Voice-Based Detection and Progression Modeling of Parkinson’s Disease

Team: Decibels & Decisions
Viral Manoj Sheth, Trisha Rane, Shobha Ganapati Bhat, Pranav Rajesh Charakondala

Abstract

Parkinson’s disease (PD) is a progressive neurodegenerative disorder that is often diagnosed only after noticeable motor symptoms appear, even though subtle changes in speech can emerge much earlier. These vocal changes make voice a promising non-invasive and low-cost biomarker for early detection and long-term monitoring. In this project, we develop a machine learning pipeline for both early detection of Parkinson’s disease and progression modeling using acoustic features extracted from voice recordings. We use two public datasets: a small, clinically collected dataset designed for classification and a large longitudinal telemonitoring dataset intended for disease severity regression. Beyond standard within-dataset evaluation, our work focuses on generalization across datasets, testing whether models trained on controlled clinical recordings can operate reliably on real-world home data. While our models achieve strong performance within individual datasets, we observe clear limitations under domain shift, highlighting dataset bias as a major challenge for deploying voice-based Parkinson’s screening systems in practice.

1. Introduction

Parkinson’s disease (PD) affects millions worldwide and is traditionally diagnosed through in-person neurological evaluations and subjective scoring systems such as the Unified Parkinson’s Disease Rating Scale (UPDRS). While effective, these assessments are time-consuming and often detect the disease only after substantial progression. Prior studies have shown that speech impairments—such as increased jitter, shimmer, and vocal instability—can emerge years before prominent motor symptoms, making voice a promising signal for early detection.

The goal of this project is to develop a machine learning pipeline for both Parkinson’s detection and progression modeling using speech data. Beyond achieving high within-dataset accuracy, we focus on a more challenging and realistic question: whether models trained on clinical recordings can generalize to home-collected speech. Addressing this challenge is essential for enabling scalable and remote Parkinson’s monitoring.

2. Data and Preprocessing

To study the clinic-to-home gap, we use two datasets from the UCI Machine Learning Repository that differ in size, purpose, and recording conditions. The UCI Parkinson’s dataset contains 195 sustained vowel recordings from 31 individuals (23 Parkinson’s patients and 8 healthy controls), with multiple recordings per subject leading to class imbalance favouring PD samples. It includes pre-extracted acoustic features such as jitter, shimmer, harmonics-to-noise ratio (HNR), recurrence period density entropy (RPDE), detrended fluctuation analysis (DFA), and pitch period entropy (PPE), collected in controlled clinical settings.

The Parkinson's Telemonitoring dataset contains 5,875 longitudinal recordings from 42 Parkinson's patients collected in home environments over several months. In addition to acoustic features, it includes Motor and Total UPDRS scores for severity modeling, but contains no healthy controls, which later limits generalization.

To support cross-dataset experiments, we aligned semantically equivalent acoustic features across datasets and retained a shared subset of common features. All features were standardized using z-score normalization, and subject-wise splitting was applied to prevent data leakage, ensuring recordings from the same individual never appeared in both training and test sets.

3. Methodology

Our analysis followed a staged approach that evolved as we uncovered limitations in initial assumptions.

3.1 Phase 1: Baseline Modeling

We established baseline performance using standard classification and regression models. For Parkinson's detection, we trained Logistic Regression, Support Vector Machines, Random Forests, and Gradient Boosting models on the UCI clinical dataset. Class imbalance was addressed using balanced class weights, and GroupKFold cross-validation was employed to prevent subject leakage across folds.

Tree-based models performed best in this controlled setting, with Random Forest and Gradient Boosting achieving accuracies above 93 percent, demonstrating that acoustic features are highly discriminative in clinical conditions. In parallel, regression models were trained on the Telemonitoring dataset to predict Motor and Total UPDRS scores. While tree-based regressors captured some nonlinear patterns, overall predictive power remained limited, underscoring the challenge of modeling disease progression from voice features alone.

3.2 Phase 2: Cross-Dataset Generalization

We next evaluated whether models trained in controlled clinical settings could generalize to real-world data. In the forward transfer experiment, a classifier trained on the UCI dataset was tested on Telemonitoring recordings. Since all Telemonitoring samples correspond to Parkinson's patients, recall was used as the primary metric, achieving 88.39 percent, only moderately lower than within-dataset performance. This result indicates that the model learned robust disease-related vocal patterns rather than dataset-specific artifacts.

