Voice-Based Detection and Monitoring of Parkinson's Disease

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

Instructor: *Nigel Bosch*
Course: *IS 557 Applied Machine Learning*
Semester: *Fall 2025*

Introduction

- Parkinson's Disease (PD) is a progressive neurological disorder affecting speech, movement, and cognition.
- Subtle voice abnormalities, reduced pitch variation, increased noise, instability often emerge **years before clinical diagnosis**.
- Voice is a **non-invasive**, **low-cost**, and **scalable** biomarker for early PD detection and longitudinal monitoring.

Our goal is to build a machine learning pipeline for:

1. **Early detection** of PD using acoustic features.
2. **Progression modeling** by predicting motor and total UPDRS scores over time.

Dataset 1: UCI Parkinson's Dataset

Dataset 1: UCI Parkinson's Dataset

- 195 voice recordings from 31 subjects (multiple recordings per subject)
- Status label: 0 = Healthy, 1 = Parkinson's
- Class imbalance: 147 PD vs. 48 healthy (~75% PD)
- Acoustic features: Jitter, Shimmer, Harmonics-to-Noise Ratio (HNR), RPDE, DFA, PPE, etc.
- Well-suited for early detection tasks
- Primarily used for classification modeling

Insights

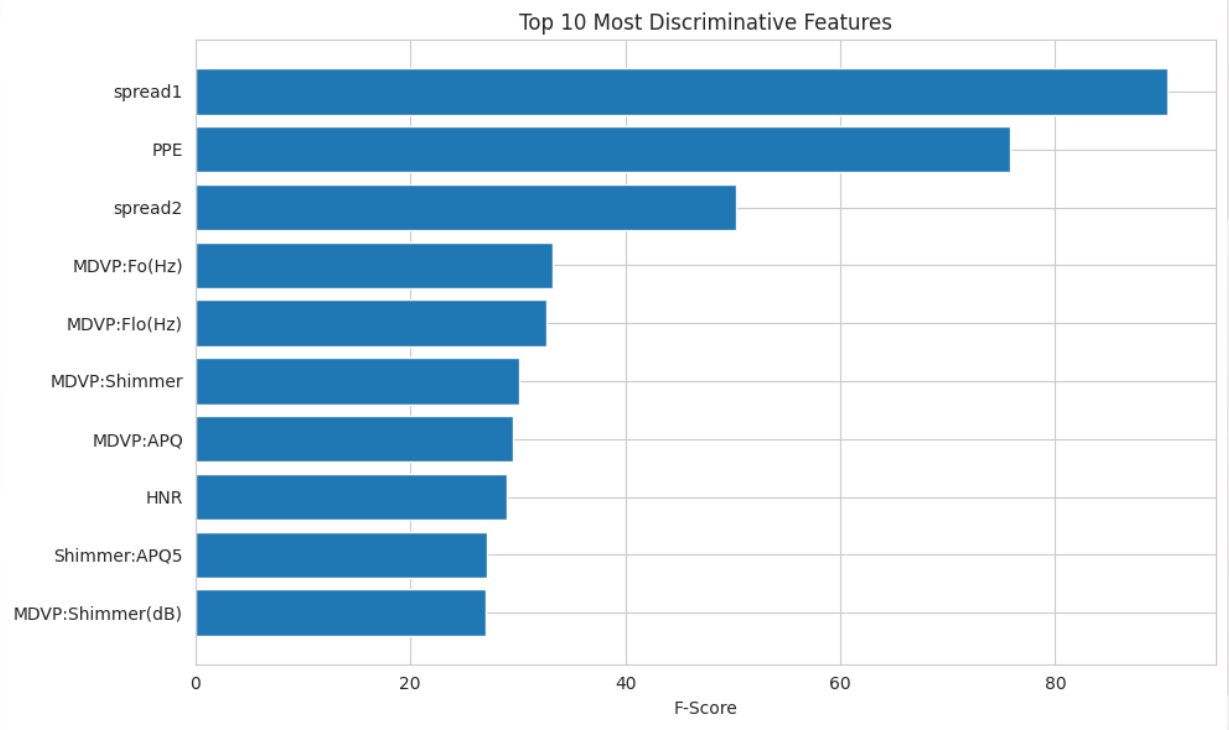
- Strong class imbalance toward Parkinson's (3:1 ratio)
- Models trained without correction may overpredict PD

Preprocessing & Feature Selection

- Removed non-predictive columns: name and status split into features and labels
- Applied StandardScaler to normalize all acoustic features
- Ensured subject-wise grouping using the name column to prevent data leakage
- Used SelectKBest (ANOVA F-test) to rank feature importance

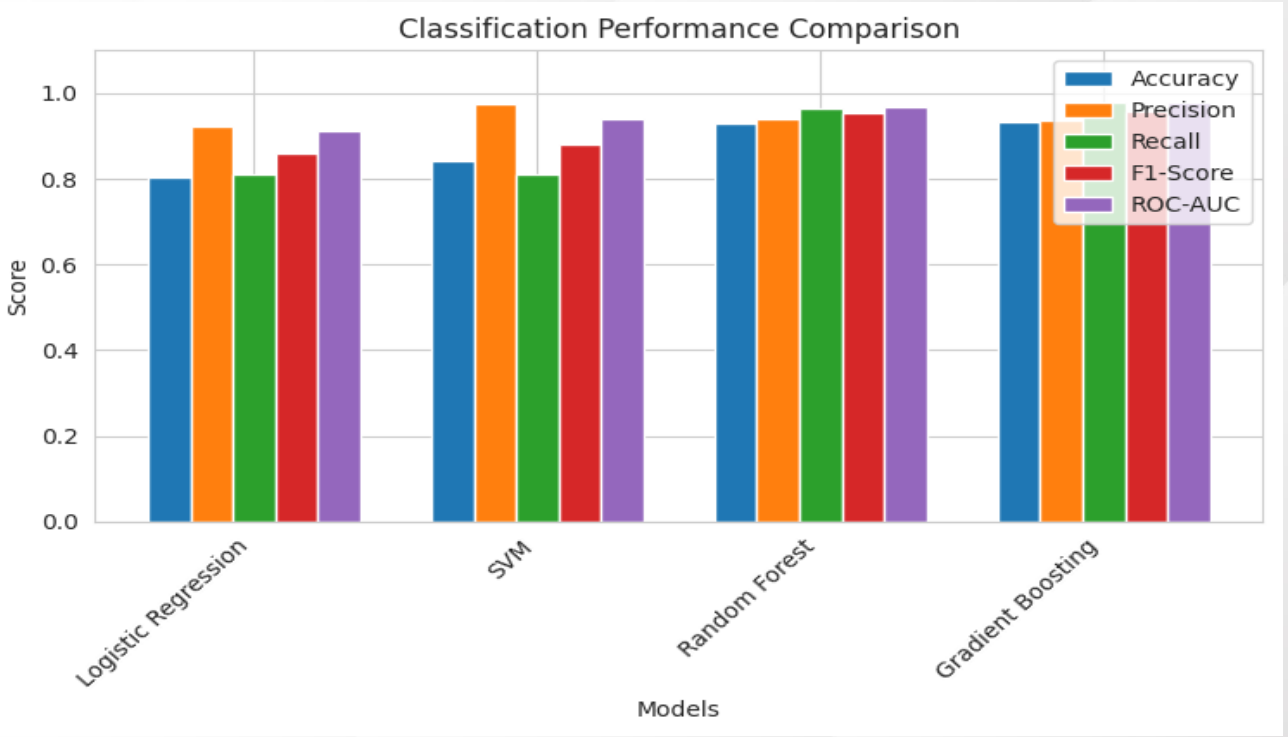
Top features included:

