Identifying biological markers in the gut microbiome associated with celiac disease using machine learning.

Petr Persianov: 5036720[1]

Supervisor(s): Thomas Abeel[1]**, Eric van der Toorn**[1]**, David Calderón 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: Persianov Petr
Final project course: CSE3000 Research Project
Thesis committee: Thomas Abeel, Eric van der Toorn, David Calderón Franco

Abstract

Celiac disease is a genetic autoimmune disorder caused by a negative reaction to gluten associated with alterations in the gut microbiome. This study explored the potential of machine learning models and feature selection methods in identifying biomarkers for celiac disease using gut microbiome data. The performance of several machine learning models was evaluated, and the impact of different feature selection methods, including MRMR, ANOVA, and information gain, was examined. The findings revealed comparable performance among the models without feature selection. However, the choice of feature selection method had varying effects on model performance, with logistic regression and support vector machines being more sensitive than random forest and XGBoost models. Notably, several identified bacteria species, such as *Bacteroides eggerthii*, *Parabacteroides johnsonii*, *Faecalibacterium prausnitzii*, and *Ruminococcus_D bicirculans*, have been previously associated with celiac disease, reinforcing their potential as biomarkers for celiac disease.

1 Introduction

Celiac disease is an autoimmune disorder that affects approximately 1.4% of the global population (Singh et al., 2018), caused by an adverse reaction to gluten, a protein found in wheat, barley, and rye (Lindfors et al., 2019). Celiac disease is associated with potential complications, including osteoporosis and intestinal lymphoma (Catassi et al., 2022). The current standard for diagnosis is a combination of blood tests and an endoscopic biopsy of the small intestine (Lebwohl et al., 2018), both invasive and painful procedures.

Recent research has suggested that the gut microbiome, the collection of microorganisms that live in the digestive tract, may play a role in the development and progression of celiac disease (Sacchetti and Nardelli, 2020). Specifically, alterations in the composition of the gut microbiome have been observed in individuals with celiac disease compared to healthy individuals (Marasco et al., 2016; Rossi et al., 2023; Sellitto et al., 2012).

However, identifying biological markers in the gut microbiome that are associated with celiac disease is a complex and challenging task. Previous research has used machine learning to analyze the gut microbiome in various disease contexts, including inflammatory bowel disease (IBD) and colorectal cancer (Marcos-Zambrano et al., 2021) using stool samples collected from patients, using non invasive and painless procedure. These studies have shown that machine learning can be a powerful tool for identifying disease-specific biological markers in the gut microbiome (Ai et al., 2019; Qin et al., 2010). By analyzing large amounts of data from the gut microbiome samples using machine learning, researchers can identify patterns and relationships that may not be trivial.

However, no previous studies have specifically focused on using machine learning to identify biological markers in the gut microbiome for celiac disease. Therefore, the knowledge gap in this area is significant, and this research aims to fill this gap by using machine learning to identify biological markers that could aid in the diagnosis and treatment of celiac disease.

The main question this research is trying to answer is the following: *Can machine learning be used to identify biological markers in the gut microbiome that are associated with celiac disease?* Additionally, this research aims to answer the following sub-questions:

- Which machine learning algorithms excel in classifying celiac disease samples using gut microbiome data?
- Which feature selection methods work best with selected machine learning methods?
- What are the specific bacterial species associated with celiac disease identified by feature selection methods?

2 Materials and Methods

In the context of disease prediction and biomarker identification using gut microbiome data and machine learning the most common type of information used as features is relative abundance. Relative abundance refers to the proportionate representation of microbial taxa within a sample, reflecting the relative prevalence of each taxon to the overall microbial community.

2.1 Datasets

The raw DNA sequencing data and associated metadata for all samples were obtained from two independent studies: *"Dataset A"* (El Mouzan et al., 2022; PRJNA757365) and *"Dataset B"* (Francavilla et al., n.d.; PRJNA904924). Dataset A comprised 80 samples, while Dataset B comprised 132 samples. The data was initially downloaded in Sequence Read Archive (SRA) format (Leinonen et al., 2011) and subsequently converted to FASTQ format using SRA-Toolkit (version 3.0.1) (National Center for Biotechnology Information (NCBI), 2023).

