Towards a reduced feature selection pipeline in 16s rRNA microbiome data using Machine Learning

David Rojas-Velazquez, Sarah Kidwai, Paula Perez-Pardo, Daniel Oberski, Johan Garssen, Aletta D Kraneveld, Alejandro Lopez-Rincon

Towards a reduced feature selection pipeline in 16s rRNA microbiome data using Machine Learning

David Rojas-Velazquez[1,3], Sarah Kidwai[1], Paula Perez-Pardo[1], Daniel Oberski[3], Johan Garssen[1,2], Aletta D. Kraneveld[1], Alejandro Lopez-Rincon[1,3]

[1] Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands
[2] Global Centre of Excellence Immunology Danone Nutricia Research, Uppsalaan 12, 3584 CT Utrecht, The Netherlands
[3] Department of Data Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

The use of machine learning (ML) in the analysis of microbiome data is becoming more common due to the ability to process high dimensional data with small number of samples [8]. There are several pipelines for microbiome data analysis that allow researchers to perform statistical analysis to identify representative sequences related to a disease. Once these sequences are identified, researchers apply ML techniques to differentiate between groups, e.g. between healthy controls and patients [6]. Nevertheless, it has been shown that working with a small number of samples can cause overfitting. A proven solution to avoid overfitting is to use nested cross-validation [8] to produce robust and unbiased results regarding the number of samples.

In this work, we propose a novel pipeline that combines DADA2 [2], which is an open-source package for modeling and correcting errors in Illumina-sequenced amplicon inferring sample sequences, and a Recursive Ensemble Feature Selection (REFS) [3,4], method to discover biomarkers, applied to 16s rRNA sequencing gut microbiome dataset PRJEB33711 [1]. The aim of the selected study [1] is to show that the gut microbiome of Inflammatory Bowel Disease (IBD) patients is less diverse compared to healthy individuals.

In [1], authors use DADA2 pipeline and phyloseq [5] to process and perform statistical analysis on 16s sequences identifying a substantial imbalance in four major bacterial phyla (Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria). As study [1] reported gut dysbiosis in IBD patients, we postulated that a novel ML approach could be used to analyze gut microbiome data for predictive IBD diagnostics. To test the effectivity of our proposed pipeline, we compared the classification accuracy of control group and IBD group using all sequences (features) in the four bacterial phyla identified in [1] (2505 features) and the features selected in REFS (5 features – bacterial phyla: Bacteroidetes and Firmicutes).

The accuracy obtained by using the 5 features selected with REFS in a nested cross-validation (10-fold cross-validation) gives us an Area Under the Curve (AUC) is 0.818 in comparison to 0.5 obtained by using the 2505 features reported in the original IBD study. The AUC helps us estimate the diagnostic accuracy of a given methodology [7]. Thus, an AUC in a 10-fold cross validation close to 1.0 is a successful discriminative test [7].

Thus, our results point that REFS can be a suitable option in the analysis of microbiome data using DADA2 to build a pipeline to process the 16s rRNA sequencing data.

References

1.- Alam, M. T., Amos, G. C., Murphy, A. R., Murch, S., Wellington, E. M., & Arasaradnam, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut pathogens*, *12*(1), 1-8.

2.-Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), 581-583.

3.-Lopez-Rincon, A., Mendoza-Maldonado, L., Martinez-Archundia, M., Schönhuth, A., Kraneveld, A. D., Garssen, J., & Tonda, A. (2020). Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification. *Cancers*, *12*(7), 1785.

4.-Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A., & Tonda, A. (2019). Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. *BMC bioinformatics*, *20*(1), 1-17.

5.-McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, *8*(4), e61217.

6.- Mortera, S. L., Vernocchi, P., Basadonne, I., Zandonà, A., Chierici, M., Durighello, M., ... & Putignani, L. (2022). A metaproteomic-based gut microbiota profiling in children affected by autism spectrum disorders. *Journal of Proteomics*, *251*, 104407.

7.-Šimundić, A. M. (2009). Measures of diagnostic accuracy: basic definitions. *ejifcc*, *19*(4), 203.

8.- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), e0224365.

