pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210591

Machine Learning Based Metagenomic Prediction of Inflammatory Bowel Disease

Andrea MIHAJLOVIĆ [a,1] Katarina MLADENOVIĆ[b],
Tatjana LONČAR-TURUKALO[b] and Sanja BRDAR[a]
[a] *BioSense Institute, University of Novi Sad, Novi Sad, Serbia*
[b] *Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia*

Abstract. In this study, we investigate faecal microbiota composition, in an attempt to evaluate performance of classification algorithms in identifying Inflammatory Bowel Disease (IBD) and its two types: Crohn's disease (CD) and ulcerative colitis (UC). From many investigated algorithms, a random forest (RF) classifier was selected for detailed evaluation in three-class (CD versus UC versus nonIBD) classification task and two binary (nonIBD versus IBD and CD versus UC) classification tasks. We dealt with class imbalance, performed extensive parameter search, dimensionality reduction and two-level classification. In three-class classification, our best model reaches F1 score of 91% in average, which confirms the strong connection of IBD and gastrointestinal microbiome. Among most important features in three-class classification are species Staphylococcus hominis, Porphyromonas endodontalis, Slackia piriformis and genus Bacteroidetes.

Keywords. microbiome, imbalance, machine learning, feature selection

1. Introduction

Inflammatory bowel disease (IBD) is an umbrella term used to describe chronic inflammation of digestive tract. It includes two types of disease: Crohn's disease (CD) and ulcerative colitis (UC). They share many common features – diarrhea, bloody stools, weight loss, abdominal pain, fever, and fatigue, even though they affect different part of digestive tract. The exact cause of IBD is unknown, but some risk factors are known. It is a genetic disease, manifested under certain external influences. Research findings imply that microbiome has a fundamental role in patients with IBD, in all aspects: the development, progression, and treatment [1,2]. Machine learning (ML) algorithms applied on microbiome data have huge potential in uncovering patterns and aiding diagnosis of diseases including IBD. Early diagnosis is crucial in helping patients particularly in cases of diseases which are caused by microbial and environmental factors, since prevention in that case could be more efficient.

In this study we investigate faecal microbiota composition, in an attempt to evaluate performance of classification algorithms in identifying IBD state. The aim is to predict patients state based on their metagenomic taxonomic profile in different time points. Since the IBD is a genetic disease, it is considered that the state remains unchanged during time. There are three possible states: CD, UC and nonIBD. Two of them (CD and

[1] Corresponding Author: Andrea Mihajlović, BioSense Institute, Dr Zorana Đinđića 16, 21000 Novi Sad, Serbia; E-mail: andrea.mihajlovic@biosense.rs.

UC) represent unhealthy individuals or IBD. We perform three-class classification (CD versus UC versus nonIBD) to distinguish between the states. In addition, two binary classification tasks (nonIBD versus IBD and CD versus UC) are examined. Our approach differs in using only metagenomic data for prediction unlike other studies on the same database [3]. More on related work can be found in [4].

2. Data

Resource for IBD data used in this paper is the Inflammatory Bowel Disease Multi'omics Database (IBDMDB, https://ibdmdb.org), part of the Integrative Human Microbiome Project [5]. For the classification task we use table of metagenomic taxonomic profiles and metadata, specifically sample and subject identifiers and true labels indicating clinical diagnosis. Dataset contains 429 samples from 27 healthy subjects, 750 samples from 65 CD subjects and 459 samples from 38 UC subjects. In total 1209 samples from 103 IBD subjects.

Microbes are grouped into 1479 operational taxonomic units (OTUs). Grouping of microorganisms is done in the following taxonomic levels: kingdom, phylum, class, order, family, genus, species and strains. Detailed description of the workflow used for producing the OTU table from the raw DNA sequences is available on the bioBakery 2.0 GitHub repository https://github.com/biobakery/hmp2_workflows. The OTU table is transposed such that each row represents one sample and each column one feature, i.e. OTU. Values in the OTU table are given as relative abundances of particular OTU in some sample, relative to all reads from that sample assigned at the same taxonomic level. Sum by level in each sample is approximately 1 (due to the rounding errors).

