author 1 conceptualized and designed the review, developed the search strategy, will perform screening and data extraction, quality assessment, cross-modality analysis, and draft the manuscript.
Email: osmar@csusm.edu
Author 2 - Tatiana Pinho - Author 2 developed the search strategy, will perform screening and data extraction, quality assessment, cross-modality analysis, and draft the manuscript.
Email: pinho001@csusm.edu