2.2 Raw data processing

To extract the relative abundance data from the originally collected raw DNA sequencing data, a pipeline consisting of three consecutive steps was employed for each sample:

1. The first step involved processing the raw sequencing data to filter out low-quality DNA reads, remove any machine contamination, and trim the DNA reads. To accomplish this, we used the Trimommatic tool (version 0.39) (Bolger et al., 2014). Trimommatic is a widely-used tool for quality control and preprocessing of sequencing data. The following parameters were used for running Trimommatic, as shown in Table 1.
2. To assign taxonomic labels to DNA sequences, the processed DNA reads were matched against a Unified Human Gastrointestinal Genome (UHGG) v2.0.1 database (Almeida et al., 2021) using the Kraken 2 tool (Wood et al., 2019) with the default settings for paired-end reads.
3. For the estimation of relative abundance, the Bracken tool (version 2.8) (Lu et al., 2017) was used. Bracken

utilizes a statistical algorithm to approximate the relative abundance for every taxon for the specified taxonomic rank detected in a given sample based on the taxonomic assignments provided by Kraken2 software (Wood et al., 2019) from the previous step in the pipeline.

Trimming step	Parameter
SLIDINGWINDOW	4:20
ILLUMINACLIP	TruSeq3-PE.fa:2:40:15
HEADCROP	5
MINLEN	35
TOPHRED33	True

Table 1: Parameters used to run Trimommatic (version 0.39).

This pipeline allowed us to obtain relative species abundance data for each sample, which served as features for the machine learning classification task.

2.3 Mapping of the original samples classes

In a problem of identifying biomarkers associated with celiac disease, the classification of samples into two distinct categories becomes crucial, constituting a binary classification problem. To facilitate this, a mapping was developed following original studies to transform the original class names into binary labels. In this binary classification context, label 1 corresponds to samples with celiac disease, while label 0 represents healthy samples. The mapping of the original classes to binary labels used in this research is outlined in Table 2. This mapping resulted in an even distribution of binary labels with 106 samples labeled as 0 and 106 labeled as 1.

Original class	Binary label
Healthy	0
tCD-TG-	1
tCD-TG+	1
Untreated CD	1
Celiac	1
Non-Celiac	0

Table 2: Mapping of original classes names of the samples to binary labels.

2.4 Machine learning models selection

Choosing an appropriate machine learning model is crucial for the success of the task. There are several machine learning algorithms used for identifying biological markers in the gut microbiome for a disease (Marcos-Zambrano et al., 2021). For this research, the following models were selected: *XGBoost* (Chen and Guestrin, 2016), *random forest (RF)* (Ho, 1995), *support vector machines (SVM)* (Cortes and Vapnik, 1995) and *logistic regression (LR)*. Random forest, support vector machines and logistic regression models were implemented using scikit-learn (sklearn) Python library (version 1.2.2). XGBoost model was implemented using xgboost Python library (version 1.7.5).

2.5 Feature Selection

In this research for feature selection, the following methods were implemented: *Information Gain (IG)* (KENT, 1983), *Analysis of Variance (ANOVA)* (Ding et al., 2014) and *Minimum Redundancy Maximum Relevance (MRMR)* (Peng et al., 2005). *Information Gain* and *ANOVA* algorithms were implemented using scikit-learn (sklearn) Python library (version 1.2.2). *MRMR* was implemented using mrmr_selection Python library (version 0.2.6).

2.6 Hyperparameters optimization

In this research, RandomizedSearchCV was used due to its simplicity and relative effectiveness (Yang and Shami, 2020). RandomizedSearchCV was implemented using the *scikit-learn (sklearn)* Python library (version 1.2.2) (Pedregosa et al., 2018). It was performed on the "train" split of the data in combination with *StratifiedKFold* cross-validation with 10 splits. Distributions for parameters used for RandomizedSearchCV can be found in Table 3, *uniform* and *randint* functions were implemented using *scipy.stats* package from python library SciPy (version 1.10.1).

