PROJECT ABSTRACT

Parkinson's Disease Detection from Voice using Machine Learning

An SVM classifier detecting Parkinson's signatures from 22 vocal biomarkers (UCI, 195 recordings), with a voice-test demo. Delivered built-to-order with notebook, model, report, PPT and viva kit.

---

01. Title

Parkinson's Disease Detection from Voice using Machine Learning

02. Introduction / Background

Up to 90% of people with Parkinson's disease develop vocal impairment — reduced loudness, breathiness, pitch instability — often years before classic motor symptoms prompt a diagnosis. Those changes are measurable: jitter, shimmer, harmonics-to-noise ratio and nonlinear measures like PPE capture them numerically from a sustained 'ahh'. The UCI Parkinson's dataset packages 195 recordings from 31 subjects with 22 precomputed biomarkers — ideal for teaching biomedical ML and its central lesson: validate by subject, not by recording. This project trains an RBF SVM with leave-one-subject-out cross-validation, explains predictions through biomarker profiles, and states its screening-only limits plainly.

03. Problem Statement

With 195 recordings from only 31 people, a random train/test split leaks speakers and inflates accuracy — the small-data trap most biomedical-ML student builds fall into. Biomedical projects also risk implying diagnostic capability they do not have. The project validates with 31-fold leave-one-subject-out CV, chooses model capacity to match the data size (SVM, not deep learning), and carries a screening-aid-only disclaimer through every deliverable.

04. Objectives

- Train an RBF-SVM on the 22 UCI vocal biomarkers.
- Validate with leave-one-subject-out cross-validation (31 folds).
- Tune hyperparameters with nested CV; benchmark against logistic regression and random forest.
- Explain each prediction through per-biomarker healthy-range comparisons.
- Build a voice-test demo with capture simulation, waveform and risk score.
- Document screening-only limits and confounders explicitly.

05. Proposed Methodology

First, the 195 recordings' 22 features are loaded with subject IDs and standardized (scaler fit on train folds only). Second, an RBF SVM is tuned via nested CV with logistic regression and random forest baselines. Third, 31-fold leave-one-subject-out validation produces accuracy, sensitivity, specificity and ROC-AUC. Fourth, permutation importance ranks the biomarkers.

Fifth, the demo is built: voice-profile selection, animated capture, biomarker extraction display, risk score with per-feature explanations, and the structured screening report.

06. System Architecture / Working

- Data: 195 recordings, 31 subjects, 22 biomarkers.
- Preprocessing: StandardScaler fit on train folds.
- Model: SVM-RBF (C=10, γ=0.01), nested-CV tuned.
- Validation: leave-one-subject-out, 31 folds; no speaker leakage.
- Demo: profile → capture simulation → biomarkers vs healthy ranges → risk score → report.
- Guardrail: screening-aid-only disclaimer on every output.

07. Software Requirements

- Python 3 with scikit-learn, NumPy, pandas, Matplotlib.
- Jupyter Notebook; trains on a laptop CPU in under 2 minutes.
- UCI Parkinson's dataset (public).
- Any modern browser for the demo.

08. Expected Outcomes

An SVM screener with a design-target LOSO accuracy of ~92%, sensitivity/specificity and ROC-AUC, biomarker importance rankings, a working voice-test demo with explanations, an explicit limits-and-ethics section, and the complete report/PPT/viva kit.

09. Applications

- Low-cost screening research and telemedicine triage studies.
- Longitudinal voice monitoring in research cohorts.
- Teaching rigorous small-data ML methodology.

10. Future Scope

- Raw-waveform deep models on larger voice corpora.
- Longitudinal tracking of biomarker drift per patient.
- Clinical validation pathway — explicitly out of scope for the teaching build.