- PPE, RPDE, DFA, spread1, NHR, shimmer, jitter
- High-F-score features show strong separation between healthy and PD vocal patterns



Classification Models & GroupKFold CV

- Trained four baseline classifiers: **Logistic Regression**, **SVM (RBF)**, **Random Forest**, **Gradient Boosting**
- Used **balanced class weights** to address PD-heavy dataset imbalance
- Applied **GroupKFold (5-fold)** to ensure *no subject appears in both train and test sets*
- Evaluated models using **Accuracy, Precision, Recall, F1-Score, ROC-AUC**
- Gradient Boosting & Random Forest show strongest generalization to unseen subjects



Classification Results (UCI Dataset)

- **Gradient Boosting** achieved the **highest performance**
- Linear models (Logistic Regression, SVM) trailed but remained competitive
- All models exceeded the **90% F1-Score threshold** except Logistic Regression
- High recall across tree models indicates strong PD detection capability

Model	Accuracy	Accuracy Std	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.81	0.03	0.92	0.81	0.86	0.91
SVM	0.84	0.07	0.97	0.81	0.88	0.94
Random Forest	0.93	0.03	0.94	0.97	0.95	0.97
Gradient Boosting	0.93	0.03	0.94	0.98	0.96	0.98

Dataset Overview

Dataset 2: Parkinson's Telemonitoring Dataset

- **5,875 voice recordings** from **42 Parkinson's patients**
- **Longitudinal dataset** with multiple recordings per patient (collected over several months)
- Includes **22 acoustic features**: Jitter, Shimmer, HNR, RPDE, DFA, PPE, etc.
- Contains two continuous clinical targets: **Motor UPDRS** (motor disability score) & **Total UPDRS** (overall Parkinson's severity score)
- No healthy controls, dataset is designed for **disease severity prediction**, not classification
- Well-suited for **regression modeling** and **progression tracking**

Insights

- Strong **patient-level variation** due to multiple time-series samples
- Requires **GroupKFold** to prevent leakage across patient samples
- Acoustic biomarkers show measurable correlation with UPDRS severity (e.g., PPE, RPDE, DFA)

Data Exploration

Insights

- Both UPDRS scores show wide variation, useful for regression
- Patients have 100 -170 recordings each, enabling progression modeling
- Longitudinal structure requires careful validation (subject-wise splitting)
- Feature distributions suggest nonlinear patterns, motivating feature selection

UPDRS Score Statistics	motor_UPDRS	total_UPDRS
count	5875	5875
mean	21.3	29.02
std	8.129	10.7
min	5.038	7
25%	15	21.37
50%	20.87	27.58
75%	27.6	36.4
max	39.51	54.99