Model	Parameter Name	Distribution
RF	'n_estimators'	randint(1, 500)
	'criterion'	['gini', 'entropy']
	'max_depth'	[None] + list(range(1,30))
	'min_samples_split'	randint(2,20)
	'max_features'	['sqrt', 'log2']
XGBoost	'n_estimators'	randint(1,500)
	'learning_rate'	uniform(0.01,1.0)
	'max_depth'	randint(1,10)
	'subsample'	uniform(0.6,0.4)
	'colsample_bytree'	uniform(0.6,0.4)
	'reg_alpha'	uniform(0,1)
	'reg_lambda'	uniform(0,1)
LR	'C'	uniform(1,1000)
	'solver'	['liblinear', 'saga']
SVM	'C'	uniform(1,100)
	'kernel'	['linear', 'poly', 'rbf', 'sigmoid']
	'degree'	randint(1,4)
	'gamma'	['scale', 'auto]

Table 3: Names and distributions for parameters used for finetuning random forest (RF), XGBoost, logistic regression (LR) and support vector machines (SVM). *uniform* and *randint* functions were implemented using *scipy.stats* package from python library SciPy (version 1.10.1).

2.7 Performance evaluation and comparison

The following metrics were chosen to evaluate the performance of the models, as they are well suited for the task and commonly used in similar studies: *F1 score* (2), *Accuracy* (1), *Area Under the Curve* (Rubenzer, 2018). However, we will primarily focus on comparing the Area Under the Curve (AUC) performance metric as it provides the most reliable

estimation of a model's performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

To evaluate the performance of individual models and compare their effectiveness, we employ 95% confidence intervals (CI). This approach offers a deeper understanding of the true performance of the models (Berrar and Lozano, 2013). Confidence intervals were constructed using bootstrapping on the test set over 1000 iterations.

In addition, we utilized 5x2 cross-validation (Dietterich, 1998) in combination with a paired t-test to compare the performance of machine learning models. The paired t-test was implemented using the *mlxtend.evaluate.paired_ttest_5x2cv* function from the Mlxtend Python library (version 0.22.0) (Raschka, 2018). The significance level was set at $\alpha = 0.05$.

2.8 Responsible research

The most important ethical concern associated with this research are reproducibility and ethical implications.

Reproducibility ensures that the findings can be independently verified and built upon. To uphold reproducibility, this research provides comprehensive details of the implementation, including the machine learning algorithms, feature selection methods and evaluation metrics used. Moreover, the research paper includes a step-by-step description of the experimental setup, data preprocessing steps, and hyperparameter settings. By providing such thorough documentation, any researcher can replicate the experiments and verify the obtained results.

Ethical considerations play a significant role in research involving human data. In this study, ethical guidelines and regulations were strictly followed. The gut microbiome datasets utilized in this research do not contain any personal information that can help identify participants.

By addressing reproducibility and ethical implications, this research aims to uphold rigorous scientific standards and contribute to the advancement of knowledge in the field of identifying biomarkers for celiac disease using machine learning.

3 Results and Discussion

3.1 Comparable performance of models without feature selection

To assess whether there is a significant difference in performance among the models without utilizing any feature selection methods, 95% confidence intervals were constructed for the models' AUC performance metric on the test set. The plot in Figure 1 presents the mean AUC scores along with their corresponding 95% confidence intervals for the XGBoost, Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM) models. It can be observed that the models exhibit comparable performance, as there is a significant overlap in the confidence intervals. Furthermore, all p-values obtained from pairwise comparisons of the models were found to be greater than the established threshold of 0.05, indicating no statistically significant difference in performance.

