Finding biological markers for Parkinson's disease

Using machine learning to analyse metagenomic data

Marilotte Koning[1]

Supervisors: Thomas Abeel[1]**, Eric van der Toorn**[1]**, David Calderon Franco**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Marilotte Koning
Final project course: CSE3000 Research Project
Thesis committee: Thomas Abeel, Eric van der Toorn, David Calderon Franco, Thomas Höllt

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder characterized by motor function loss and potential mental and behavioral changes. The identification of biomarkers in the gut microbiota of PD patients can significantly aid in fast and accurate diagnosis. This study investigates the application of machine learning (ML) models, including Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM), to discover biomarkers in the gut metagenomic data of PD patients. The ML models were optimized using various feature selection techniques, and a comparative analysis of the most influential species in sample discrimination was conducted to verify potential PD-associated biomarkers. The results demonstrate that all three ML models exhibit moderate performance, indicating their limited discriminatory power. However, the comparison of significant species across different classifiers demonstrates substantial overlap and indicates PD-associated species that align with existing literature findings. These outcomes provide promising evidence that LR, RF, and SVM classifiers can effectively identify biomarkers for PD. However, confounding analysis on a small subset of the dataset failed to identify meaningful PD-associated species. Therefore, caution is advised when interpreting the findings of ML model, considering factors such as classifier performance, dataset limitations, potential biases, influence of feature selection methods, and inherent model differences. We validate the potential usefulness of ML approaches for biomarker discovery and highlight areas for further investigation into building a sufficiently accurate ML model for PD diagnosis.

1 Introduction

The human microbiome is a vast collection of genes which, due to its unique composition, can be viewed as our 'other genome', and has recently gained interest for the identification of biomarkers for human diseases (Hajjo et al., 2022). Biomedical research has investigated the relationship between changes in the gut flora and changes in host immunity (Thursby and Juge, 2017). These studies observed a divergence from the normal microbiome composition when the host is affected by diseases ranging from chronic gastrointestinal diseases to neurodevelopmental disorders.

One of the diseases that has been of interest for biomarker discovery is Parkinson's disease (PD) (Wallen et al., 2022; Bedarf et al., 2017; Qian et al., 2020; Mao et al., 2021; Boktor et al., 2023). PD is a neurodegenerative disorder that affects the central nervous system, leading to motor function loss and potential mental and behavioral changes. However, diagnosing PD, especially in its early stages, is challenging due to the inaccuracies associated with clinical diagnosis, confirmed through postmortem neuropathological assessment (Adler et al., 2021). Adler et al. (2021) explain that factors such as disease duration, responsiveness to dopaminergic medication, and the presence of PD-associated motor symptoms have shown some impact on diagnostic accuracy, but it remains a significant challenge. In this context, the discovery of biomarkers in the gut microbiota of PD patients holds great potential for facilitating fast and accurate diagnosis.

Previous studies have already verified a significant difference in the gut microbiota of patients with PD compared to healthy controls (Wallen et al., 2022; Mao et al., 2021; Bedarf et al., 2017; Qian et al., 2020; Boktor et al., 2023). Several of these studies have already utilized the recently introduced shotgun metagenomic sequencing approach, which enables taxonomical profiling of all microbial genomes within a sample (Quince et al., 2017). Using this technique, Wallen et al. (2022) found that both at genus and species level the dysbiosis in the PD gut microbiome appeared to involve about 30% of the tested taxa. After analyzing the possible effect of other factors such as alcohol or laxative usage, Wallen et al. (2022) also confirmed that 32 species were exclusively associated with PD. These findings provide compelling evidence of a significant distinction between samples of PD and control subjects. However, it is important to note that the methods used for identifying these relevant species continue to vary within the field.

This research involves a comparison of multiple Machine Learning (ML) models that have been used in the process of PD biomarker discovery before, namely Logistic Regression (LR) by Bedarf et al. (2017), Random Forest (RF) by Mao et al. (2021) and Support Vector Machines (SVM) by Qian et al. (2020). However, none of these studies have obtained their results based on these ML models, but solely use their performances to corroborate their findings.

The main objective of this research is to evaluate the usefulness of ML methods in discovering biomarkers for PD. This is done by determining and comparing the effectiveness of LR, RF and SVM in classifying PD patients based on the metagenomic profiles of their gut samples. Additionally, this research aims to identify the most influential species in sample discrimination, which can potentially serve as biomarkers for PD. By corroborating the findings with existing literature, we intend to validate the usefulness of ML approaches for biomarker discovery.

2 Materials and methods

2.1 Metagenomic data collection, availability and preprocessing

The dataset used is from a previous study conducted by Wallen et al. (2022). The data is derived from shotgun metagenomic sequencing samples and has been through quality control and taxonomic profiling, to quantify the presence of various species within the samples. This dataset comprises samples obtained from 490 individuals diagnosed with PD and 234 neurologically healthy control subjects. Although slightly unbalanced, the substantial size of the study provides an advantage over smaller datasets. The study is accompanied by extensive subject metadata encompassing various factors such as age, sex, lifestyle, presence of other diseases, medication usage, and more. Importantly, all the data, including the subject metadata, is publicly available without any restrictions and can be accessed at Zenodo [https://zenodo.org/record/7246185].