In contrast, the reverse transfer experiment performed poorly. A regression model trained on Telemonitoring data assigned high severity scores to both Parkinson's and healthy UCI subjects, resulting in near-random performance. This behaviour was traced to an intercept bias caused by the absence of healthy controls in the Telemonitoring dataset, highlighting the importance of dataset diversity over dataset size.

3.3 Phase 3: Domain Shift and Interpretability

To better understand the asymmetric behavior observed in the transfer experiments, we analyzed the structure of the feature space and the model's decision process.

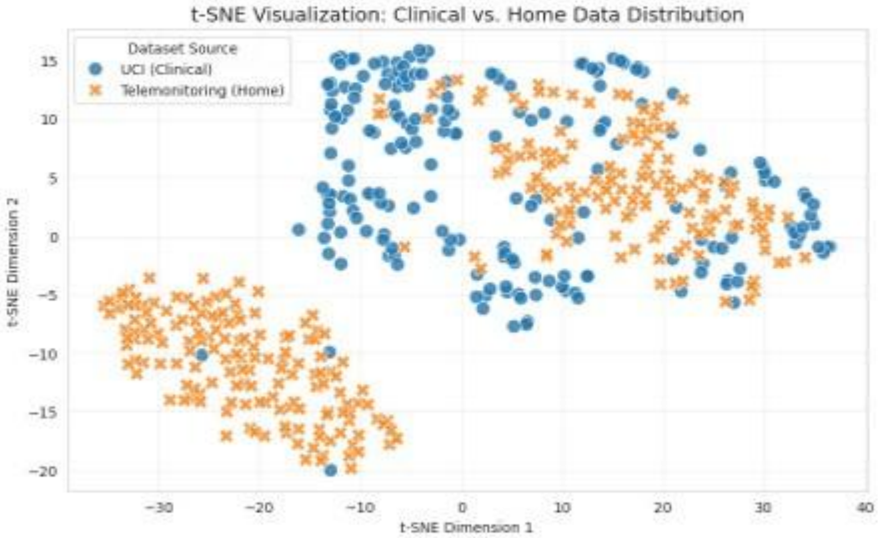

Figure 1. t-SNE visualization showing a clear separation between clinical (UCI) and home (Telemonitoring) recordings, indicating strong domain shift.

We first visualized the combined feature space using t-SNE. As shown in Figure 1, the clinical and home datasets formed clearly separated clusters, confirming that the two datasets occupy distinct regions of the feature space. This visualization provided direct evidence of a strong domain shift caused by differences in recording environments and data collection conditions.

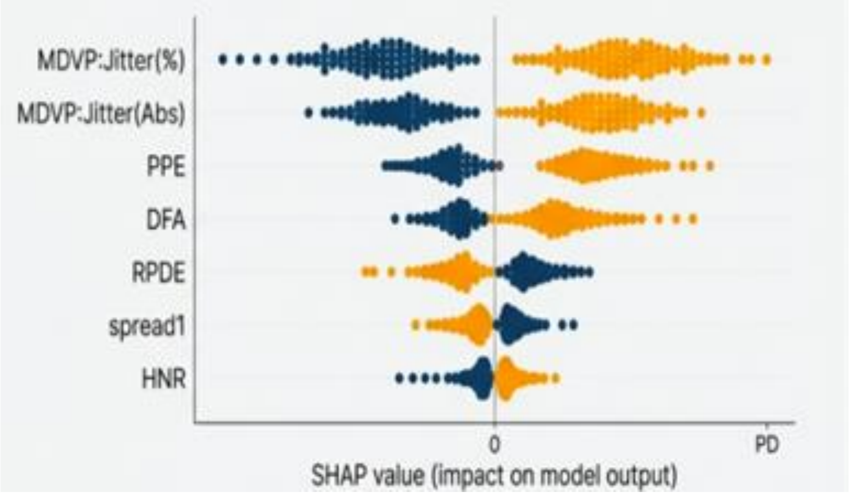

Figure 2. SHAP summary plot highlighting Pitch Period Entropy (PPE), jitter, and DFA as the most influential features for Parkinson's detection.

To further validate that the model relied on clinically meaningful signals, we applied SHAP to interpret the Random Forest classifier. Figure 2 shows that Pitch Period Entropy (PPE), jitter, and DFA were the most influential features driving predictions toward Parkinson's disease. These results align well with established clinical understanding of vocal instability in Parkinson's patients and increased confidence that the model's decisions were not arbitrary.

3.4 Phase 4: Unified Hybrid Modeling

Based on the insights from the previous phases, we designed a unified hybrid model to reduce domain bias while preserving clinically relevant signals. We constructed a combined training dataset consisting of all UCI clinical samples and a randomly undersampled subset of the Telemonitoring data. This approach prevented the much larger Telemonitoring dataset from overwhelming the clinical signal while still exposing the model to variability from home recordings.

