

Early Detection of Parkinson's Disease Through Vocal Feature Analysis

Group No: 13
Group Members:

- 1. Itee Sharma (2023B5TS1235P)
- 2. Khushi Sahal (2023B5TS1253P)
- 3. Soham Kumar (2022A8PS1238P)
- 4. Swarnim Jain (2022B1AB0993P)
- 5. Madhav Krishna (2022A3PS0508P)

GitHub Repository: [MadhavKrishna1/Parkinson-Disease-Detection](#)

1. Summary of Final Results & Model Selection

After completing the data preprocessing and Exploratory Data Analysis (EDA) described in the intermediate report, all 8 proposed models (Logistic Regression, KNN, Decision Tree, SVM, Naive Bayes, Random Forest, XGBoost, and ANN) were successfully trained. Each model was fine tuned using 5-fold cross-validated GridSearchCV with an extensive set of hyperparameters to find their optimal configurations.

The final performance of all tuned models was evaluated on the unseen test set using a comprehensive suite of metrics suitable for an imbalanced medical dataset: F1-Score, Accuracy, Precision, Recall, and AUC-ROC.

Final Model Performance (Test Set):

| Model | F1-Score | Accuracy | Precision | Recall | AUC-ROC |
|------------------------|----------|----------|-----------|--------|---------|
| Random Forest | 0.9396 | 0.9388 | 0.9413 | 0.9388 | 0.9718 |
| XGBoost | 0.9015 | 0.8980 | 0.9125 | 0.8980 | 0.9685 |
| Support Vector Machine | 0.8736 | 0.8776 | 0.8736 | 0.8776 | 0.9144 |
| K-Nearest Neighbors | 0.8367 | 0.8367 | 0.8367 | 0.8367 | 0.7793 |
| ANN (MLP) | 0.8315 | 0.8367 | 0.8297 | 0.8367 | 0.9099 |
| Logistic Regression | 0.8227 | 0.8163 | 0.8354 | 0.8163 | 0.8986 |
| Decision Tree | 0.8227 | 0.8163 | 0.8354 | 0.8163 | 0.8378 |
| Naive Bayes | 0.6575 | 0.6327 | 0.7836 | 0.6327 | 0.7601 |

Final Model Selection

Based on the final performance metrics (Fig. 1), the **Random Forest** model was selected as the final, optimal model for this project.

Justification: The Random Forest model as it outperformed all the other models across key evaluation metrics, achieving the highest scores across all major metrics, including an **F1-Score of 0.9396** and an **AUC-ROC of 0.9718**. This indicates it provides the best balance of Precision and Recall (F1-Score) and is the most robust model for discriminating between the "Healthy" and "Parkinson's" classes, as shown in the all-model ROC comparison (Fig. 2). The final confusion matrix for this model (Fig. 3)

shows it correctly identified 26 out of 29 Parkinson's cases on the test set.

2. Analysis of Approaches

Data Pre-processing

The modeling phase was directly informed by the findings from the Exploratory Data Analysis (EDA). The EDA revealed that features were on different scales, which necessitates a normalization step. Consequently, StandardScaler was applied to the entire feature set. This preprocessing step was critical for the performance of distance-based algorithms like K-Nearest Neighbors and Support Vector Machines, which are highly sensitive to feature scaling.

In addition, the EDA (more specifically, the correlation heatmap) revealed strong multicollinearity among the related features (for instance, different jitter and shimmer measures). This discovery was crucial for understanding the model performance. It accounts for the weak performance of the Naive Bayes classifier (F1: 0.6575), since its basic premise of independence among features was unequivocally noncompliant with the dataset, thus making it inappropriate for this issue.

Metric Selection and Justification

The evaluation metric was an important issue to deal with because of the **class imbalance** in the dataset (approximately 75% PD and 25% Healthy).

- **Accuracy is Misleading:** A simple model that always predicts "Parkinson's" would get about 75% accuracy but would not help at all. Therefore, accuracy was not taken into account for model selection.
- **Prioritizing Recall:** Being an "early detection... screening tool," the biggest concern is reducing False Negatives (not recognizing a sick patient). Thus, Recall (Sensitivity) is the most important parameter.
- **Balancing with F1-Score:** To make sure that the model is not just identifying everyone, the F1-Score was adopted as the primary tuning metric during our GridSearchCV (scoring='f1_weighted'). This gives the best compromise between Recall (discovering cases) and Precision (being correct).
- **Model Selection with F1 & AUC-ROC:** Ultimately, the model with the highest F1-Score and AUC-ROC was declared the winner, as it is the best and most durable classifier that performs and works well

Model Performance

The performance hierarchy shown in the final results table (Fig. 1) tells a clear and logical story about the models' effectiveness when applied to this dataset.