Before conducting any analysis, the dataset underwent preprocessing steps. Initially, it contained relative abundances for all taxonomical levels, but only species-level data was extracted for further analysis. The dataset contained five swab samples that have been removed. Subsequently, a filtering process was applied to retain species present in over 5% of the samples, resulting in a final set of 259 species. This filtering approach aligns with the methodology used in the study by Wallen et al. (2022), from which the data was sourced. It was assumed that species present in less than 5% of the samples may not have significant importance. Notably, the removed species accounted for an average abundance of only 2.1% per sample, indicating their relatively minor contribution to the overall dataset.

2.2 Machine learning models used for classification

The ML methods Logistic Regression (LR) (Yu et al., 2011), Random Forest (RF) (Breiman, 2001) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995) have been used for classification. These classifiers were based on the *scikit-learn* Python library implementation (version 1.2.2) (Pedregosa et al., 2011), with initial default settings utilized for LR and RF. However, for the SVM classifier, the kernel parameter was specifically set to employ a linear kernel. This choice was made to ensure the availability of feature importance scores needed for species ranking, which non-linear SVMs do not straightforwardly provide.

Classifier	Hyperparameters
LR	C: 0.0, 0.1, 0.2, ..., 10.0 Penalty: l1, l2, elasticnet, none Solver: newton-cg, lbfgs, liblinear, sag, saga
RF	n_estimators: 0, 5, 10, ..., 500 max_depth: 0, 5, 10, ..., 500
SVM	C: 0.0, 0.1, 0.2, ..., 10.0 Kernel: linear

Table 1: The grid values used for hyperparameter tuning of the LR, RF and SVM models.

Hyperparameter tuning was conducted on all classifiers before making predictions, in order to optimize their performance. Again, the *scikit-learn* Python library was used for this purpose, employing the *RandomizedSearchCV* object. Table 1 presents the grid values used for hyperparameter tuning. The tuning was based on the ”f1 macro” metric, aiming to strike a balance between correctly identifying positive instances (recall) and minimizing false positives (precision). Stratified 5-fold cross-validation was utilized within the *RandomizedSearchCV* object by passing a *StratifiedKFold* object from the same library, with the shuffle attribute set to ”True”. To ensure result reproducibility, both the *RandomizedSearchCV* and *StratifiedKFold* objects were assigned a random state of 42.

2.3 Feature selection

Feature selection techniques were applied to enhance classifier performance. Three specific methods were employed: Recursive Feature Elimination (RFE) (Guyon et al., 2002), Mean Decrease Accuracy (MDA) (Han et al., 2016), and Minimum Redundancy Maximum Relevance (MRMR) (Ding and Peng, 2003). The choice of these techniques was based on previous usage within the PD biomarker research field (Huang et al., 2023; Mao et al., 2021; Qian et al., 2020).

RFE and MDA were performed using the *scikit-learn* Python library implementation (version 1.2.2) (Pedregosa et al., 2011). RFE was performed using the built-in *RFE* object. For MDA the *permutation_importance* object was used, with a fixed random state of 42. MRMR was performed using the *mrmr_selection* Python library (version 0.2.7) (Mazzanti, 2021). All three methods were set to maintain only the 50 most relevant features, which had been found to result in the best performance by manual inspection.

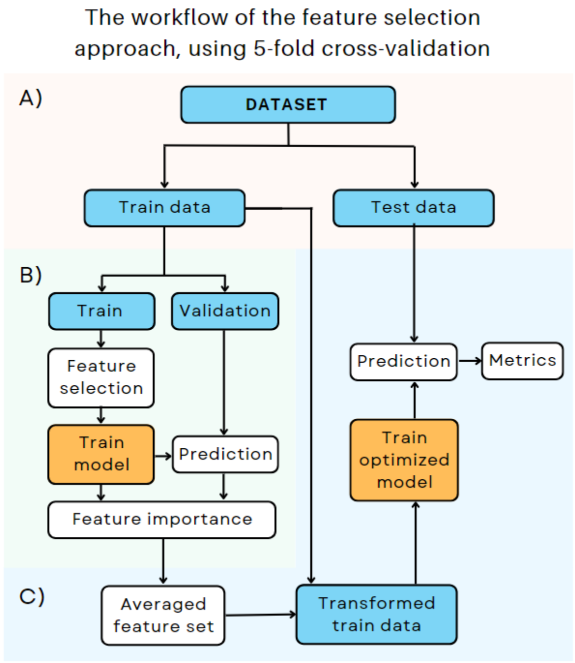

Figure 1: The workflow of the feature selection approach. (A) Initially, the dataset is divided into a training set and a test set with a ratio of 70/30. (B) Subsequently, 5-fold cross-validation is conducted on the training set, incorporating feature selection in each fold to obtain prediction outcomes and feature importance scores. (C) Lastly, the final feature set is constructed and the performance of the classifiers is evaluated on the holdout test set.

Figure 1 depicts the procedure for feature selection. Initially, the dataset is split into a 70% training set and a 30% test set. Subsequently, 5-fold cross-validation is conducted on the training set using the *StratifiedKFold* object from the *scikit-learn* Python library, with a random state set to 42. Within each fold, a portion of the data is separated to serve as a validation set. Feature selection is then applied using the training data. The selected features are then used to train the classifier, after which feature importance scores are extracted using either the *coef_* or *feature_importance_* attributes of the respective model. These scores indicate the relevance of the feature during classification and have subsequently been adjusted by multiplying them by the accuracy score obtained from predicting the validation set. By following this approach we aim to tune the features based on the accuracy of the model. The final feature set is constructed by averaging the scores across all folds, assigning zero importance to features not selected in a particular fold, and then selecting the top 50 features. This final feature set is then used to train the optimized model. Finally, the performance of the classifier is evaluated by predicting the holdout test set. In both the cross-validation and final evaluation stages, hyperparameter tuning is performed before training the model in order to optimize its performance. This entire procedure is employed for each individual classifier.