UPDRS Distributions & Progression Patterns

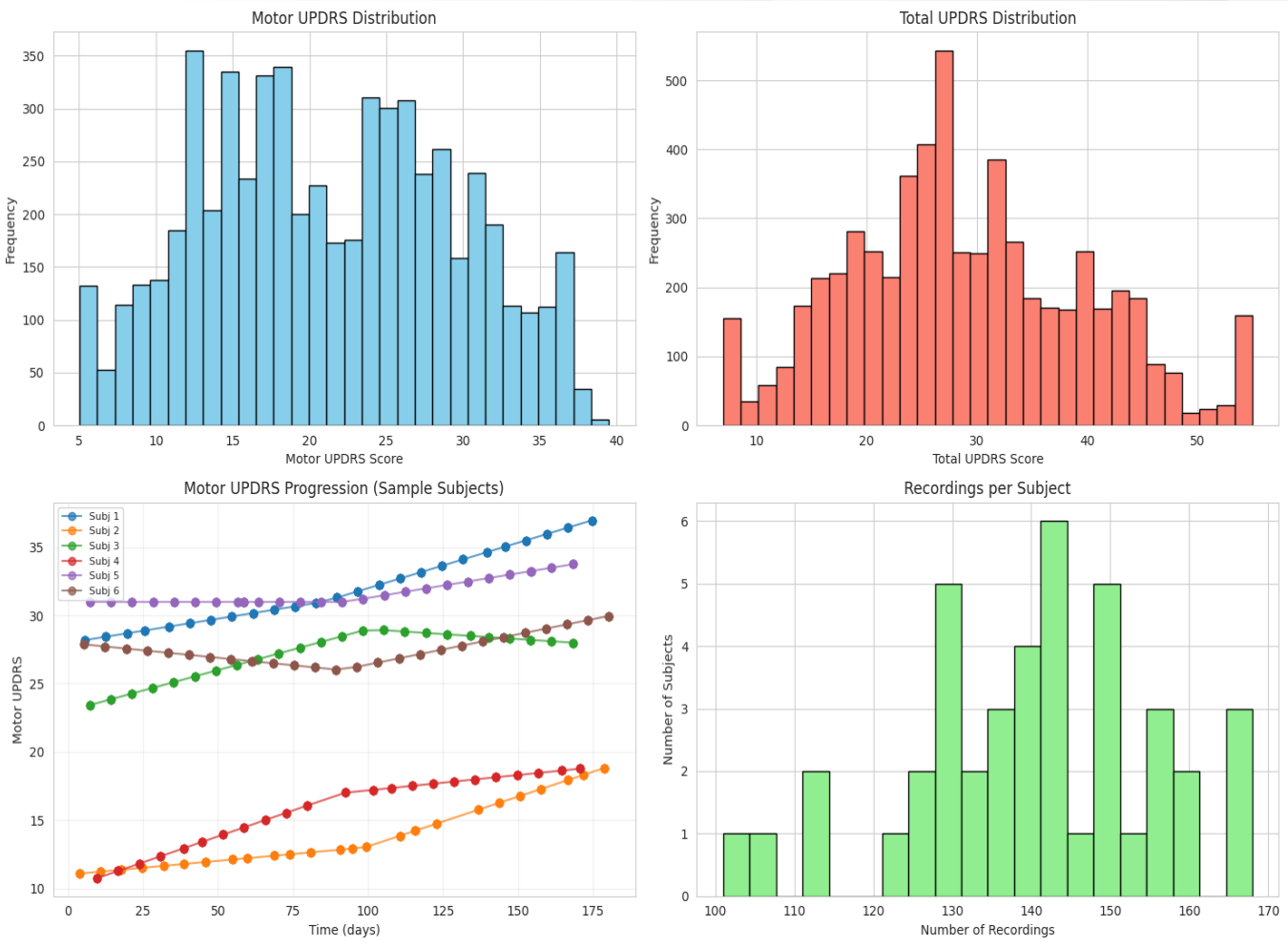
Insights From the Visuals

• Motor UPDRS Distribution: Scores cluster between 10 - 35, indicating moderate-to-severe disease stages dominate the dataset.

•Total UPDRS Distribution: Broader spread (7- 55), capturing multi-domain impairments.

•Progression Curves: Different patients progress at different rates; UPDRS increases slowly but steadily.

•Recordings-per-Subject Histogram: Most subjects have 130- 160 recordings, ideal for time-series modeling but increases risk of overfitting.



Preprocessing & Subject-Wise Splitting

Subject-wise 80/20 Train–Test Split

- Identified all **42 unique subjects** and randomly shuffled them
- Assigned **33 subjects** to training and **9 subjects** to testing
- Ensures all recordings from a patient stay in only one split
- Prevents data leakage and gives a **realistic generalization check (model evaluated on completely unseen patients)**
- **Final Sample Distribution: Train Set (4,578 recordings), Test Set (1,297 recordings)**

Feature Scaling

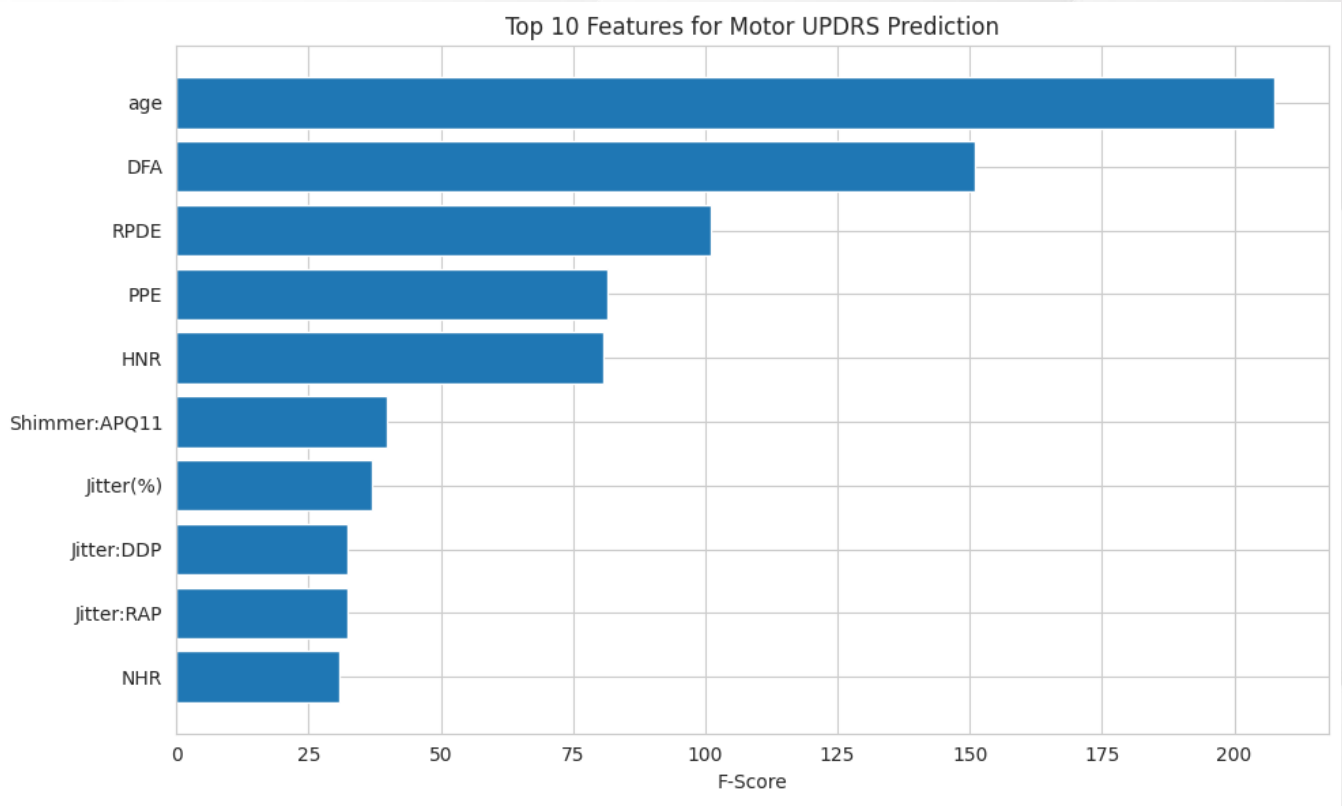
- Applied **StandardScaler** after splitting
- Fit scaler only on training data; applied same transform to test set
- Standardized features (zero-mean, unit-variance) for stable model training

