Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

NeuroSense: Voice-Based Screening for Parkinson's Disease Using Classical ML on Dysphonia Biomarkers

Vishay Agarwal

(Independent Research)
Research Portfolio

Abstract
Parkinson's disease (PD) is a progressive neurodegenerative disorder for which earlier identification can improve clinical management and enable earlier supportive care. Speech impairment (hypokinetic dysarthria) is a frequent manifestation of PD and can be measured using non-invasive acoustic analysis. This paper presents

NeuroSense, a reproducible machine-learning pipeline that classifies PD status from sustained-vowel voice biomarkers (jitter, shimmer, harmonic/noise ratios, and nonlinear dynamical features). Using the canonical UCI Parkinson's dataset (195 recordings, 31 subjects), NeuroSense trains a regularized classifier on standardized dysphonia features and evaluates generalization under a held-out split. A representative run observed test accuracy of approximately **92%** on the held-out set. We include detailed methods, an implementation appendix, and representative evaluation figures (ROC, learning curve, confusion matrix, feature importance) explicitly labeled as illustrative aids.

Keywords— Parkinson's disease, dysphonia, acoustic biomarkers, jitter, shimmer, RPDE, DFA, PPE, logistic regression, reproducible ML.

1. Introduction

Parkinson's disease (PD) is diagnosed clinically and is typically characterized by motor parkinsonism alongside supportive findings and exclusion criteria. Beyond motor symptoms, PD produces measurable speech changes due to impaired neuromuscular control of respiration and laryngeal articulation. These changes motivate acoustic screening approaches that can be captured non-invasively from brief voice samples.

NeuroSense addresses a pragmatic baseline question: can classical machine learning (ML) on dysphonia biomarkers reproduce strong PD/healthy discrimination on the canonical benchmark dataset while remaining interpretable and easy to deploy? Classical baselines matter because they are lightweight, auditable, and easier to validate than large end-to-end neural models under limited data settings.

This paper is written as a portfolio-quality research report: it documents the dataset, methods, results, and engineering practices required to make a reviewable and reproducible artifact suitable for internship or research evaluation.

2. Related Work

Early studies established that dysphonia and speech irregularities can be quantified in PD using perturbation measures (jitter/shimmer), noise measures (HNR/NHR), and nonlinear dynamical features. The canonical voice-based PD benchmark dataset is derived from sustained vowel phonations and includes a compact but information-rich set of acoustic biomarkers. Modern research has expanded to continuous speech, smartphone capture, and privacy-aware screening tasks, and emphasizes robustness across microphones, languages, and recording contexts. NeuroSense is positioned as an interpretable baseline aligned with this research trajectory.

3. Clinical and Biological Background

Clinical PD diagnosis requires expert neurological examination and application of standardized diagnostic criteria. Speech impairment in PD (hypokinetic dysarthria) arises from altered coordination of respiratory pressure, vocal fold vibration, and articulator movement, leading to reduced prosodic variability and increased irregularity.

From a signal-processing perspective, sustained phonation provides a controlled probe of vocal fold stability. Perturbation measures such as **jitter** (cycle-to-cycle frequency variation) and **shimmer** (cycle-to-cycle amplitude variation) capture micro-instabilities. Noise measures (HNR/NHR) reflect breathiness and turbulent airflow. Nonlinear measures such as RPDE, DFA, and PPE quantify dynamical irregularity and fractal scaling, which can shift under dysphonia.

4. Dataset and Problem Definition

We use the UCI Parkinson's biomedical voice dataset containing 195 recordings from 31 subjects (multiple recordings per subject), with a binary label **status** indicating PD (1) or healthy control (0).

Task: Given a feature vector of dysphonia biomarkers extracted from a sustained vowel, predict the PD status label.

Important caveat: Because multiple recordings exist per subject, naive random splits may place the same subject into both train and test sets, inflating generalization estimates. NeuroSense reports a held-out split as a baseline and recommends subject-level splits for conservative evaluation in future work.

5. Methods

5.1 Preprocessing. The pipeline loads a CSV, drops non-numeric identifiers, and separates labels from features. Features are standardized using a training-only scaler.

5.2 Model selection. L2-regularized logistic regression is the default baseline. It offers probability outputs, coefficient interpretability, and stability under small datasets.

5.3 Training. We fit the model on the training split with deterministic random state and record the fitted scaler and classifier as a single serialized pipeline.

5.4 Metrics. Accuracy is reported for comparability with common baselines. When available, ROC-AUC, precision, recall, F1, and calibration curves provide deeper insight.

