Machine-Learning Benchmarking of Voice-Based Biomarkers for Parkinson's Disease

Xiaowen Luo[1], *Ryszard Auksztulewicz*[1], *Sonja Kotz*[1]

[1]Department of Neuropsychology and Psychopharmacology, Maastricht University, the Netherlands

xiaowen.luo@maastrichtuniversity.nl, ryszard.auksztulewicz@maastrichtuniversity.nl, sonja.kotz@maastrichtuniversity.nl

Abstract

Parkinson's disease (PD) is often accompanied by measurable changes in voice and speech, reflecting the disruption of multiple aspects of motor speech production, including phonation, articulation, prosody, and temporal coordination. Because voice can be collected non-invasively, repeatedly, and with relatively low cost, it has become an increasingly promising digital biomarker for PD. Voice-based measures are of interests not only for patient–control classification, but also for longer-term goals such as tracking symptom progression and enabling scalable remote assessment. At the same time, this area remains methodologically challenging. Performance can vary substantially depending on the speech tasks, acoustic features, preprocessing strategy, and modeling framework used. In addition, many previous studies have relied on relatively small or narrowly defined datasets, limiting conclusions about robustness across tasks and analytic pipelines. Establishing a transparent and reproducible baseline is therefore valuable. The present study addresses this need by developing a baseline machine-learning framework to distinguish individuals with PD from healthy controls using pre-extracted acoustic features from the Bridge2AI-Voice dataset.

We analyzed data from 105 PD and 134 healthy controls from the Bridge2AI-Voice database. This resource includes a heterogeneous set of voice and speech tasks (e.g., prolonged vowel, picture description, diadochokinesis, etc.), together with 131 pre-extracted acoustic features (e.g., loudness, MFCCs, F0-related measures, etc.). Because participants differed in the number of recording sessions completed and in the tasks available within each session, preprocessing first aimed to derive a comparable participant-level representation. One representative session was selected per participant based on task coverage and data completeness. The data were then organized into a participant × feature × task array to preserve the dataset's multitask structure. For classifier input, this three-dimensional representation was aggregated across available tasks into a two-dimensional participant × feature matrix, yielding a single feature vector per participant. This aggregation strategy was applied consistently across all three classifiers.

We benchmarked three standard classifiers: Linear Discriminant Analysis (LDA), linear Support Vector Machine (SVM), and Random Forest (RF). These models were selected to provide a transparent baseline comparison across different modeling assumptions: a shrinkage-based linear discriminant model, a regularized linear margin-based classifier, and a nonlinear ensemble method. All three were applied to the same participant-level feature matrix and used median imputation to handle missing values. LDA and SVM were evaluated with true nested cross-validation, using repeated stratified outer folds for performance estimation and repeated stratified inner folds for model selection. For both linear models, preprocessing also included Yeo–Johnson transformation and standardization. Within this framework, LDA tuned the discriminant solver and shrinkage setting, whereas SVM tuned the regularization parameter of a linear-kernel model. RF was evaluated using the same repeated stratified outer cross-validation framework, but without additional feature transformation.

All three classifiers showed meaningful discrimination between PD and healthy controls. LDA achieved a balanced accuracy of 0.836 ± 0.082 and ROC-AUC of 0.875 ± 0.084, followed by SVM (0.820 ± 0.084, 0.867 ± 0.087) and RF (0.814 ± 0.091, 0.897 ± 0.073). Taken together, these results indicate that the pre-extracted acoustic feature set contains useful information for PD–control discrimination. At this stage, these values should be interpreted as an initial benchmark. More formal model comparison will be needed in future work, including statistical evaluation of cross-validated performance and assessment of whether observed differences remain consistent across alternative preprocessing strategies, task-specific analyses, and additional datasets.

To improve interpretability, we assessed feature stability across outer cross-validation folds. For LDA and SVM, stable features were defined as those whose absolute coefficients ranked in the top 10% within a fold and recurred in at least 90% of outer folds. For RF, stability was assessed using test-fold permutation importance, applying the same recurrence criterion. Although coefficient-based stability in the linear models and permutation-based stability in RF are not identical measures, they remain comparable in terms of cross-validated recurrence. Across models, variability in F1 frequency, along with the mean and standard deviation of voiced segment length, emerged as shared stable features. This suggests that formant-related and temporal voice characteristics contribute robustly to PD–control discrimination in this dataset. Beyond their predictive value, these features may also help identify which aspects of speech production are most consistently altered in PD, particularly articulatory control and speech timing.

Overall, these findings provide an initial benchmark for voice-based PD classification using a large, heterogeneous voice dataset. The present framework can support more systematic comparison of additional models in future work. It also provides a basis for extending the analysis into clinically meaningful applications, such as evaluating symptom severity and tracking longitudinal disease progression. Building on this baseline, future work will test whether these features retain their performance in task-specific analyses, define a smaller and more interpretable feature set, and generalize assessments that remain robust across datasets.

Index Terms: Parkinson's disease, voice and speech biomarker, digital health