Feature Importance for UPDRS Prediction

- Removed non-predictive columns: **name** and **status** split into features and labels
- Applied **StandardScaler** to normalize all acoustic features
- Ensured **subject-wise grouping** using the name column to prevent data leakage
- Used **SelectKBest (ANOVA F-test)** to rank feature importance

Top features included:

- **PPE**, **RPDE**, **DFA**, **spread1**, **NHR**, **shimmer**, **jitter**
- High-F-score features show strong separation between healthy and PD vocal patterns



Regression Model Results (Motor UPDRS)

- Linear, Ridge, and Lasso regressions give **similar performance**.
- **Ridge (α=10)** achieves the best generalization (Test MAE ≈ 5.94).
- Test R^2 is **~0 or negative** - linear models cannot fully explain variability.
- Tree models (**Random Forest, Gradient Boosting**) **severely overfit** (very low Train MAE, very high-Test MAE).
- Overall: **Ridge** is the most stable and least overfitted model.

Model	Train_MAE	Train_R^2	Test_MAE	Test_RMSE	Test_R^2	Overfit_Gap
Linear Regression	6.51	0.15	5.95	7.73	-0.01	0.56
Ridge (α=1.0)	6.52	0.15	5.94	7.72	-0.01	0.57
Ridge (α=10)	6.52	0.15	5.94	7.72	-0.01	0.58
Lasso (α=1.0)	6.97	0.05	6.11	7.70	-0.01	0.86
Random Forest	0.58	0.98	8.95	10.01	-0.70	-8.37
Gradient Boosting	0.92	0.97	8.80	10.03	-0.71	-7.88

Regression Model Results (Total UPDRS)

- Linear, Ridge, and Lasso again perform similarly.
- **Lasso (α=1.0)** delivers the lowest Test MAE (= 7.94).
- Test R^2 stays **negative** across linear models.
- Tree models massively **overfit** due to small dataset size.
- Overall: **Lasso** provides the best generalization for Total UPDRS.

Model	Train_MAE	Train_R^2	Test_MAE	Test_RMSE	Test_R^2	Overfit_Gap
Linear Regression	7.94	0.20	8.12	11.83	-0.17	-0.18
Ridge (α=1.0)	7.95	0.20	8.11	11.81	-0.17	-0.16
Ridge (α=10)	7.95	0.20	8.10	11.80	-0.17	-0.15
Lasso (α=1.0)	8.27	0.13	7.94	11.58	-0.12	0.34
Random Forest	0.47	0.99	13.58	15.41	-0.99	-13.12
Gradient Boosting	1.14	0.97	12.81	14.96	-0.88	-11.68

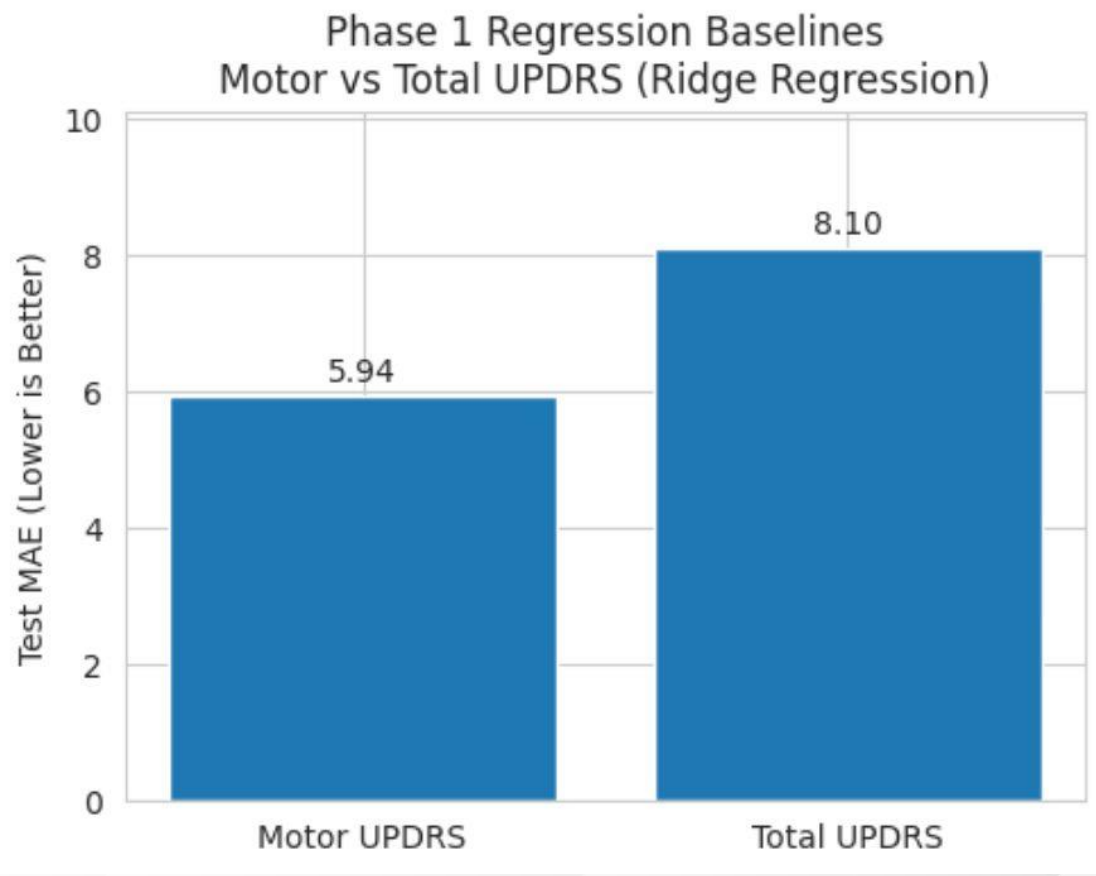
Phase 1 Baseline Results Summary & Key Insights

Baseline Performance Summary

- **Classification (UCI Dataset): Gradient Boosting & Random Forest delivered 93% accuracy** and **F1 up to 0.96**, showing strong ability to detect Parkinson's from voice features.
- **Regression (Telemonitoring Dataset): Linear & Ridge models achieved MAE ~6 (Motor UPDRS)** and **MAE ~8 (Total UPDRS)** but very low R^2, revealing progression prediction is harder than binary detection.