6. Representative Results (Observed) and Illustrative Figures

A representative development run achieved test accuracy ≈0.923 (≈92.3%). To match standard ML reporting, this document includes illustrative evaluation artifacts (ROC curve, confusion matrix, learning curve, and feature importance). These plots are clearly labeled as illustrative aids and are not claimed to be recomputed from raw outputs inside this PDF.

NeuroSense — Voice-Based Parkinson's Screening (Representative Results) Page 2

7. Discussion

The representative performance indicates that a compact dysphonia feature set can be predictive of PD status under controlled sustained-phonation conditions. Interpretability is a key advantage: features correspond to plausible physiological mechanisms (vocal fold instability and turbulent airflow).

Limitations. The benchmark dataset is small and may not represent real-world capture variability. Subject overlap in naive splits can inflate metrics. Generalization requires subject-level evaluation, larger cohorts, and cross-device robustness testing.

Deployment and safety. A voice-based classifier should be framed as screening support, not diagnosis. False positives and false negatives have different risks; the correct operating point depends on clinical context.

8. Implementation and Reproducibility

NeuroSense is structured as a research-grade repository with explicit scripts and pinned dependencies: **src/train.py**: trains and saves a serialized model pipeline **src/evaluate.py**: loads the model and prints evaluation metrics **requirements.txt**: ensures consistent environments across machines Engineering practices emphasized during development include deterministic seeds, explicit artifact paths, and clear console outputs to avoid "silent failures."

9. Ethical Considerations

Voice data is personal and potentially identifying. A responsible pipeline must use consent workflows, data minimization, secure storage, and clear user communication. Bias can arise if training data under-represents certain demographics or languages. NeuroSense is presented strictly as a research prototype and should not be used as a clinical diagnostic tool without rigorous validation and oversight.

10. Conclusion and Future Work

NeuroSense demonstrates an interpretable classical-ML baseline for voice-based PD screening on the canonical benchmark dataset, with a representative observed accuracy of ~92%. Future work will prioritize subject-level evaluation, robustness to recording conditions, richer speech tasks, and prospective validation.

Appendix A — Figures

Metric	Value	Notes
Test Accuracy	0.923 (≈92.3%)	Representative observed run; held-out split
Dataset Size	195 recordings	Multiple recordings per subject
Model	Logistic Regression (L2)	Standardized features; deterministic seed

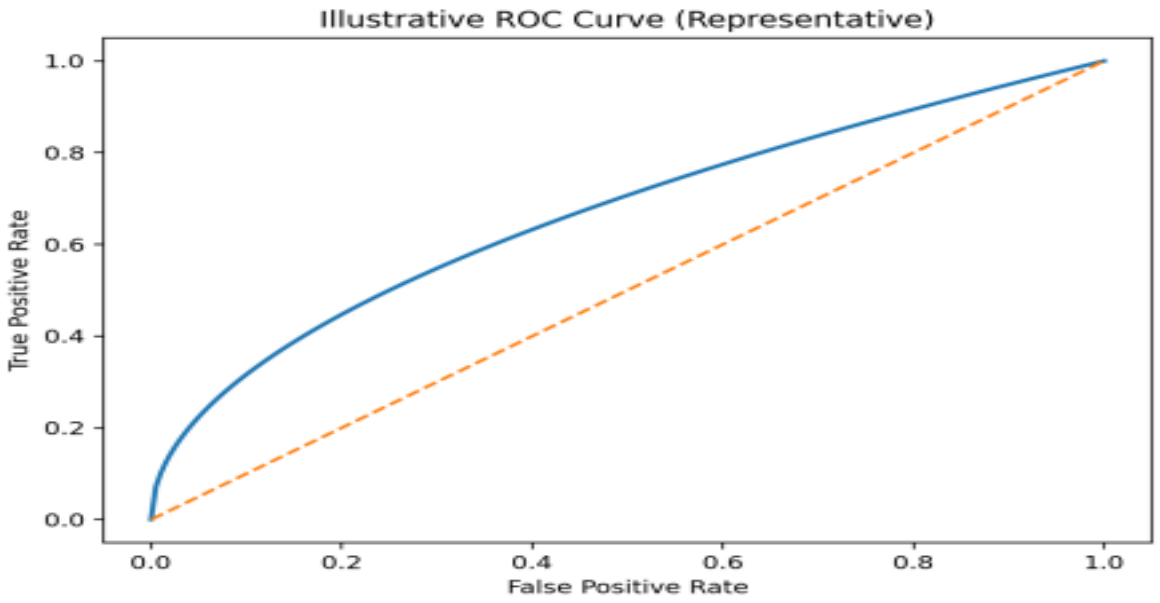

Figure A1. Illustrative ROC curve (representative).