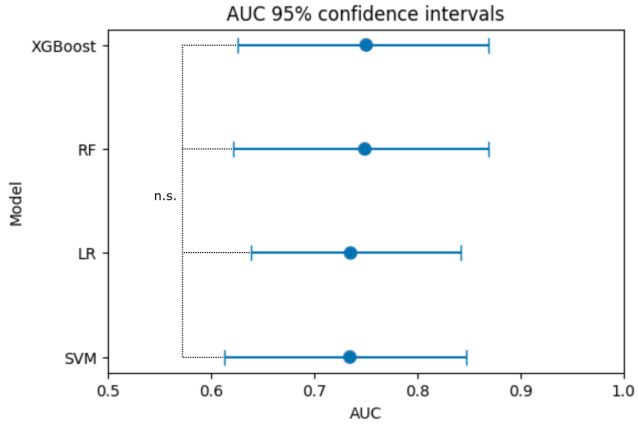

Figure 1: Mean AUC score on bootstrapped test set with 95% confidence intervals for XGBoost, Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM) models without any feature selection. Points represent the mean AUC score, error bars indicate 95% confidence intervals and 'n.s.' indicates that the difference in performance is not statistically significant.

Overall, the analysis suggests that without feature selection, the XGBoost, RF, LR, and SVM models demonstrate similar performance in terms of AUC scores. These models can be considered comparable options when feature selection is not employed and can be used as the baseline for evaluating the effectiveness of feature selection methods.

3.2 Impact of feature selection methods on model performance

In order to evaluate the impact of different feature selection methods (FSM) on model performance, all models were fine-tuned and trained using 100 features selected by feature selection methods on the training set, and their performance was evaluated on the test set.

Figure 2 presents the mean performance of all models with different feature selection methods, along with 95% confidence intervals. The findings suggest that there is no statistically significant difference in performance between the XGBoost and RF models with any of the feature selection methods. However, for the LR model, a statistically significant difference is observed when using the MRMR and IG feature selection methods. Additionally, a statistically significant difference in performance is found when using the SVM model in combination with the MRMR method.

The findings suggest that feature selection methods have a limited impact on the performance of the RF and XGBoost models. However, for LR and SVM, the choice of feature selection method can significantly influence performance.

3.3 Intersections between sets of selected features

Given the similar performance of the three feature selection methods (FSM), namely MRMR, ANOVA, and information gain, for RF and XGBoost models, it was of interest to examine the overlap between the sets of selected features. Such

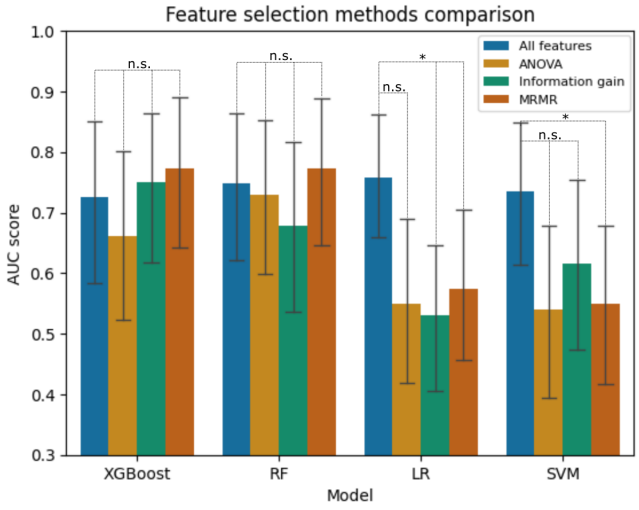

Figure 2: Mean AUC score on bootstrapped test set with 95% confidence intervals for XGBoost, Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM) models using different feature selection methods. Bars represent mean AUC score, error bars indicate 95% confidence intervals. 'n.s.' indicates that the difference in performance is not statistically significant. '*' indicates that there is a statistically significant difference with $p < 0.05$.

overlap could potentially indicate a higher level of importance for these features.

A total of 11 features were selected by all three FSM: *Bacteroides eggerthii*, *CAG-41 sp900066215*, *Campylobacter_A concisus*, *Faecalibacterium prausnitzii_H*, *Faecalibacterium prausnitzii_I*, *Gemmiger qucibialis*, *Parabacteroides johnsonii*, *Paraprevotella clara*, *Roseburia sp900552665*, *Ruminococcus_D bicirculans*, *SFEL01 sp004557245*. To determine the actual importance of these features, we compared the performance of models using the full set of features against models using only these 11 selected features. The results are presented in Figure 3, displaying the mean AUC scores with 95% confidence intervals. Notably, there was no statistically significant difference in performance.