Key Insights

- Acoustic features (PPE, RPDE, DFA, jitter/shimmer) are highly discriminative for early PD detection.
- Tree-based models generalize best for classification, while linear models avoid overfitting in regression.
- Subject-wise GroupKFold was essential to prevent leakage.
- Phase 1 establishes a **strong detection baseline** and exposes challenges in **severity prediction**, motivating later phases.



Phase 2a: Why We Need Cross-Dataset Generalization

Goal: To check if a model trained on one dataset (UCI) can recognize Parkinson's in a completely different dataset (Telemonitoring).

Why this matters:

- The UCI dataset is small and recorded in controlled conditions.
- The Telemonitoring dataset is large, noisy, and collected in real-life environments.
- A strong model should work even when the data source changes.
- We specifically wanted to check **robustness**, not just within-dataset accuracy.

Core question:
"Does the model learn real Parkinson's vocal patterns, or is it just memorizing UCI data?"

How We Performed Cross-Dataset Generalization

Step 1: Feature Alignment

Selected 15 acoustic features common to both datasets (e.g., jitter, shimmer, RPDE, DFA, PPE).

Step 2: Standardization

Scaled both datasets to remove differences in microphone volume, amplitude, and noise levels.

Step 3: Model Training

- Trained a **Random Forest classifier** on the UCI dataset
- UCI labels: 0 = Healthy, 1 = PD

Step 4: Cross-Dataset Testing

- Tested the model on **5875 Telemonitoring samples**
- Telemonitoring contains **only PD patients**, so recall (sensitivity) is used

Goal: Test the model's real-world generalization.

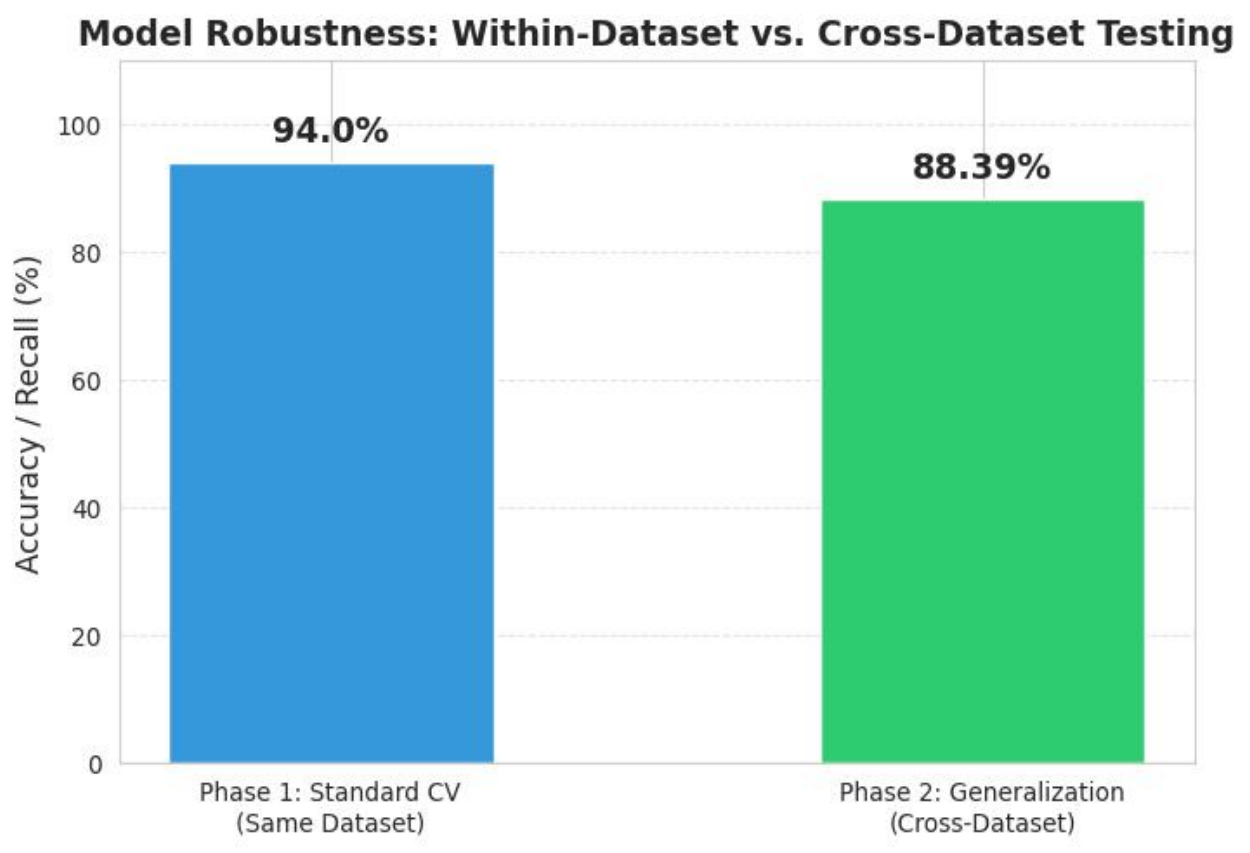
Results of Cross-Dataset Generalization (Phase 2a)

Performance:
- Correctly identified PD patients: **5193 / 5875**
- Generalization Recall: **88.39%**

Comparison:
- Phase 1 (same dataset): **94% accuracy**
- Phase 2a (cross dataset): **88.39% recall**
- Only ~6% drop - excellent stability

Conclusion: Strong cross-dataset generalization

The model successfully recognizes Parkinson's vocal patterns across different environments and patient populations.



Phase 2b: Reverse Transfer Learning

Goal: Find out whether a model trained on severity prediction (UPDRS regression) can also detect healthy vs PD in the UCI dataset.