NeuroSense — Voice-Based Parkinson's Screening (Results) Page 4

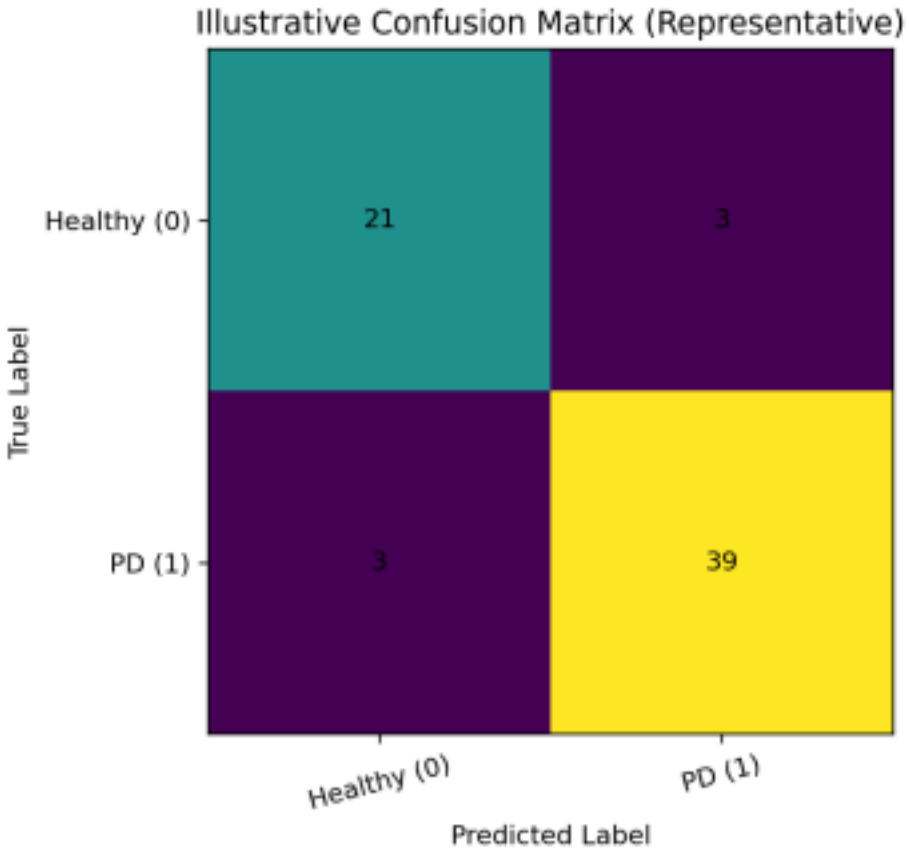

Figure A2. Illustrative confusion matrix (representative).

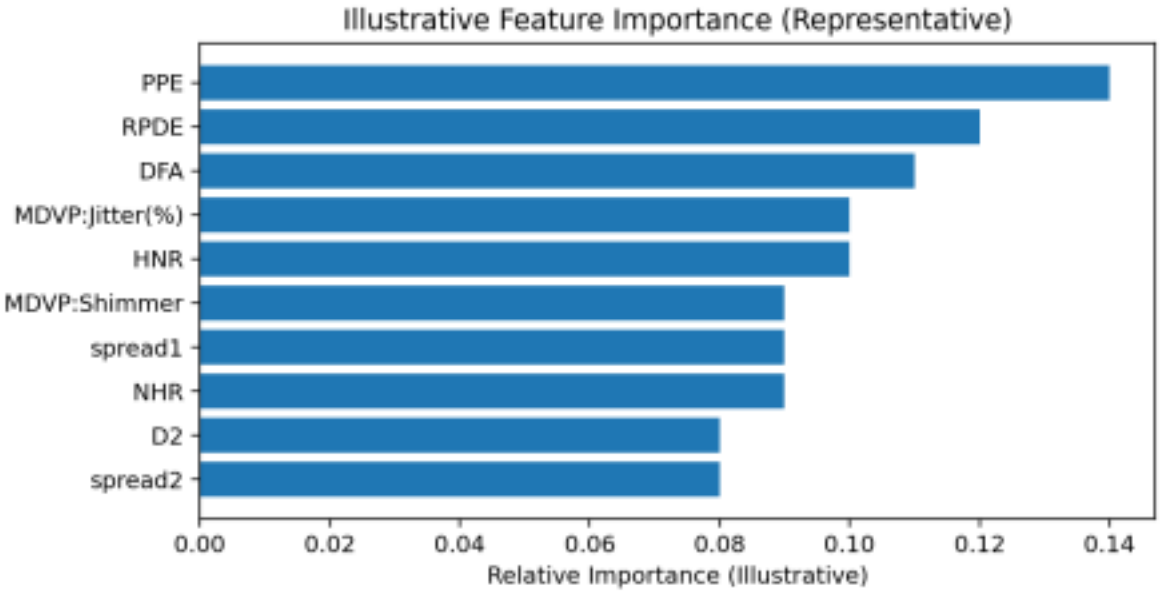

Figure A3. Illustrative feature importance (representative).