These findings suggest that the subset of 11 commonly selected features is sufficient to achieve comparable performance to using the full set of features. This indicates the potential of these bacteria species as biomarkers for celiac disease.

3.4 Correspondence of selected features with existing findings

In order to investigate if these species can be considered potential biomarkers for celiac disease we analyzed findings of multiple previous studies on celiac disease. Figure 4 provides a visual representation of the mean relative abundance, highlighting statistically significant differences.

Bacteroides eggerthii and *Parabacteroides johnsonii* have been identified as significantly associated with celiac disease (El Mouzan et al., 2022). This study utilized one of the datasets employed in our research (PRJNA757365), providing support for the potential importance of these bacteria species as biomarkers. The difference in relative abundance for *Parabacteroides johnsonii* is statistically significant (p-value = 0.0026). However, for *Bacteroides eggerthii*, the difference in mean relative abundance between healthy and celiac samples is not statistically significant (p-value = 0.051). Nonetheless, this does not diminish the potential significance of this feature, as it may work in conjunction with other bacteria to provide valuable insights.

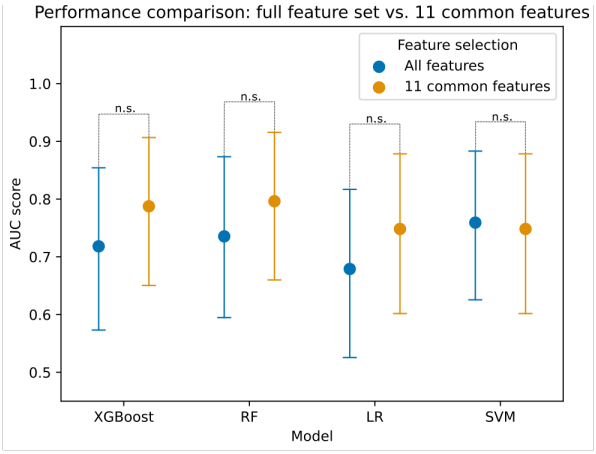

Figure 3: Performance comparison of models on the full set of features vs 11 commonly selected features by MRMR, ANOVA and information gain methods. Points represent the mean AUC score, error bars indicate 95% confidence intervals and 'n.s.' indicates that the difference in performance is not statistically significant.

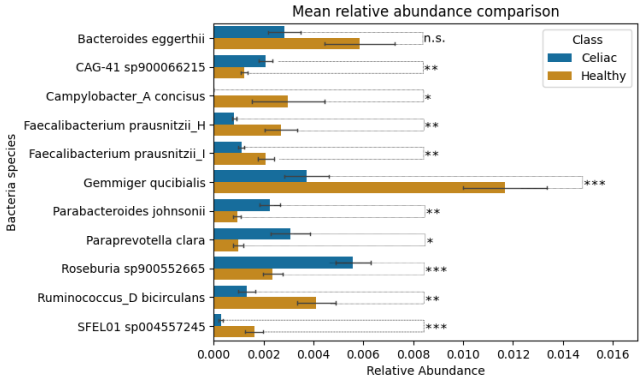

Figure 4: Comparison of mean relative abundance of bacteria between healthy and celiac samples. "n.s." indicates $p > 0.05$, "*" indicates $p \leq 0.05$, "**" indicates $p \leq 0.01$ and "***" indicates $p \leq 0.001$.

Furthermore, *Faecalibacterium prausnitzii_H* and *Faecalibacterium prausnitzii_I* exhibit statistically significant differences in abundance between healthy individuals and those with onset celiac disease (Leonard et al., 2021). These findings align with our research, as both species demonstrate a p-value lower than 0.01, indicating a substantial statistical difference. Notably, these bacteria species are known for their anti-inflammatory effects, which further supports the findings given the association between celiac disease and inflamma-

tion (Quévrain et al., 2016).

Paraprevotella clara displays increased abundance in the breast milk microbiota of mothers whose children later develop the celiac disease (Benítez-Páez et al., 2020). Consistent with these findings, our research reveals a statistically significant difference in the mean relative abundance of *Paraprevotella clara* between the two groups, with a p-value lower than 0.05. The identification of *Paraprevotella clara* as an important bacterium in both the gut and breast milk microbiota is highly intriguing. However, the precise implications of *Paraprevotella clara* in the development or progression of the celiac disease remain uncertain. Further comprehensive research is necessary to elucidate the specific mechanisms and significance of *Paraprevotella clara* in both the gut and breast milk microbiota within the context of celiac disease.