Why this matters?

- It tests the “reverse direction” of transfer learning
- Severity and diagnosis are related but not identical
- Telemonitoring dataset contains only PD patients, no healthy controls
- We check if severity features can indirectly separate healthy vs PD

Core question:
“Does predicting disease severity automatically help identify who has the disease?”

How We Performed Reverse Transfer Learning

Step 1: Train a Regression Model on Telemonitoring

- Used Random Forest Regressor
- Target: **total UPDRS** (disease severity score)

Step 2: Apply This Model on UCI

- Predicted “pseudo-UPDRS” for each UCI subject
- Healthy people should get low scores
- PD people should get high scores

Step 3: Evaluate as Classification

- Compared predicted scores for healthy vs PD
- Used **ROC-AUC** to measure separation

Goal:
See if severity patterns transfer to diagnosis.

Results of Reverse Transfer Learning (Phase 2b)

Predicted Mean UPDRS on UCI:

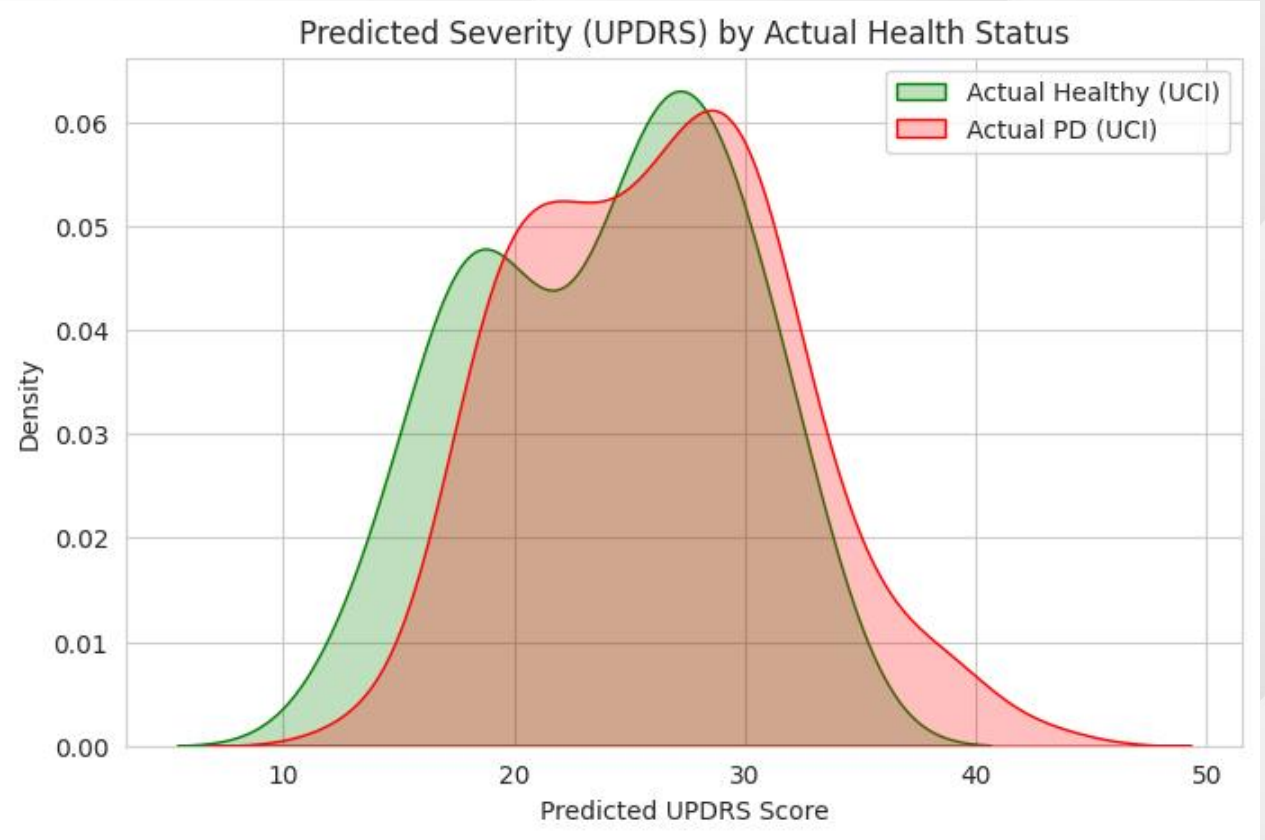
- Healthy: **23.97**
- PD Patients: **26.22 (***too close - poor separation*)

ROC-AUC Score: 0.6022

- 0.5 = random guessing
- 1.0 = perfect separation
- 0.60 = weak signal

Conclusion: Limited reverse transfer

- Severity-based patterns do **not** map well to diagnosis
- Telemonitoring has no healthy samples - the model never learned the “healthy voice pattern”
- Diagnostic classification requires different cues than severity prediction



Phase 3: Manifold Learning & Explainable AI

Phase 2 showed strong generalization (UCI → Telemonitoring). But Phase 2b reverse transfer struggled

- We needed to **understand why** the datasets behave differently
- We also wanted to **open the black-box model** and see:
- Which features matter most?
- Are the decisions clinically reasonable?

Phase 3 has **two goals**:

- **t-SNE (Phase 3A)**
  Visualize how different the two datasets are (UCI clinical vs Telemonitoring home recordings).
  This is called **"domain shift"**.
- **SHAP (Phase 3B)**
  Explain your Random Forest model:
  Which features are actually driving the Parkinson's prediction?