Annex 1

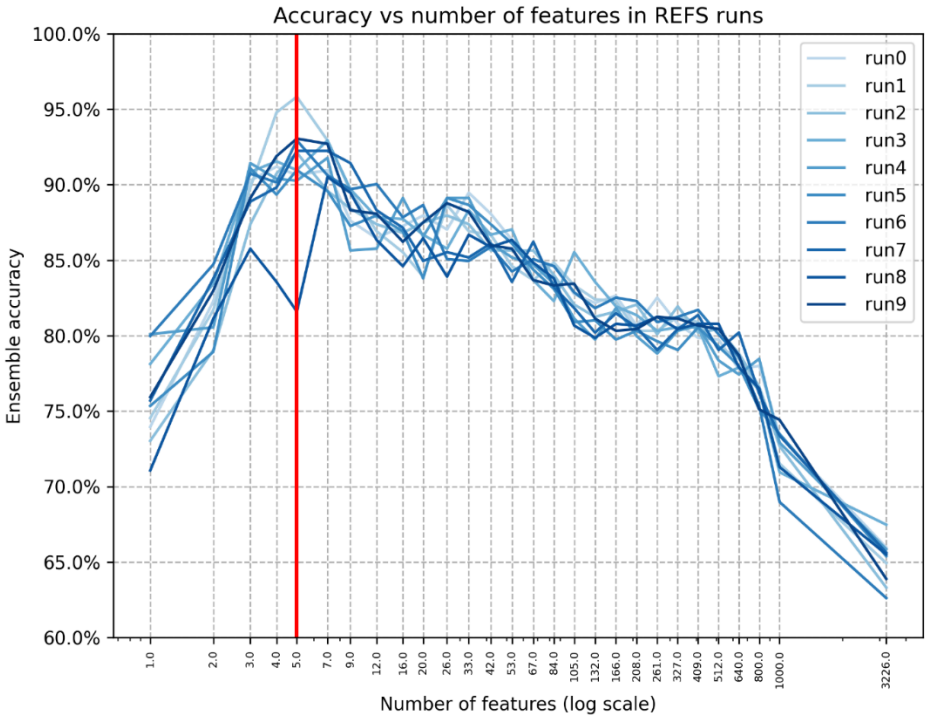

Figure 1. Best accuracy with de minimum features selected by using REFS.

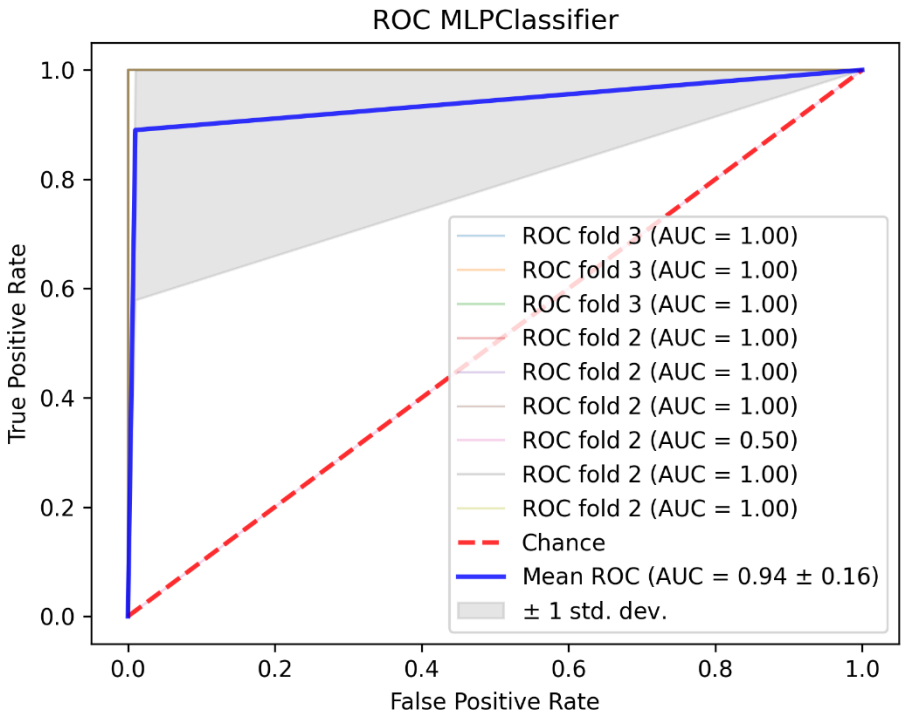

Figure 2. The ROC corresponding to the best classification rate using the 5 features resulting from REFS