Ruminococcus_D bicirculans has been shown to exhibit lower abundance in samples with celiac disease compared to healthy samples (Francavilla et al., n.d.). This study was conducted using one of the datasets employed in our research (Francavilla et al., n.d.; PRJNA904924). Our research aligns with these findings, as we also observe a statistically significant difference in the mean relative abundance of *Ruminococcus_D bicirculans* with p-value = 0.0011.

However, none of the following bacteria species has been previously shown to be associated with celiac disease: *CAG-41 sp900066215*, *Campylobacter_A concisus*, *Gemmiger qucibialis*, *Roseburia sp900552665*, *SFEL01 sp004557245*. It is important to note that the absence of previous associations could be due to potential misclassification during the data preprocessing step or the use of a different database for classification, resulting in different classification names for these bacteria species. While this does not necessarily negate the possibility of these bacteria species being associated with celiac disease, it is crucial to emphasize that there is currently no supporting evidence. Notably, for all of these species, the difference in mean relative abundance between healthy and celiac samples was statistically significant. However, further research is necessary to determine whether these bacteria species can be considered biomarkers for celiac disease.

4 Conclusions and future work

4.1 Conclusions

This study successfully demonstrated the feasibility of using machine learning to identify biological markers for celiac disease in the gut microbiome. The findings revealed several important insights.

First, the models employed in this study, including XGBoost, Random Forest, Logistic Regression, and Support Vector Machines, showed comparable performance in predicting celiac disease when feature selection was not applied.

Second, impact of feature selection methods varied, particularly for Logistic Regression and Support Vector Machines, indicating the need for careful consideration when choosing an appropriate feature selection approach. Interestingly, a subset of 11 species commonly selected by all three feautes selection methods proved to be as effective as using the full set of feature, highlighting the potential of these bacterial species as biomarkers for celiac disease.

Furthermore, several of these bacteria species, such as *Bacteroides eggerthii*, *Parabacteroides johnsonii*, *Faecalibacterium prausnitzii*, and *Ruminococcus_D bicirculans*, have been previously associated with celiac disease, supporting their relevance in this context. However, other species, such as *CAG-41 sp900066215*, *Campylobacter_A concisus*, *Gemmiger qucibialis*, *Roseburia sp900552665*, and *SFEL01 sp004557245*, lack previous evidence of association. Although, statistically significant differences in abundance between healthy and celiac samples indicate that further research is necessary to establish their role as potential biomarkers for celiac disease.

4.2 Future work and limitation

Despite the insightful findings obtained in this study, there are several limitations and possible improvements for future research. The limitations include the population size and the choice of machine learning models and feature selection methods.

Firstly, the population size in this study may limit the statistical power and generalizability of the results. Increasing the sample size and including more diverse populations can provide a more comprehensive understanding of the gut microbiota's role in celiac disease. It is important to include individuals from different ethnic backgrounds and geographical locations to account for potential variations in gut microbiota composition (Gaulke and Sharpton, 2018; Gupta et al., 2017).

Secondly, the choice of feature selection methods and machine learning models can impact the performance and interpretability of the results. While various feature selection methods and models were explored in this study, there may be alternative approaches that could yield different results. For example, future research can consider using the recursive feature elimination (RFE) method for feature selection (Song et al., 2016). This can potentially improve the selection of bacteria species in the gut microbiome associated with celiac disease.