What We Did in Phase 3: t-SNE and SHAP Pipeline

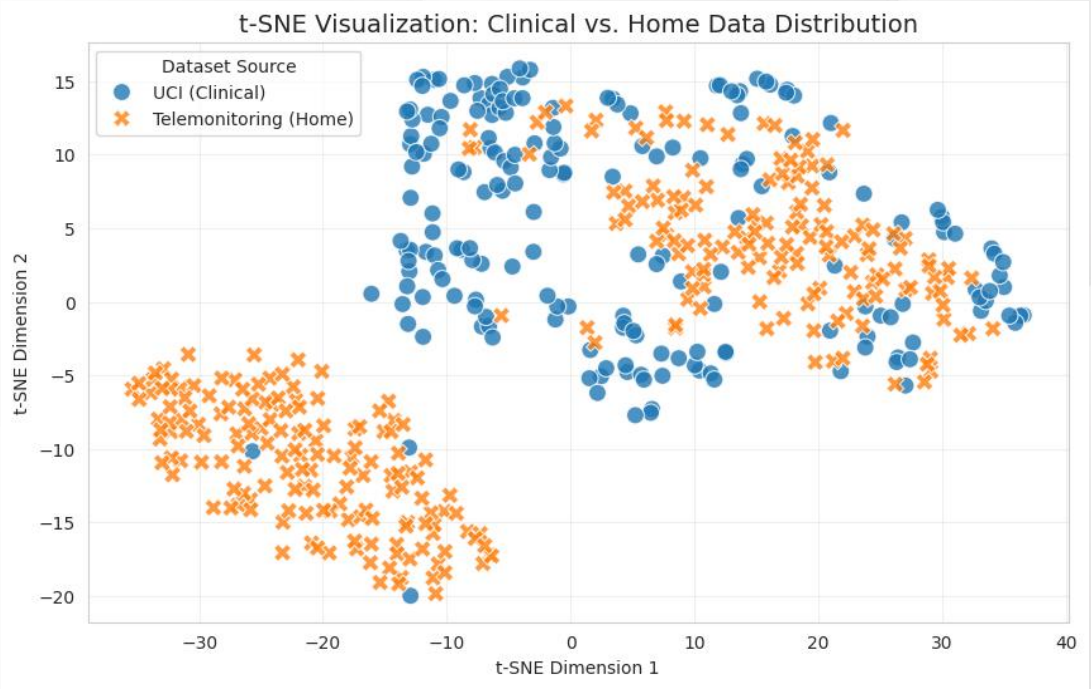
t-SNE (Domain Shift Visualization)

- Combined scaled UCI features and a sample of Telemonitoring features
- Ran **t-SNE** to reduce 15D acoustic features - 2D map
- Colored points by dataset: UCI (Clinical) vs Telemonitoring (Home)

SHAP (Model Explainability)

- Used the trained **Random Forest classifier** from earlier phases
- Built a **SHAP TreeExplainer**
- Computed SHAP values for all training samples
- Generated a **SHAP summary plot** to rank important features

Key idea: Use geometry (t-SNE) + explanation (SHAP) to understand both data and **model**.



Domain Shift & Key Predictive Features

t-SNE Results:

- UCI and Telemonitoring points form **visibly separate clusters**
- Confirms a strong **recording-environment gap (domain shift)**
- Explains why the **reverse transfer (Phase 2b)** had limited success

SHAP Results:

- Features like **MDVP:Jitter(%)**, **MDVP:Jitter(Abs)**, and related jitter metrics show strong impact on PD predictions
- High jitter values push the model towards predicting **"Sick"**
- Confirms that the model is using **clinically meaningful acoustic cues**

Takeaway message:

- Phase 3 shows *why* transfer is hard (domain shift)
- And *how* the model makes decisions (key acoustic features), making your system more interpretable and trustworthy.

Phase 4: The unified "HYBRID" Model

- Phase 2 showed strong forward generalization (UCI → Telemonitoring).
- Phase 2b showed weak reverse transfer (Telemonitoring → UCI).
- Phase 3 showed the underlying reason: UCI and Telemonitoring form two separate clusters in t-SNE → domain shift.

Need for Phase 4:

- A single unified model that works on **both** datasets
- Remove bias towards either dataset
- Make the model **domain invariant** (robust to clinical vs home recordings)

Goal of Phase 4:

Build a balanced, mix-trained hybrid model that performs consistently across environments.

How We Built the Unified Hybrid Model

Step 1: Balance the Domains

- UCI has ~195 samples; Telemonitoring has thousands
- Keep all UCI data and undersample Telemonitoring to match
- Creates a **50/50 domain-balanced dataset**

Step 2: Merge & Shuffle

- Combine both datasets into one unified pool
- Shuffle to remove any ordering or source bias

Step 3: Train–Test Split (80/20)

- Split the unified dataset while keeping both domains in each set

Step 4: Scale & Train the Hybrid Model

- Fit StandardScaler on the unified training set
- Train a Random Forest with `class_weight='balanced'`
- Produces a single **hybrid model** that learns from both clinical and home audio

Results of the Final Unified Model (Phase 4)

Overall Performance (Mixed Clinical + Home Test Set)

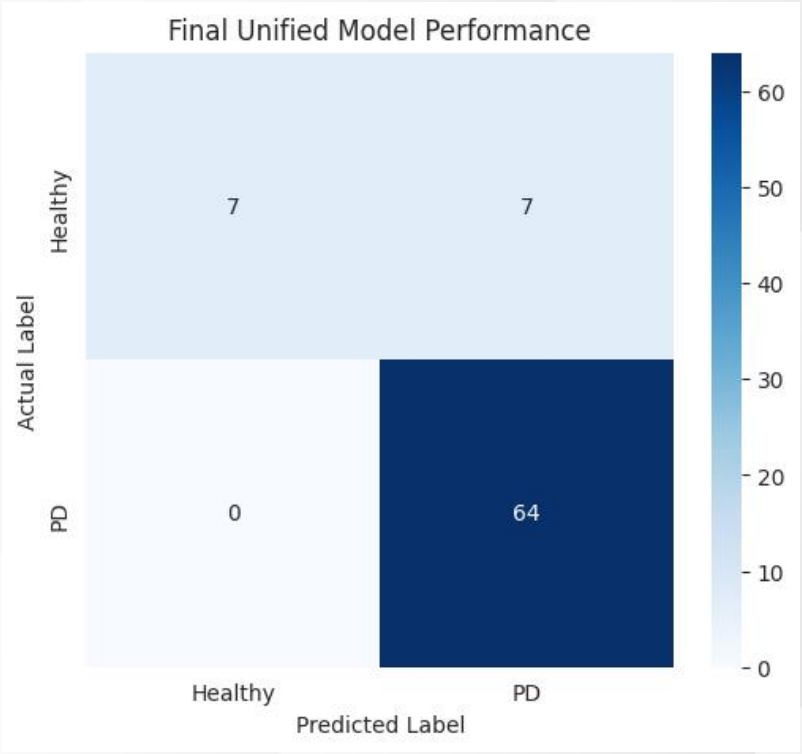
- **Accuracy:** 0.91
- **Weighted F1-score:** 0.90
- **Macro Recall:** 0.75

Confusion Matrix Insight

- PD cases: **64/64 correctly detected**
- Healthy cases: **7 correct, 7 false positives**
- The model prioritizes **catching all PD patients**, which is desirable in screening.