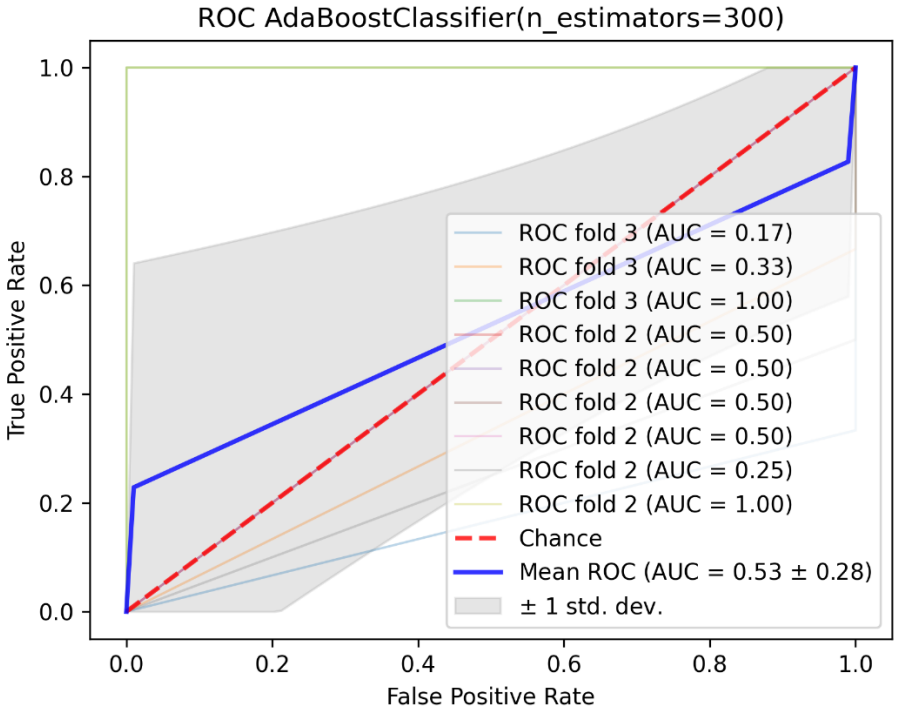

Figure 3. The ROC corresponding to the best classification rate using the 2505 features Reported in the IBD study

Annex 2

Table 1: Taxonomy assignment of the five features selected by using REFS

Feature	Domain	Phylum	Class	Order	Family	Genus	Species
GACGAACGCTGGCGGCGCGCCTAACACATGCAAGTCGAACGAGCGAGAGAGAGCTTG CTTTCTCGAGCGAGTGGCGAACGGGTGAGTAACGCGTGAGGAATCTGCCTCAAAGAG GGGGACAACAGTTGGAAACGACTGCTAATACCGCATAAGCCCACGGTGCCGCATGGC ACAGAGGGAAAAGGAGTAATCCGCTTTGAGATGGCCTCGCGTCCGATTAGCTAGTTGG TGAGGTAACGGCCCACCAAGGCGACGATCGGTAGCCGGACTGAGAGGTTGAACGGCC ACATTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTG CACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTCTTCGGAT	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	NA	NA
GATGAACGCTAGCTACAGGCTTAACACATGCAAGTCGAGGGGCAGCATGAACTTAGCT CGCTAAGTTTGATGGCGACCGGCGCACGGGTGAGTAACACGTATCCAACCTGCCGATG ACTCGGGGATAGCCTTTCGAAAGAAAGATTAATACCCGATGGCATAGTTCTTCCGCAT GGTAGAACTATTAAAGAATTTCGGTCATCGATGGGGATGCGTTCCATTAGGTTGTTGG CGGGGTAACGGCCCACCAAGCCTTCGATGGATAGGGGTTCTGAGAGGAAGGTCCCCC ACATTGGAACTGAGACACGGTCCAAACTCCTACGGGAGGCAGCAGTGAGGAATATTG GTCAATGGGCGATGGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTAT	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	NA
GACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAAGAGAGGAGCTT GCTCTTCTTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACCTGCCTTGTAGC GGGGGATAACTATTGGAAACGATAGCTAATACCGCATAACAATGGATGACACATGTCA TTTATTTGAAAGGGGCAATTGCTCCACTACAAGATGGACCTGCGTTGTATTAGCTAGTA GGTGAGGTAATGGCTCACCTAGGCGACGATACATAGCCGACCTGAGAGGGTGATCGG CCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCT TCGGCAATGGGGGCAACCCTGACCGAGCAACGCCGCGTGAGTGAAGAAGGTTTT	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	NA
GACGAACGCTGGCGGCGCGCCTAACACATGCAAGTCGAACGGAGCTTAGAGAGCTTG CTTTTTAAGCTTAGTGGCGAACGGGTGAGTAACGCGTGAGTAACCTGCCCTGGAGTGG GGGACAACAGTTGGAAACGACTGCTAATACCGCATAAGCCCACGGTACCGCATGGTAC TGAGGGAAAAGGATTTATTCGCTTTAGGATGGACTCGCGTCCAATTAGCTAGTTGGTG AGGTAACGGCCCACCAAGGCGACGATTGGTAGCCGGACTGAGAGGTTGAACGGCCAC ATTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGGATATTGCA CAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTTTTCGGAT	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	NA	NA
GATGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGGGGTGCTCATGACGGA GGATTCGTCCAACGGATTGAGTTACCTAGTGGCGGACGGGTGAGTAACGCGTGAGGA ACCTGCCTTGGAGAGGGGAATAACACTCCGAAAGGAGTGCTAATACCGCATGATGCA GTTGGGTCGCATGGCTCTGACTGCCAAAGATTTATCGCTCTGAGATGGCCTCGCGTCTG ATTAGCTAGTAGGCGGGGTAACGGCCCACCTAGGCGACGATCAGTAGCCGGACTGAG AGGTTGACCGGCCACATTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCA GTGGGGAATATTGGGCAATGGGGGAGACCCTGATGCAGCGACGCCGCGTGAGCGAA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Oscillospiraceae	Flavonifractor	NA