Finding biological markers for Parkinson's disease

Using Machine Learning to analyze Shotgun Metagenomic Sequencing data from the gut microbiome

1. Introduction

Several studies have already found significant differences between the **metagenomic data** of Parkinson's patients compared to healthy controls [1],[2],[3]. But not many studies have used **Machine Learning** for **biomarker discovery.**

2. Research Question

Can machine learning models be used to discover/verify biological markers for Parkinson's disease based on gut metagenomic data?

4. Feature selection

Three **feature selection techniques** used:

- Recursive Feature Elimination
- Mean Decrease Accuracy
- Minimum Redundancy Maximum Relevance

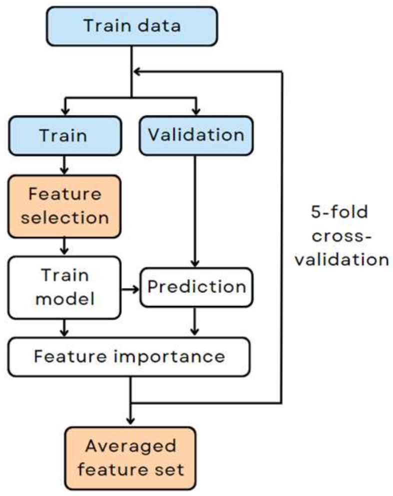
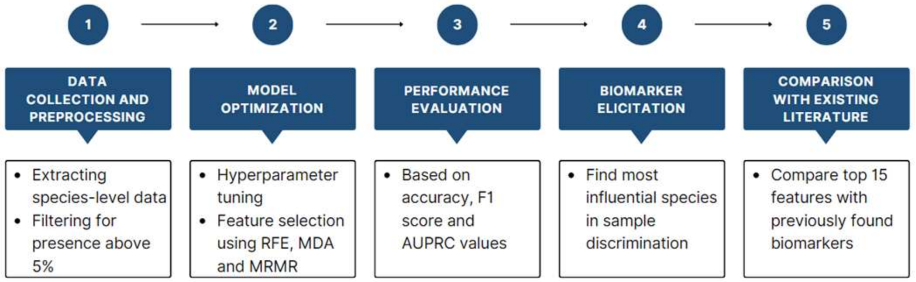


FIGURE 1 ▲
The **5-fold cross-validation** approach for conducting feature selection.

3. Methodology

The three machine learning models used are **Logistic Regression (LR), Random Forest (RF)** and **Support Vector Machines (SVM).**



1. DATA COLLECTION AND PREPROCESSING
   - Extracting species-level data
   - Filtering for presence above 5%
2. MODEL OPTIMIZATION
   - Hyperparameter tuning
   - Feature selection using RFE, MDA and MRMR
3. PERFORMANCE EVALUATION
   - Based on accuracy, F1 score and AUPRC values
4. BIOMARKER ELICITATION
   - Find most influential species in sample discrimination
5. COMPARISON WITH EXISTING LITERATURE
   - Compare top 15 features with previously found biomarkers

FIGURE 2 ▲ The five-step methodology of the research.

5. Results

Highest classifier performances were achieved without feature selection on RF, but with MRMR feature selection on LR and SVM.

Optimized classifiers show **moderate performances** as illustrated in Figure 3. Although the RF model exhibited the best performance among all classifiers, it displayed a tendency to **overestimate PD cases.**

However, a comparative analysis of the top features indicates a **significant overlap** between classifiers and with previously found biomarkers in existing literature (Figure 4).

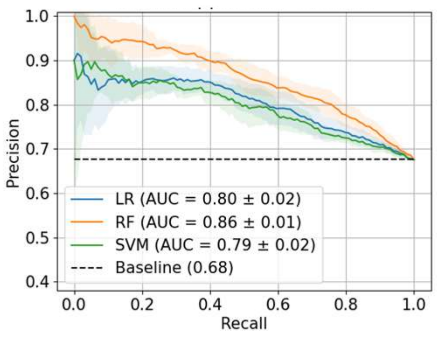


FIGURE 3 ▲ Precision-Recall curve illustrating moderate performance compared to baseline (AUPRC 0.68).

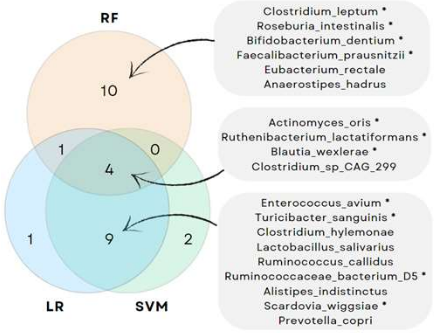


FIGURE 4 ▲ Overlapping top features between classifiers and existing literature.

A **confounding analysis** on a small subset of the data shows a decrease in model performances and biomarker identification lacks confirmation from existing literature.

6. Limitations

Limitations of ML approaches for PD biomarker discovery:

- Reliance on input data
- **Overfitting** and **biases** (Figure 5)
- **Interpretability** issues and the need for validation of the results
- **Misclassification** of metagenomic data due to diagnosis inaccuracies

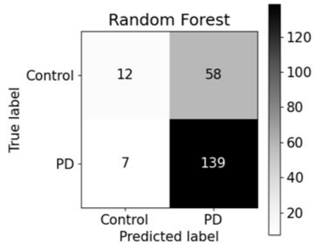


FIGURE 5 ▶
Confusion matrix for the RF model illustrates bias towards PD-overestimation due to overfitting.

7. Future work

Recommendations for further research include:

- Large-scale clinical trial with **postmortem neuropathological disease validation.**
- Analysis using a balanced and generalizable dataset.
- Analyzing improvement when **including current diagnostic measures,** such as motor symptoms.
- **Review** the usefulness of all available ML models for metagenomic analysis.

8. Conclusion

Despite achieving moderate performance, LR, RF, and SVM classifiers provided **compelling evidence** of their capability to identify PD biomarkers.

The findings of this research contribute to the understanding of ML approaches for biomarker discovery in PD and highlight areas for further investigation.

References

[1] Wallen, Z.D., Demirkan, A., Twa, G. et al. Metagenomics of Parkinson's disease implicates the gut microbiome in multiple disease mechanisms. Nat Commun 13, 6958 (2022).

[2] Bedarf, J. R., Hildebrand, F., Coelho, L. P., Sunagawa, S., Bahram, M., Goeser, F., Bork, P., and W ̈ullner, U.v(2017). Functional implications of microbial and viral gut metagenome changes in early stage l-dopa-na ̈ıve parkinson's disease patients. Genome Medicine, 9(1):39.

[3] Mao, L., et al. (2021). "Cross-Sectional Study on the Gut Microbiome of Parkinson's Disease Patients in Central China." Frontiers in Microbiology 12.

Author: **Marilotte Koning** m.l.koning-1@student.tudelft.nl Responsible Professor: **Thomas Abeel** Supervisors: **Eric van der Toorn, David Calderon Franco**