2.4 Performance evaluation

Bootstrapping was employed to obtain a robust evaluation of the overall performances of the ML models. This technique involved averaging the model performances over ten runs using different random states for the train-test splits, as depicted in step A of Figure 1. The train-test splits were conducted using the *train_test_split* object from the *scikit-learn* Python library. The random states ranged from 0 to 100 with increments of ten. This approach helps capture the overall performance and provides insights into the stability of the models across different random splits.

The impact of feature selection on performance was assessed using three key metrics: accuracy, F1 score, and the area under the precision-recall curve (AUPRC). These metrics collectively provide a comprehensive evaluation of the model's performance. Accuracy measures the overall correctness of predictions, while F1 score aims to minimize both false positives and false negatives. Additionally, AUPRC summarizes the classifier's overall ability to identify positive instances while maintaining a high precision, which is particularly valuable in imbalanced datasets.

The accuracy and F1 score were calculated using the *accuracy_score* and *f1_score* functions available in the *scikit-learn* Python library. The formulas for accuracy and F1 score are as follows:

$$\textbf{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\textbf{F1 Score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

Where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

The final evaluation of the classifier's performance is presented as the precision-recall (PR) curve. The PR curves and their corresponding AUPRC value have been calculated using the *precision_recall_curve* and *auc* functions available in the *scikit-learn* Python library.

Statistical tests were used to validate the significance of relative performances. Paired t-tests were used since the tests were conducted on the same train-test splits resulting in a sample dependence. A total of twelve tests were performed, comparing the performances of LR, RF, and SVM with different feature selection methods and between classifiers. The p-values were adjusted for multiple testing using the Bonferroni method. This statistical approach validated the effectiveness of feature selection techniques in enhancing classifier performances. In some cases, when ten runs did not provide a clear significance, the bootstrapping range was increased to 300, resulting in thirty runs to reduce variability.

Additionally, Mann-Whitney U tests were performed to compare variability within the data of the PD and control groups. Again, the p-values were adjusted for multiple testing using the Bonferroni method. The use of Mann-Whitney U tests allowed for a comprehensive analysis of group variability and helped identify significant differences between the PD and control groups. Both the paired t-tests and the Mann-Whitney U tests have been performed using the *SciPy* Python library (version 1.10.1) (Virtanen et al., 2020).

The figures within this paper illustrating classifier performances and important findings were generated using either the *Matplotlib* Python library (version 3.7.1) (Hunter, 2007) or the *Seaborn* Python library (version 0.12.2) (Waskom, 2021).

2.5 Comparison of important species

The possible PD-associated species have been elicited by examining the most influential features in sample discrimination. This has been done by extracting feature importance scores using the *coef_* and *feature_importance_* attributes of the respective models. These scores indicate the relevance of the feature during classification. We decided to compare the results of multiple runs to obtain a more reliable result, accounting for the observed variability between runs. Ten runs of the optimized models were conducted, using random states ranging from 0 to 100 with increments of ten. The average feature importance has been used to rank the species. For models employing feature selection, we verified the reliability of the species by assessing their consistency in the selected feature sets. In models without feature selection, we checked for outliers by examining the standard deviation of the scores. We compared the top 15 ranked species from all classifiers and compared the results with existing literature.

3 Results and discussion

3.1 T-SNE data visualization indicates substantial similarity between the PD and control groups

To gain insights into the distribution and separability of the data, t-Distributed Stochastic Neighbor Embedding (t-SNE) (Hinton and Roweis, 2002) was utilized as a visualization technique. In the t-SNE plot displayed in Figure 2, it is observed that the points related to PD patients and controls show an almost complete overlap. No apparent patterns or distinct groups could be observed, indicating a lack of clear separation between the two groups.

To assess the extent of overlap, the locations of the PD and control groups in the two-dimensional space were analyzed. The mean and standard deviation of their locations are summarized in Table 2. Statistical p-values were calculated using

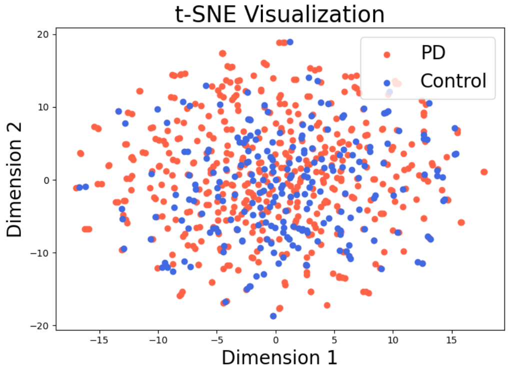

Figure 2: The two-dimensional t-SNE visualizing major overlap between the PD and control groups. No apparent patterns or distinct groups could be observed, indicating a lack of clear separation between the two groups.

a Mann-Whitney U test to determine the significant similarity between the dimensions of the two groups. Both p-values were found to be below the 0.05 significance threshold, indicating a substantial similarity. This finding suggests a limited variability between the PD and control groups, making it challenging to differentiate the two groups based solely on the current dataset.