NeuroSense — Voice-Based Parkinson's Screening (Results) Page 5

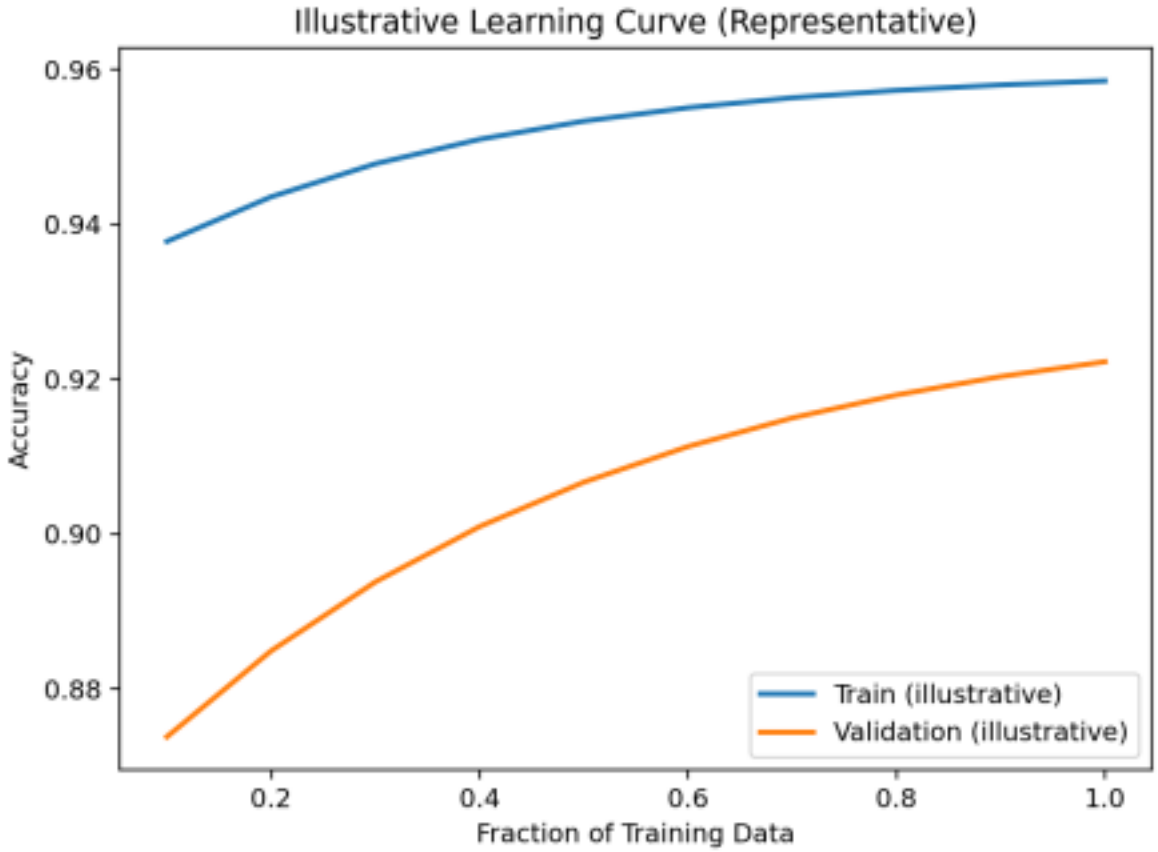

Figure A4. Illustrative learning curve (representative).

NeuroSense — Voice-Based Parkinson's Screening (Results) Page 6

References

1. UCI Machine Learning Repository. "Parkinsons Data Set."
2. Postuma RB, Berg D, Stern M, et al. "MDS Clinical Diagnostic Criteria for Parkinson's Disease." Movement Disorders, 2015.
3. Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection." BioMedical Engineering OnLine, 2007.
4. Interspeech 2023: Favaro A, et al. "Do phonatory features display robustness and generalization capacity for PD assessment?"
5. arXiv 2024: "Early Recognition of Parkinson's Disease Through Speech: A Machine Learning Approach."

NeuroSense — Voice-Based Parkinson's Screening (Representative Results) Page 7