References

Ai, D., Pan, H., Han, R., Li, X., Liu, G., & Xia, L. C. (2019). Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer [Publisher: MDPI]. *Genes*, *10*(2), 112. https://doi.org/10.3390/genes10020112

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome [Publisher: Nature Publishing Group]. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Benítez-Páez, A., Olivares, M., Szajewska, H., Pieścik-Lech, M., Polanco, I., Castillejo, G., Nuñez, M., Ribes-Koninckx, C., Korponay-Szabó, I. R., Koletzko, S., Meijer, C. R., Mearin, M. L., & Sanz, Y. (2020). Breast-milk microbiota linked to celiac disease development in children: A pilot study from the Pre-

ventCD cohort. *Frontiers in Microbiology*, *11*, 1335. https://doi.org/10.3389/fmicb.2020.01335

Berrar, D., & Lozano, J. A. (2013). Significance tests or confidence intervals: Which are preferable for the comparison of classifiers? [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0952813X.2012.680252]. *Journal of Experimental & Theoretical Artificial Intelligence*, *25*(2), 189–206. https://doi.org/10.1080/0952813X.2012.680252

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Catassi, C., Verdu, E. F., Bai, J. C., & Lionetti, E. (2022). Coeliac disease [Publisher: Elsevier BV]. *The Lancet*, *399*(10344), 2413–2426. https://doi.org/10.1016/s0140-6736(22)00794-2

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Ding, H., Feng, P.-M., Chen, W., & Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis [Publisher: Royal Society of Chemistry]. *Molecular BioSystems*, *10*(8), 2229–2235. https://doi.org/10.1039/C4MB00316K

El Mouzan, M., Assiri, A., Al Sarkhy, A., Alasmi, M., Saeed, A., Al-Hussaini, A., AlSaleem, B., & Al Mofarreh, M. (2022). Viral dysbiosis in children with new-onset celiac disease [tex.eprint: 35030192 tex.eprinttype: pmid tex.pmcid: PMC8759644]. *PLoS ONE*, *17*(1), e0262108. https://doi.org/10.1371/journal.pone.0262108

Francavilla, A., Ferrero, G., Pardini, B., Tarallo, S., Zanatto, L., Caviglia, G. P., Sieri, S., Grioni, S., Francescato, G., Stalla, F., Guiotto, C., Crocella, L., Astegiano, M., Bruno, M., Calvo, P. L., Vineis, P., Ribaldone, D. G., & Naccarati, A. (n.d.). Gluten-free diet affects fecal small non-coding RNA profiles and microbiome composition in celiac disease supporting a host-gut microbiota crosstalk [tex.eprint: 36751856 tex.eprinttype: pmid tex.pmcid: PMC9928459]. *Gut Microbes*, *15*(1), 2172955. https://doi.org/10.1080/19490976.2023.2172955

Gaulke, C. A., & Sharpton, T. J. (2018). The influence of ethnicity and geography on human gut microbiome composition [Number: 10 Publisher: Nature Publishing Group]. *Nature Medicine*, *24*(10), 1495–1496. https://doi.org/10.1038/s41591-018-0210-8

Gupta, V. K., Paul, S., & Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in Microbiology*, *8*. Retrieved June 25, 2023, from https://www.frontiersin.org/articles/10.3389/fmicb.2017.01162

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

KENT, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, *70*(1), 163–173. https://doi.org/10.1093/biomet/70.1.163

Lebwohl, B., Sanders, D. S., & Green, P. H. R. (2018). Coeliac disease [Publisher: Elsevier BV]. *The Lancet*, *391*(10115), 70–81. https://doi.org/10.1016/s0140-6736(17)31796-8

Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*, D19–D21. https://doi.org/10.1093/nar/gkq1019

Leonard, M. M., Valitutti, F., Karathia, H., Pujolassos, M., Kenyon, V., Fanelli, B., Troisi, J., Subramanian, P., Camhi, S., Colucci, A., Serena, G., Cucchiara, S., Trovato, C. M., Malamisura, B., Francavilla, R., Elli, L., Hasan, N. A., Zomorrodi, A. R., Colwell, R., & Fasano, A. (2021). Microbiome signatures of progression toward celiac disease onset in at-risk children in a longitudinal prospective cohort study. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(29), e2020322118. https://doi.org/10.1073/pnas.2020322118

Lindfors, K., Ciacci, C., Kurppa, K., Lundin, K. E. A., Makharia, G. K., Mearin, M. L., Murray, J. A., Verdu, E. F., & Kaukinen, K. (2019). Coeliac disease [Publisher: Springer Science and Business Media LLC]. *Nature Reviews Disease Primers*, *5*(1). https://doi.org/10.1038/s41572-018-0054-z