	PD		Control		
Dimension	Mean	STD	Mean	STD	p-values
Dimension 1	-0.42	7.35	0.77	6.49	0.036
Dimension 2	0.74	7.93	-1.05	6.94	0.0044

Table 2: Mean and standard deviation (STD) of the t-SNE dimensionality reduction for PD and control groups. The p-values indicate a significant similarity between the locations of the two groups, suggesting a lack of clear separation. However, greater standard deviation associated with the PD point suggests more variability within this group.

However, Table 2 does show a slightly greater standard deviation associated with the PD points, observed as a slight dispersion among the points associated with PD patients in Figure 2. This dispersion suggests some variability within the PD group, potentially indicating the presence of subgroups or variations within the PD population. Further investigation is necessary to explore the underlying factors contributing to this dispersion and to assess their potential implications for disease heterogeneity or progression.

3.2 Performance evaluation of optimized LR, RF, and SVM classifiers suggest limited discriminatory power and bias due to dataset imbalance

We conducted an evaluation of three ML models, namely Logistic Regression (LR) (Yu et al., 2011), Random Forest (RF) (Breiman, 2001), and Support Vector Machine (SVM) (Cortes and Vapnik, 1995). Our objective was to investigate the effectiveness of these ML models in accurately classifying PD patients based on their metagenomic samples and potentially uncovering important PD biomarkers through feature importance analysis during the classification process. Upon optimizing all classifiers using various feature selection techniques, our findings indicate that the discriminatory power between PD and control cases is limited. Although the RF model exhibited the best performance among all classifiers, it displayed a tendency to overestimate PD cases, which could potentially be attributed to the imbalance of the dataset.

To optimize the performance of the classifiers, three feature selection methods were employed: Recursive Feature Selection (RFE) (Guyon et al., 2002), Mean Decrease Accuracy (MDA) (Han et al., 2016) and Minimum Redundancy Maximum Relevance (MRMR) (Ding and Peng, 2003). We compared the performance of the classifiers with and without these feature selection techniques using accuracy, F1 score, and Area Under the Precision-Recall Curve (AUPRC) values. The results, summarized in Table 3, show that RF did not benefit from any feature selection method, while LR and SVM exhibited improved performances, particularly with MRMR, although the differences were small. This indicates that MRMR effectively selected informative features relevant for classification.

To assess the significance of these improvements, we conducted Bonferroni-corrected paired t-tests, with the observations indicated in Table 3. It is important to interpret these improvements with caution, as increased test size was necessary to obtain clear evidence of significance. The results suggest that MRMR did not have a major impact on the performances, but some metrics showed significance. Therefore, we decided to continue the biomarker identification process with MRMR feature selection applied to both the LR and SVM classifiers, while no feature selection technique was selected for RF.

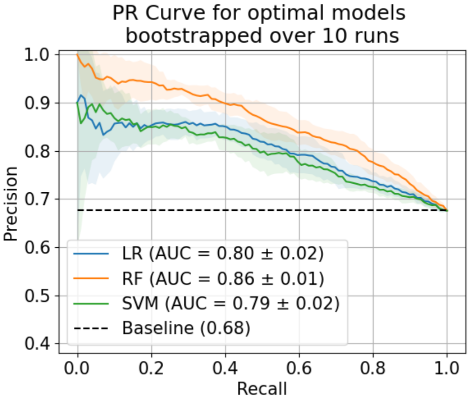

Figure 3: The PR curves illustrate the performance of LR, RF and SVM in classifying PD patients (positive label). LR, RF, and SVM achieved AUPRC values of 0.80, 0.86, and 0.79, respectively. Baseline lies at 0.68 which is the partition of PD cases in the dataset, representing a random classifier. All classifiers exhibit moderate discriminatory power, with RF performing best.

To interpret the performances of the optimized classifiers in distinguishing PD patients, we used precision-recall (PR) curves and calculated the corresponding AUPRC values. As shown in Figure 3, LR, RF, and SVM achieved AUPRC val-

	Without FS			RFE			MDA			MRMR		
	LR	RF	SVM	LR	RF	SVM	LR	RF	SVM	LR	RF	SVM
Accuracy	0.65	0.70	0.63	0.64	0.70	0.64	0.67	0.70	0.65	0.67 **	0.71	0.66 **
F1 score	0.74	0.81	0.71	0.74	0.80	0.73	0.77	0.81	0.75	0.77 **	0.81	0.76 *
AUPRC	0.78	0.86	0.78	0.79	0.84	0.78	0.78	0.84	0.78	0.80 *	0.85	0.79 **

Table 3: Comparison of classifier performances with and without feature selection (FS) techniques. Only MRMR exhibited a significant improvement, although small, verified using paired t-testing with Bonferroni correction. One asterisk (*) indicates a significant p-value observed after 10 bootstrapped runs, while two asterisks (**) indicate significance observed after increasing to 30 runs to decrease variance between test samples.

ues of 0.80, 0.86, and 0.79, respectively. These values indicate that RF performs better than both LR (paired t-test, p-value = 0.0014) and SVM (paired t-test, p-value = 0.0004) in accurately identifying positive instances while minimizing false positives. However, it is important to interpret the overall predictive power with caution. The AUPRC values, although relatively high, should be considered in comparison to the baseline value of 0.68, which represents the partition of PD cases among the entire dataset and therefore exhibits the performance of a random classifier. The classifiers only perform between 0.12 to 0.18 percent better than this baseline value, suggesting that the ability to distinguish PD patients given the current dataset remains limited.