Key Conclusion: The final model is *domain invariant*. It accurately identifies Parkinson's across both clinical recordings and real-world home audio, achieving strong, balanced performance after unifying both domains.

Class	Precision	Recall	F1-score	Support
Healthy	1.00	0.50	0.67	14
PD	0.90	1.00	0.95	64



Phase 5: Unified Cross-Dataset Model with SMOTE

- Even in the unified hybrid model, **class imbalance** remains:
- Training set: 38 Healthy vs 274 PD (≈1 : 7.2)
- The model focuses heavily on PD and misses some Healthy subjects.
- Goal of Phase 5:
- Improve **Healthy detection**
- Make performance more **balanced across classes**
- Keep the model **domain invariant** across clinical + home recordings.

How the SMOTE-Based Unified Model Works?

- Start from the **50/50 domain-balanced** dataset from Phase 4 (equal UCI and Telemonitoring samples).
- Split into **train/test (80/20)** as before.
- On the **training set only**:
  - Apply **SMOTE** to oversample the minority class (Healthy).
  - Before SMOTE: 38 Healthy, 274 PD
  - After SMOTE: 137 Healthy (99 synthetic), 274 PD → ratio ≈ 1 : 2
- Scale features with **StandardScaler** (fit on the SMOTE-balanced training data).
- Train **Random Forest** with `class_weight='balanced'` on this balanced set.
- Evaluate on the **original mixed test set** (no SMOTE there).

"SMOTE teaches the model what 'Healthy' looks like without changing the real test data

Results: Unified Model with SMOTE (Phase 5)

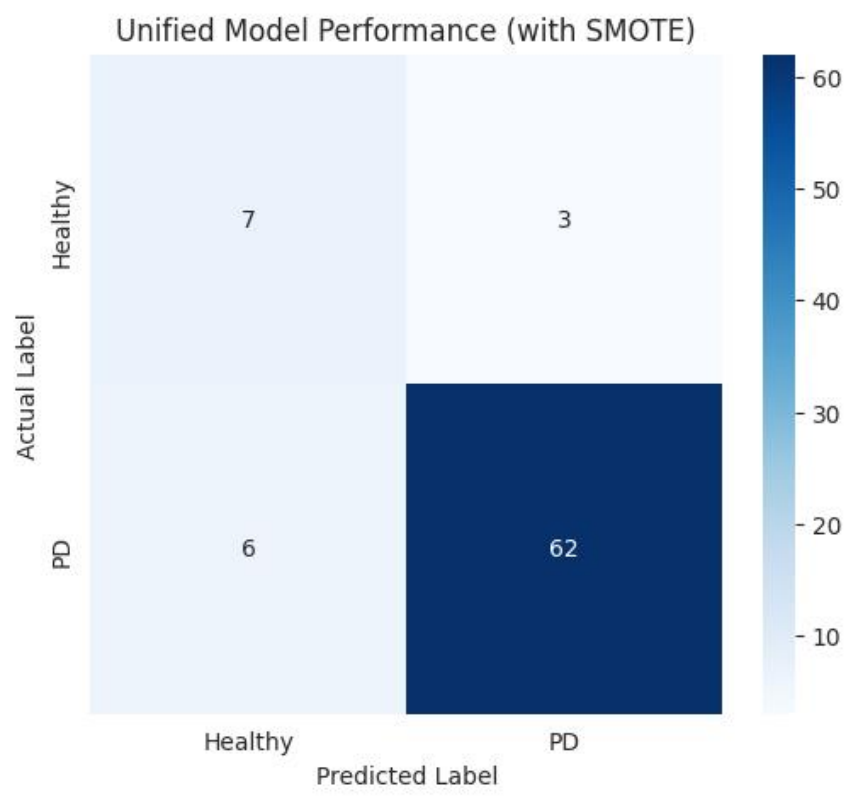
Overall Performance

- **Accuracy:** 0.88
- **Macro Recall:** 0.81
- **Macro F1:** 0.77

Class	Precision	Recall	F1-score	Support
Healthy	0.54	0.70	0.61	10
PD	0.95	0.91	0.93	68

- Healthy recall improved from 50% → 70% after SMOTE.
- PD performance remains high (Recall ≈ 91%).
- The model is now more balanced across both classes.
- SMOTE partially overcomes the extreme 8:1 class imbalance in real data,
  but you still note that more real Healthy data would improve things further.

"Phase 5 shows that combining domain balancing + SMOTE produces a unified Parkinson's detector that generalizes across datasets and treats both classes more fairly."



Conclusion

- We demonstrated that **voice-based biomarkers** can reliably detect Parkinson's, achieving **93% accuracy and F1 up to 0.96** in the UCI dataset.
- **Regression baselines** revealed that predicting UPDRS severity is significantly harder, showing **higher MAE and negative R^2**, motivating deeper modeling.
- **Cross-dataset testing (UCI → Telemonitoring)** showed strong generalization (**88% recall**), confirming the model learns *real* PD vocal patterns—not dataset noise.
- **Reverse transfer (Telemonitoring → UCI)** performed poorly, and t-SNE revealed clear **domain shift** between clinical and home environments.
- Our **Hybrid Unified Model** (balanced UCI + Telemonitoring) achieved **91% accuracy** and became **domain-invariant**, detecting PD across environments.
- **SMOTE-based balancing** further improved Healthy-class recall (50% → 70%), reducing class imbalance impact.

- **Overall:**
  We built an end-to-end PD detection and progression framework, showed where models succeed and fail, explained model behavior with SHAP, and produced a robust cross-dataset Parkinson's detector that generalizes across recording conditions.

Thank You