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: Estimating species abundance in metagenomics data [Publisher: PeerJ Inc.]. *PeerJ Computer Science*, *3*, e104. https://doi.org/10.7717/peerj-cs.104

Marasco, G., Di Biase, A. R., Schiumerini, R., Eusebi, L. H., Iughetti, L., Ravaioli, F., Scaioli, E., Colecchia, A., & Festi, D. (2016). Gut microbiota and celiac disease. *Digestive Diseases and Sciences*, *61*(6), 1461–1472. https://doi.org/10.1007/s10620-015-4020-2

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., … Truu, J. (2021). Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology*, *12*. Retrieved April 30,

2023, from https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511

National Center for Biotechnology Information (NCBI). (2023). *SRA toolkit* (Version 3.0.1). https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2018, June 5). Scikit-learn: Machine learning in python. https://doi.org/10.48550/arXiv.1201.0490

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing [Publisher: Nature Publishing Group UK London]. *nature*, *464*(7285), 59–65. https://doi.org/10.1128/mSystems.00230-19

Quévrain, E., Maubert, M. A., Michon, C., Chain, F., Marquant, R., Tailhades, J., Miquel, S., Carlier, L., Bermúdez-Humarán, L. G., Pigneur, B., Lequin, O., Kharrat, P., Thomas, G., Rainteau, D., Aubry, C., Breyner, N., Afonso, C., Lavielle, S., Grill, J.-P., … Seksik, P. (2016). Identification of an anti-inflammatory protein from faecalibacterium prausnitzii, a commensal bacterium deficient in crohn's disease [Publisher: BMJ Publishing Group Section: Inflammatory bowel disease]. *Gut*, *65*(3), 415–425. https://doi.org/10.1136/gutjnl-2014-307649

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack [Publisher: The Open Journal]. *The Journal of Open Source Software*, *3*(24). https://doi.org/10.21105/joss.00638

Rossi, R. E., Dispinzieri, G., Elvevi, A., & Massironi, S. (2023). Interaction between gut microbiota and celiac disease: From pathogenesis to treatment [tex.article-number: 823 tex.pubmedid: 36980164]. *Cells*, *12*(6). https://doi.org/10.3390/cells12060823

Rubenzer, S. (2018, March 1). Area under the curve explained. In S. J. Rubenzer (Ed.), *Assessing negative response bias in competency to stand trial evaluations* (p. 0). Oxford University Press. https://doi.org/10.1093/med-psych/9780190653163.005.0001

Sacchetti, L., & Nardelli, C. (2020). Gut microbiome investigation in celiac disease: From methods to its pathogenetic role. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *58*(3), 340–349. https://doi.org/10.1515/cclm-2019-0657

Sellitto, M., Bai, G., Serena, G., Fricke, W. F., Sturgeon, C., Gajer, P., White, J. R., Koenig, S. S. K., Sakamoto, J., Boothe, D., Gicquelais, R., Kryszak, D., Puppa, E., Catassi, C., Ravel, J., & Fasano, A. (2012). Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. [tex.pmcid: PMC3303818]. *PloS one*, *7*(3), e33387. https://doi.org/10.1371/journal.pone.0033387

Singh, P., Arora, A., Strand, T. A., Leffler, D. A., Catassi, C., Green, P. H., Kelly, C. P., Ahuja, V., & Makharia, G. K. (2018). Global prevalence of celiac disease: Systematic review and meta-analysis [Publisher: Elsevier BV]. *Clinical Gastroenterology and Hepatology*, *16*(6), 823–836.e2. https://doi.org/10.1016/j.cgh.2017.06.037

Song, H., Yoo, Y., Hwang, J., Na, Y.-C., & Kim, H. S. (2016). Faecalibacterium prausnitzii subspecies–level dysbiosis in the human gut microbiome underlying atopic dermatitis [Publisher: Elsevier]. *Journal of Allergy and Clinical Immunology*, *137*(3), 852–860. https://doi.org/10.1016/j.jaci.2015.08.021

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061