The moderate performances observed across all classifiers in this study can likely be partially attributed to the limited sample size of the data. Considering the high dimensionality of metagenomic data, the limited number of samples might not provide an adequate amount of training data for the ML models to effectively capture the complex and non-linear relationships present. As a result, the statistical power of the models may be compromised, leading to moderate performance outcomes.

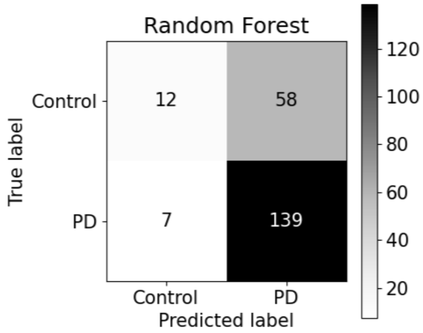

Figure 4: Confusion matrix for the classification performance of the RF classifier exhibits an overestimation of PD values. Approximately 9% of the total samples are classified as being control, while the true distribution lies at 32%.

Furthermore, it is crucial to take into account the influence of dataset imbalance on the classification performance. Imbalanced datasets can pose challenges for classification models as they tend to bias the models towards the majority class and may lead to lower performance metrics for the minority class. In our study, the dataset exhibits an imbalance, with the PD group being over-represented. The consequences of this imbalance can be observed in Figure 4, which depicts the confusion matrix for the RF classifier. It can be observed that only 19 cases are classified as control, which accounts for approximately 9% of the total samples, contrasting with the true distribution of 32%. This discrepancy further emphasizes the impact of dataset imbalance on the classifier's performance.

Considering the limitations imposed by the dataset size and imbalance, it is important to interpret the classification results cautiously and explore methods to mitigate the effects of dataset imbalance in future analyses. The PR curve analysis provides valuable insights into the performance of the classifiers, especially in scenarios with imbalanced datasets, which in this case indicates moderate performance on all classifiers with RF performing best.

3.3 Results align with existing literature and validate the usefulness of LR, RF and SVM in identifying biomarkers for PD

Building upon previous studies investigating metagenomic data in PD patients (Wallen et al., 2022; Bedarf et al., 2017; Qian et al., 2020), our objective was to validate the effectiveness of LR, RF, and SVM classifiers in identifying such PD biomarkers. To achieve this, we optimized these classifiers and identified the species deemed highly important during the classification process. Our findings reveal a significant overlap between the identified species and the results reported in the existing literature. This observation also strengthens the reliability and relevance of the identified species as potential biomarkers for PD.

The analysis involved a comparative examination of the most influential species in discriminating the samples during classification. Given the comparable performances of the ML models, we considered that identifying biomarkers based on consensus between the models would yield more reliable results. Consistency in feature selection was assessed through multiple runs, and the averaged results were considered to reduce variability. The overlap between the top fifteen species of each classifier, particularly those also found in existing literature, are presented in Figure 5. Species with significant changes in relative abundance were confirmed using Mann-Whitney U testing with Bonferroni correction. By retraining the models with only these fifteen species, notable improvements were observed in the LR and SVM classifiers, with AUPRC values of 0.88 and 0.86 respectively, while the RF classifier maintained a stable AUPRC value of 0.85. These findings confirm that these specific species alone possess significant discriminatory power for their respective classifiers, affirming their association with PD.

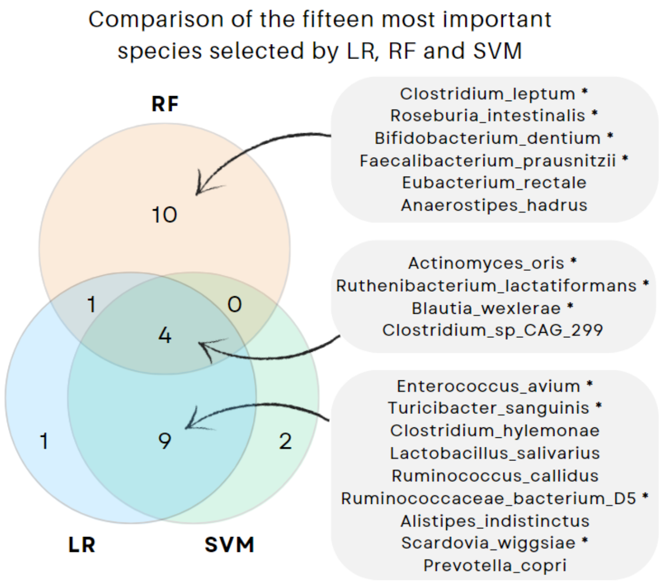

Figure 5: Comparison of the top fifteen highly important features selected by LR, RF, and SVM classifiers reveals a substantial overlap. The species indicated represent biomarkers previously reported in the literature. The asterisk (*) denotes a confirmed significance in relative abundance change between the PD and control groups.