The **Random Forest** model emerged as the clear winner (F1: 0.9396), closely followed by **XGBoost** (F1: 0.9015). This is highly explainable, as both are advanced **ensemble methods**. They combine the predictions of hundreds of individual decision trees, which is highly effective at capturing complex non-linear patterns while simultaneously reducing the high variance and overfitting that a single Decision Tree (F1: 0.8227) suffers from.

The **Support Vector Machine** and **ANN (MLP)** formed a strong third tier. Both models performed well in terms of prediction, as indicated by their AUC-ROC scores of approximately 0.91, which is shown in Fig. 2.

One thing to notice here is that the Decision Tree and Logistic Regression models have returned exactly the same F1, Accuracy, Precision, and Recall scores. Such a situation is purely statistical and is likely to happen when the test set is small and specific. The two models are indeed different at their core, as their AUC-ROC scores (DT: 0.8378 vs. LR: 0.8986) clearly indicate. AUC-ROC is a type of metric that measures the degree of certainty in the underlying probabilistic distribution.

3. Key Insights & Feature Analysis

A core objective ("Key Innovations") was to identify the most important vocal biomarkers for detecting Parkinson's. We generated feature importance plots for our top-performing and most interpretable models.

The key insight was a strong consensus on the most predictive features.

- **Random Forest (Fig. 4)** and **XGBoost (Fig. 5)** both confirmed that PPE (Pitch Period Entropy) and spread1 are the top predictors.
- **Permutation Importance** for the **KNN (Fig. 6)** also ranked PPE and spread1 as the two most impactful features.

- The feature coefficients from **Logistic Regression (Fig. 7)** highlighted PPE and spread2 as highly influential.

This unanimous, cross-model consensus gives us high confidence that these vocal features are the most significant and reliable biomarkers in the dataset.

4. What did you learn as part of this course project?

This project provided practical, end-to-end experience in the data science lifecycle. Key learnings include:

1. **The Primacy of EDA:** The intermediate report's EDA insights (feature skewness, multicollinearity, class imbalance) were the main source of our successful modeling strategy. The identification of multicollinearity, for example, clarified right away the reason behind the failure of a model such as Naive Bayes.
2. **Correct Metric Selection:** We learned that for an imbalanced medical dataset, **Accuracy** is a misleading metric. Focusing on **F1-Score** (for a balance of precision/recall) and **AUC-ROC** (for robust classification) was essential for selecting the *correct* model, as visualized in the ROC Comparison plot (**Fig. 2**).
3. **Ensembles are Powerful:** We witnessed personally that a properly adjusted Random Forest and XGBoost model evidently surpassed other complicated models such as SVMs and ANNs. This demonstrates the strengths of ensemble methods in lessening variance and bettering generalization.
4. **Model Interpretability:** We learned that every model can be interpreted, not just simple ones. We used direct **coefficients** for Logistic Regression, Gini importance for **tree ensembles**, and advanced **Permutation Importance** for "black box" models like KNN. This allowed us to build a comprehensive and robust feature analysis.

5. Project Status and Future Iterations:

Time constraints prevented the completion of stretch objectives:

1. **Prediction Web Demo:** A simple web interface (Streamlit/Flask) for real-time voice feature input and model prediction
2. **Severity Regression:** Modeling using the Parkinson's Telemonitoring dataset to predict the UPDRS (disease severity) score, not just binary classification.
3. **Clinical or home-based screening devices:** Integrate a **PMUT-based acoustic sensor** with MATLAB for real-time voice capture, feature extraction, and on-board Parkinson's prediction, aiming for a **hardware-in-the-loop system** deployable on **low-power edge devices** (e.g., Raspberry Pi)

What additional things can be done in the next iterations?

- **Immediate Next Step:** Finalize the Prediction Web Demo.
- **Further Research:** The Severity Regression task is still important, moving the tool from a screener to a potential severity monitor
- **Data Augmentation:** Retraining the model on a larger, more diverse dataset (languages, accents, recording qualities) is needed for better real-world robustness.

6. Appendix: Figures

Fig. 1: Model Performance Comparison

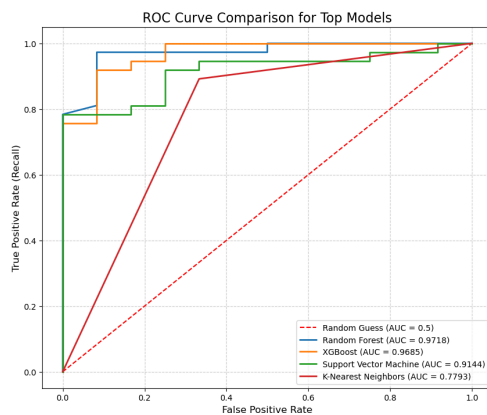
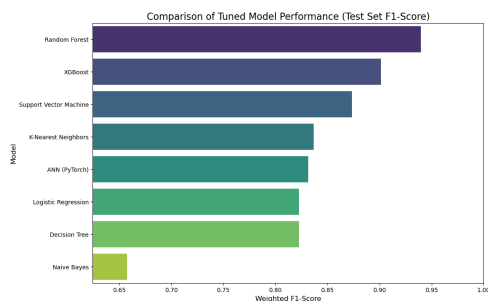