A Random Forest trained on this unified dataset produced a single domain-aware detector. The final model achieved approximately 90 percent overall accuracy, with very high sensitivity for Parkinson's detection. While specificity for healthy subjects remained lower, the unified approach demonstrated that careful data engineering can partially bridge the gap between clinical and real-world environments. This unified model serves as the basis for the final evaluation presented in the Results section.

3.5 Phase 5: Addressing Class Imbalance with Synthetic Oversampling

Although the unified hybrid model in Phase 4 achieved strong overall performance and high sensitivity to Parkinson's Disease (PD), it remained conservative, with comparatively low recall for healthy speakers due to class imbalance in the combined dataset. To address this limitation, we introduced a fifth phase focused on rebalancing class distributions.

In Phase 5, we applied SMOTE to the training data to synthetically oversample the healthy class, ensuring that oversampling was restricted to the training split to prevent data leakage. Retraining the unified classifier on the balanced dataset led to a substantial improvement in healthy recall, increasing from approximately 50 percent to nearly 70 percent, while maintaining high PD recall. These results demonstrate that targeted class-balancing techniques can improve fairness across classes without significantly compromising disease detection performance.

4. Results

To reduce domain bias, we constructed a unified hybrid dataset by combining all clinical samples with a balanced subset of Telemonitoring data. A Random Forest trained on this hybrid dataset achieved approximately 90 percent overall accuracy. Sensitivity for Parkinson's detection reached 98 percent, reflecting strong disease detection performance. Prior to class balancing, healthy-class recall remained lower at around 50 percent, indicating a conservative screening-oriented model.

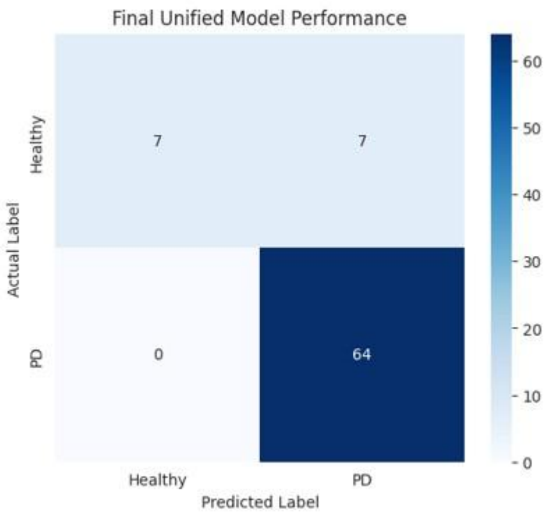

Figure 3. Confusion matrix of the final unified model, showing high sensitivity for Parkinson's detection with moderate specificity.

After applying synthetic oversampling to address class imbalance, healthy-class recall improved to approximately 70 percent while maintaining high sensitivity for Parkinson's detection. These results demonstrate that targeted class-balancing techniques can improve performance for healthy subjects without substantially compromising the model's ability to detect Parkinson's disease.

5. Discussion and Limitations

Prior studies have shown that machine learning models can achieve high accuracy for Parkinson's disease detection using voice data, especially when trained and evaluated within a single, controlled dataset. Our work builds on this line of research by examining how well such models perform when applied across datasets collected in different recording environments, which is critical for real-world use.

Our findings suggest that dataset bias and recording environment differences can have a larger impact on performance than model complexity alone. Even relatively large datasets may fail to generalize if they lack important population diversity, such as the inclusion of healthy control subjects. We also observed that relying only on pre-extracted acoustic features limits the model's ability to capture temporal speech patterns, which are likely important for understanding disease progression.

6. Conclusion and Future Work

In this project, we developed an end-to-end machine learning pipeline for voice-based Parkinson's disease detection and progression modeling. While strong performance was achieved within and across datasets, robust generalization remains challenging due to domain shift and dataset bias. Future work should focus on collecting balanced home-environment datasets, incorporating raw audio modeling, and applying domain adaptation techniques. Overall, our work highlights both the promise and the current limitations of voice-based biomarkers for real-world Parkinson's monitoring.

7. Team Collaboration

Our team worked collaboratively throughout the project, with responsibilities distributed across different stages of the analysis. We jointly explored the datasets, aligned features, and established baseline models, then focused on specific components including classification, disease severity regression, cross-dataset generalization, and interpretability using SHAP. Regular discussions allowed us to review results, address challenges, and refine the project direction as our understanding of generalization evolved.