Our analysis revealed that LR and SVM classifiers exhibited an almost complete overlap in the top-rated species, all of which have previously been associated with PD according to the study by Wallen et al. (2022). These species were consistently selected by the MRMR feature selection, with all but three being present in eight to ten runs. Notably, *Clostridium hylemonae* and *Scardovia wiggsiae*, which did lack consistency, have however been identified as associated with PD in multiple previous studies (Wallen et al., 2022; Mao et al., 2021). Additionally, Mao et al. (2021) has supported the association of *Alistipes indistinctus* and *Lactobacillus salivarius* with PD, both of which were selected by the LR and SVM classifiers.

Furthermore, among the nine species with the highest enrichment or depletion reported by Wallen et al. (2022), the unanimously selected species included three of them: *Actinomyces oris*, *Ruthenibacterium lactatiformans* and *Blautia wexlerae*. Two more species, *Enterococcus avium* and *Ruminococcaceae bacterium D5*, were selected by the LR and SVM classifiers, and three additional species, *Bifidobacterium dentium*, *Roseburia intestinalis* and *Bifilobacterium dentium*, were chosen by the RF classifier. Only *Streptococcus mutans* was failed to be detected by any of the classifiers, along with the other seven PD-associated *Streptococcus* species.

Our ML analysis did not corroborate the results from the study by Bedarf et al. (2017), except for the species *Prevotella copri*. We did not find substantial evidence supporting the influence of *Akkermansia muciniphila*, *Alistipes shahii*, *Eubacterium biforme* or *Clostridium saccharolyticum*, which aligns with the conclusions drawn by Wallen et al. (2022), the source of our data. This disparity in results could potentially be attributed to differences in taxonomic profiling techniques (MOCAT2 vs. MetaPhlAn3) or the geographical locations (Bonn, EU vs. Deep South, US) of the research and its participants.

Overall, our findings provide compelling evidence that LR, RF and SVM classifiers are capable of identifying biomarkers for PD. This is supported by the substantial overlap observed between the identified features and the existing literature. The species that demonstrated consensus among all classifiers consistently exhibited a strong association with PD. This observation also strengthens the reliability and relevance of the previously identified species as potential biomarkers for PD.

However, the process of biomarker identification solely based on ML classification encounters various challenges. The quality and representativeness of input data are crucial for reliable model performance, necessitating large-scale datasets with detailed metadata. In this study, limitations arise from the relatively small sample size and dataset imbalance. Metagenomic analysis involves high-dimensional data due to the multitude of species present, thus a larger dataset would provide more reliable and robust results. Additionally, dataset imbalance can impact model performance, leading to overfitting and the selection of features without genuine associations with the target disease. It is worth noting that the RF classifier in this research has shown an overestimation of PD cases, likely due to this dataset imbalance. While its top-rated features mostly align with existing literature, five unrelated species were also selected, highlighting the need for cautious interpretation when relying solely on ML-based approaches.

Another challenge in ML analysis for biomarker discovery is the involvement of feature selection or dimensionality reduction techniques. The success of these methods depends on the identification of relevant features that are informative for the classification task. However, in complex biological systems, identifying the most discriminative features associated with a specific disease can be difficult due to various factors such as inter-individual variability, genetic heterogeneity, and environmental influences.

In the context of this research, the impact of feature selection methods can be observed, particularly with the application of MRMR only to the LR and SVM classifiers. Each feature selection technique has its own criteria for selecting relevant features, which could explain the almost complete overlap of selected species between the LR and SVM classifiers, as well as the lack of overlap with the RF classifier, as depicted in Figure 5. It is important to consider that the different classifiers may inherently work in distinct ways, which also may contribute to these differences in results.

In conclusion, while ML analysis shows promise for biomarker discovery, the inherent difficulties in identifying new biomarkers through this approach should be acknowledged. Addressing these challenges requires careful data selection, study design and a comprehensive understanding of the underlying biology. While ML models may identify potential biomarkers based on statistical associations, further experimental validation is necessary to confirm their biological relevance and clinical significance. It is crucial to interpret the results cautiously and consider the limitations and relative performances of the classifiers in order to avoid overgeneralization or misinterpretation of the findings.

3.4 Confounding analysis demonstrates low model performances and fails to align with existing literature

A similar process has been conducted on the dataset excluding confounding factors relating to using alcohol, laxatives, pain medication, depression medication, anxiety medication, mood medication, probiotics, antihistamines, and sleep aids. This resulted in a significantly smaller dataset consisting of 107 samples, with 38 control subjects and 69 PD subjects. The models' performances on the reduced dataset demonstrated minimal discriminatory power, and there was limited overlap between the identified species and existing literature.

The LR, RF, and SVM models achieved AUPRC performance values of 0.71, 0.72, and 0.71, respectively, which indicated limited ability to discriminate between PD and control cases compared to the baseline random classifier (AUPRC of 0.64). Despite applying feature selection techniques, statistical analysis failed to demonstrate significant improvement.

However, after retraining the models using only the fifteen selected species, noticeable improvements were observed across all classifiers. The AUPRC values for LR, RF, and SVM increased to 0.85, 0.85, and 0.82, respectively, accompanied by an approximately 10% improvement in accuracy. Although these findings suggested a significant association between these species and PD, they lacked supporting evidence from existing literature.

Comparison of the top-rated species among all classifiers revealed an overlap between LR and SVM classifiers, with eight common species, but only five were associated with PD of which two were exclusively PD-related according to Wallen et al. (2022). Three species showed consensus across all classifiers, but only one was previously identified as a biomarker, namely *Ruminococcaceae bacterium D5*, which was however also associated with laxative usage. The results of the RF classifier exhibited no further overlap with those of LR and SVM, identifying only four species in existing literature, with one being exclusively PD-related. The LR classifier matched three exclusively PD-related species, while the SVM classifier identified two. Interestingly, the SVM classifier found an association with *Akkermansia muciniphila*, as reported by Bedarf et al. (2017).