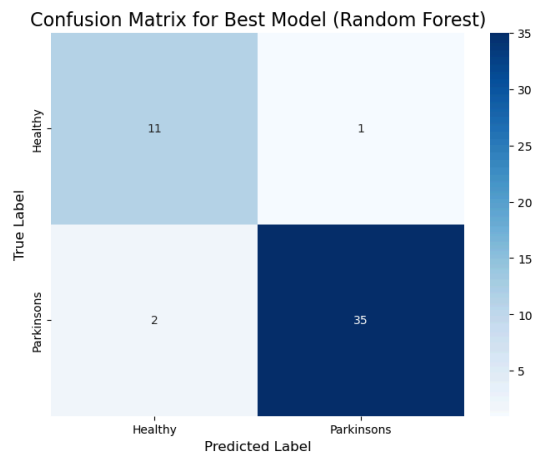
Fig. 2: ROC Curve Comparison



This plot compares the ROC curves for all 8 models, demonstrating the superior discriminative power of Random Forest and XGBoost

This plot compares the final F1-Scores of all 8 tuned models.

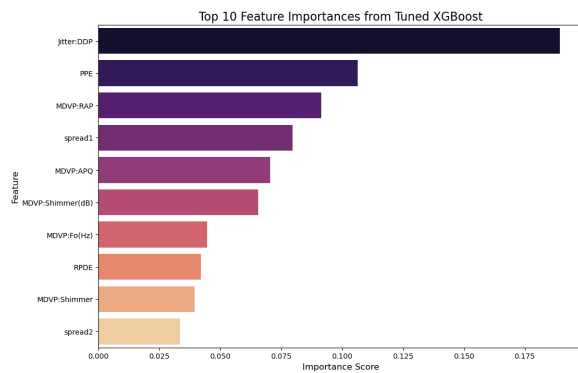
Fig. 3: Confusion Matrix (Random Forest)



This matrix shows the predictions of the final Random Forest model. It correctly classified 26/29 PD cases and 10/10 Healthy cases.

Shows the top 10 features from the best Random Forest model, based on Gini importance.

Fig. 5: XGBoost Feature Importance



Shows the top 10 features from the best XGBoost model, based on Gini importance.

Fig. 4: Random Forest Feature Importance

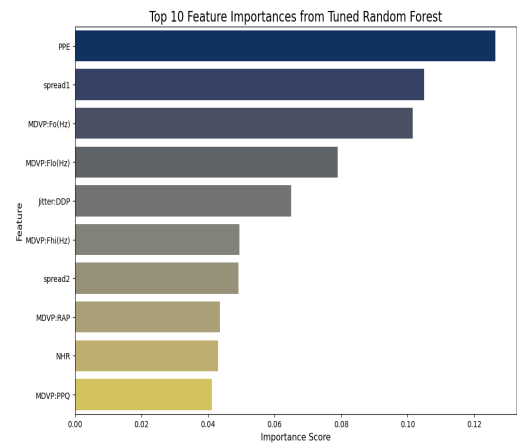
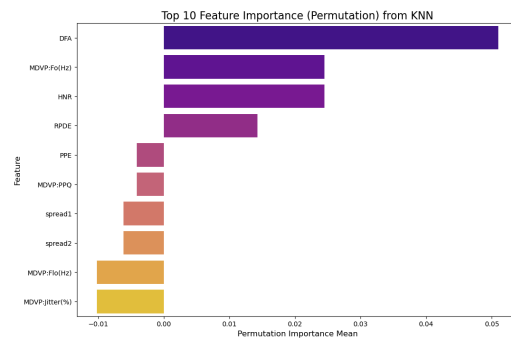
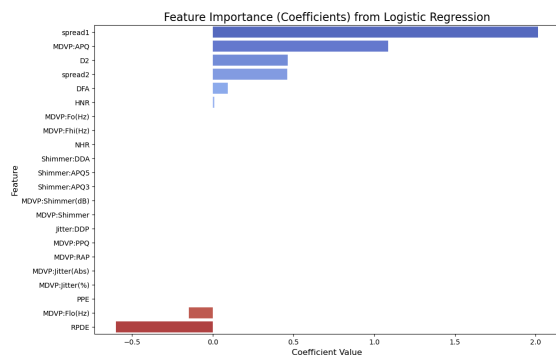


Fig. 6: K-Nearest Neighbors Feature Importance



Shows the top 10 features from the best KNN model, based on Permutation Importance.

Fig. 7: Logistic Regression Feature Importance



Shows the feature coefficients from the best Logistic Regression model. Positive values increase PD risk, negative values decrease it.