Considering the relatively low performances of the models and the lack of confirmation from existing literature, it is evident that the LR, RF, and SVM models struggle to detect meaningful correlations. This limitation is likely due to the small dataset size which is known to have a major impact on ML model performances.

4 Responsible research

Responsible research practices are crucial for ensuring the integrity and validity of scientific studies. This section addresses four key considerations in conducting responsible research within the context of the current study.

Data source and attribution: The data utilized in this study is obtained from a previous study conducted by Wallen et al. (2022), and proper attribution and acknowledgment of the original data source are essential to upholding research integrity. Transparency regarding the origin of the data fosters collaboration and ensures that credit is given to the appropriate researchers. However, it is crucial to acknowledge that the data cannot be verified for the authenticity and uniqueness of individual participants and relies on the credibility and reliability of the researchers who provided the data.

Addressing biases: To maintain objectivity and validity in the research, it is important to be aware of and mitigate biases. Confirmation bias, which favors information that confirms preexisting beliefs, is a potential bias in this study because the results are compared to existing literature to confirm that the ML models behave successfully. To mitigate it, the results were compared only at the end of the process, while remaining objective during model optimization and training without favoring the inclusion of previous PD-associated species.

Additionally, the sampling process employed to collect the data might have introduced selection biases, whether conscious or unconscious. For example, the study focuses on a single geographical location, failing to obtain a totally generalizable result. Mitigating this bias has however been considered outside of the scope of this research, and has been considered further research.

Lastly, survivorship bias, which influences us to focus on the characteristic of the best-performing outcome and fail to consider other perspectives, has been mitigated. This has been done by not solely focusing on the best-performing model (RF), but also considering results from LR and SVM to obtain a more reliable outcome.

Sample size and sampling limitations: The sample size of the data used in this study may be too small for the high dimensionality of the data, affecting the statistical power and generalizability of the findings. The unequal distribution of PD and control subjects in the dataset should also be acknowledged, which introduces bias by potentially overestimating PD cases. The potential impacts of these sampling limitations have been carefully considered during result interpretation and are discussed in detail in section 3.

Reproducibility and data Accessibility: Ensuring the reproducibility of research findings is a fundamental aspect of responsible research. In line with this principle, the data used in this research is easily accessible and findable at Zenodo [https://zenodo.org/record/7246185].

Furthermore, this report places significant emphasis on a detailed explanation of the materials and methods, as explained in section 2. Every step leading to the obtained results has been meticulously described, including the specific Python libraries and objects utilized, as well as any modified input parameters. Random states have been employed and provided for transparency. By providing such a comprehensive methodology and ensuring data accessibility, this study promotes reproducibility and facilitates research validation.

Incorporating these responsible research practices, including acknowledging data sources, addressing biases, considering sample size and sampling limitations, and promoting data accessibility, enhances the transparency and reliability of the study.

5 Conclusion

This research aimed to evaluate the effectiveness of machine learning (ML) models in discovering biomarkers for Parkinson's disease (PD). An analysis of Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) ML models was conducted, comparing their most influential species in sample discrimination. This work provided an overview of the species that exhibited consensus among multiple classifiers, reinforcing the possible significance of these species as biomarkers by corroborating the findings with existing literature. This observation has strengthened the reliability of ML model approaches for biomarker analysis.

This research revealed that given our current dataset, including 490 PD patients and 234 healthy controls, all classifiers exhibit moderate performance in distinguishing between PD and control cases, indicating limited discriminatory power. Although the RF model exhibited the best performance, it displayed a tendency to overestimate PD cases, potentially due to dataset imbalance. This overfitting of the data might have caused the selection of species without genuine associations with PD. Caution should be exercised in interpreting the results, considering the limitations and relative performances of the classifiers to avoid overgeneralization or misinterpretation.

Despite achieving moderate performance, LR, RF, and SVM classifiers provided compelling evidence of their capability to identify PD biomarkers. This is supported by the substantial overlap observed between the identified species and the existing literature. The species that demonstrated consensus among all classifiers consistently exhibited a strong association with PD. This observation further supports the reliability and relevance of those species as potential biomarkers. However, confounding analysis failed to corroborate previous findings but also exhibited very limited classifier performances, likely due to the limited size of the remaining subset of the data.

Although these results overall provide promising evidence of the usefulness of ML models for PD biomarker discovery, research into this subject is only in the early stages and needs further investigation before it can be used for a fast and reliable diagnosis. The limitations of ML models, such as reliance on input data, overfitting, bias, and interpretability challenges, must be acknowledged. These limitations also emphasize that validation of the results based on biological relevance and clinical significance of the potential biomarkers is essential.

To develop a sufficiently accurate ML model for PD diagnosis, several requirements must be met. These include a comprehensive and balanced dataset with metagenomic profiles of a large number of PD and control subjects, accompanied by detailed subject metadata to account for confounding factors. Furthermore, given the current inaccuracies in clinical diagnosis, incorporating postmortem neuropathological assessment would be valuable in order to reduce the risk of misleading results due to the misclassification of metagenomic data. Moreover, considering additional factors such as disease duration, medication responsiveness, and PD-associated motor symptoms could potentially enhance the ML analysis, as previous research has indicated their positive impact on diagnostic accuracy. Finally, enhancing the interpretability of the model's decision-making process would aid in result verification and its validation of biological relevance.

Based on the requirements for an accurate and useful model, several suggestions for future research can be proposed. One of the key recommendations is conducting a large-scale clinical trial that collects a balanced metagenomic dataset including both PD patients and healthy controls, which can subsequently be verified through postmortem neuropathological assessment. It is important to ensure diversity among the trial participants in terms of age, sex, and race to achieve a more generalizable model for PD biomarker discovery. By achieving high and unbiased performances on ML classifiers, this comprehensive dataset can serve as a valuable framework for biomarker discovery using ML approaches. Research into the influence of including data on various typical PD symptoms on ML performance should be analyzed.

Furthermore, it is essential to expand research beyond the LR, RF, and SVM models discussed in this study. A thorough review of existing ML models should be undertaken to assess their relevance and potential in high-dimensional metagenomic analysis. This broader exploration will contribute to a more comprehensive understanding of the available ML models and their applicability in this field.

In summary, ML models have the potential to identify biomarkers for PD, but their application requires careful consideration of dataset characteristics, study design, and the need for experimental validation. The findings of this research contribute to the understanding of ML approaches for biomarker discovery in PD and highlight areas for further investigation.

References

Adler, C. H., Beach, T. G., Zhang, N., Shill, H. A., Driver-Dunckley, E., Mehta, S. H., Atri, A., Caviness, J. N., Serrano, G., Shprecher, D. R., Sue, L. I., and Belden, C. M. (2021). Clinical diagnostic accuracy of early/advanced parkinson disease: An updated clinicopathologic study. *Neurol Clin Pract*, 11(4):e414–e421. 2163-0933 Adler, Charles H Beach, Thomas G Zhang, Nan Shill, Holly A Driver-Dunckley, Erika Mehta, Shyamal H Atri, Alireza Caviness, John N Serrano, Geidy Shprecher, David R Sue, Lucia I Belden, Christine M Journal Article United States 2021/09/07 Neurol Clin Pract. 2021 Aug;11(4):e414-e421. doi: 10.1212/CPJ.0000000000001016.

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

Boktor, J. C., Sharon, G., Verhagen Metman, L. A., Hall, D. A., Engen, P. A., Zreloff, Z., Hakim, D. J., Bostick, J. W., Ousey, J., Lange, D., Humphrey, G., Ackermann, G., Carlin, M., Knight, R., Keshavarzian, A., and Mazmanian, S. K. (2023). Integrated multi-cohort analysis of the parkinson's disease gut metagenome. *Movement Disorders*, 38(3):399–409.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Ding, C. and Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology - JBCB*, 3:523–528.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.

Hajjo, R., Sabbah, D. A., and Al Bawab, A. Q. (2022). Unlocking the potential of the human microbiome for identifying disease diagnostic biomarkers. *Diagnostics (Basel)*, 12(7). 2075-4418 Hajjo, Rima Orcid: 0000-0002-7090-5425 Sabbah, Dima A Orcid: 0000-0003-1428-5097 Al Bawab, Abdel Qader Orcid: 0000-0002-6215-7949 2019-2020/23/07/Deanship of Scientific Research at Al-Zaytoonah University of Jordan/ Journal Article Review Switzerland 2022/07/28 Diagnostics (Basel). 2022 Jul 19;12(7):1742. doi: 10.3390/diagnostics12071742.

Han, H., Guo, X., and Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 219–224.

Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Huang, B., Chau, S. W. H., Liu, Y., Chan, J. W. Y., Wang, J., Ma, S. L., Zhang, J., Chan, P. K. S., Yeoh, Y. K., Chen, Z., Zhou, L., Wong, S. H., Mok, V. C. T., To, K. F., Lai, H. M., Ng, S., Trenkwalder, C., Chan, F. K. L., and Wing, Y. K. (2023). Gut microbiome dysbiosis across early parkinson's disease, rem sleep behavior disorder and their first-degree relatives. *Nature Communications*, 14(1):2501.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Mao, L., Zhang, Y., Tian, J., Sang, M., Zhang, G., Zhou, Y., and Wang, P. (2021). Cross-sectional study on the gut microbiome of parkinson's disease patients in central china. *Frontiers in Microbiology*, 12.

Mazzanti, S. (2021). mrmr (minimum-redundancy-maximum-relevance) for automatic feature selection at scale. https://github.com/smazzanti/mrmr.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qian, Y., Yang, X., Xu, S., Huang, P., Li, B., Du, J., He, Y., Su, B., Xu, L.-M., Wang, L., Huang, R., Chen, S., and Xiao, Q. (2020). Gut metagenomics-derived genes as potential biomarkers of parkinson's disease. *Brain*, 143(8):2474–2489.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9):833–844.

Thursby, E. and Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal*, 474(11):1823–1836.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wallen, Z. D., Demirkan, A., Twa, G., Cohen, G., Dean, M. N., Standaert, D. G., Sampson, T. R., and Payami, H. (2022). Metagenomics of parkinson's disease implicates the gut microbiome in multiple disease mechanisms. *Nature Communications*, 13(1):6958.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Yu, H.-F., Huang, F.-L., and Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1):41–75.