3. Methods

3.1. Workflow

The adopted approach consists of three main steps: (i) choosing of suitable ML model and forming of model pipeline; (ii) hyperparameters search; and (iii) training and evaluation of the best model. Further, mentioned steps are explained in details. The proposed workflow is presented in the Figure 1. It is worth noting that in the subsequent steps the care was taken with respect to which sample belongs to which subject along with its diagnosis. The model used is described in Section 3.2. Parameter searching was conducted using a group cross-validation approach. In the given problem, group cross-validation is used in order to ensure that all samples from one subject are either in the training set or in the validation set. Initial set of parameters was created at random, in order to narrow parameters searching space. Best performing parameters from the initial random search were used as a guidance for the more thorough grid search and thus further fine-tuned. Training was performed for each model in 100 iterations with 10-fold group cross-validation. This ensures insight into stability of performance and more comprehensive evaluation.

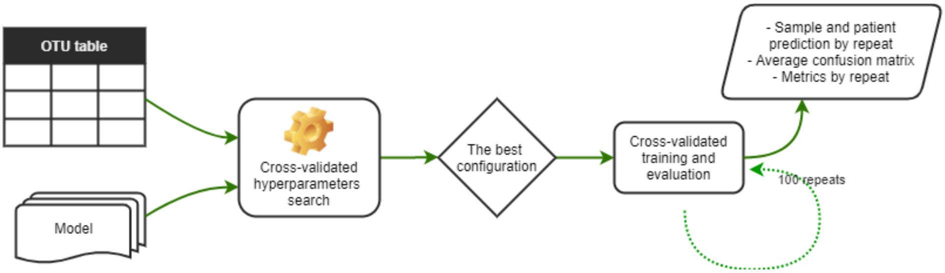
OTU table
Model
Cross-validated hyperparameters search
The best configuration
Cross-validated training and evaluation
100 repeats
- Sample and patient prediction by repeat
- Average confusion matrix
- Metrics by repeat

Figure 1. The workflow overview.

Upon sample-wise binary classification, subjects are labelled as positive (in one case IBD, in other UC) if average decision probability of their samples that model outputted is above the certain threshold. The threshold was as well treated as a parameter and varied. In three-class classification, the decision becomes somewhat complicated. For that purpose, we employ two thresholds and tune them. Firstly, the average decision probability for each class is estimated by the model used. By summing two probabilities for IBD classes (CD and UC) and comparing with average probability for nonIBD class, we decide if the subject is an IBD (above *th1*) or not. Furthermore, if it is an IBD, we decide which type of disease subject might have by normalizing average probabilities for CD and UC on the IBD event and decide if the subject is UC or not (above *th2*).

3.2. Model

Data imbalance have an impact on the classification since some events becomes so rare that it is impossible for classifier to learn useful patterns about them. One possibility is to use class weighting if applicable with the learning algorithm. Another solution are widely adopted resampling techniques, which assume either down- and over- sampling. Downsampling reduces the number of samples in the majority class to balance the classes, while over-sampling increases the number of samples in the minority class. Apart from the random sampling with replacement, there is a popular method to over-sample minority class(es) *the Synthetic Minority Oversampling Technique* (SMOTE) [6].

From many investigated algorithms, a Random Forest (RF) classifier [7] is selected for detailed evaluation in our classification tasks. RF randomizes decision tree through the features/samples sub-sampling (bootstrap) and groups trees to make a final decision on the basis of majority voting / averaging. In learning extremely imbalanced data, there is a significant probability that a bootstrap sample contains few or even none of the minority class, resulting in a tree with poor performance for predicting the minority class [8]. To overcome this issue, Chen et al. [8] proposed algorithm for balancing downsampled data in a bootstrap process called *Balanced Random Forest* (BRF). Using a random selection of features to split each node (tree growing), in both RF and BRF, each tree, gives an internal estimate which leads to feature importance [9]. This property of (B)RF will be very useful for microbiome analysis after classification since it can tell us which features most helped in making decision.

In the dataset used classes are highly imbalanced, with IBD samples making up 74% of the dataset and almost two times more CD samples than UC. To approach class imbalance problem, three model pipelines were evaluated and compared: (1) class weighting of RF; (2) concatenation of SMOTE and RF; and (3) BRF. The feature selection was used in all scenarios to reduce data dimensionality. Experiments included hand-picked taxa and/or selecting k best scored features (SKB), an (*Univariate Feature Selection* method based on *ANOVA* F-test [10]).

The workflow is implemented in Python 3.0. For the ML algorithms *scikit-learn* 0.24.2 library [11] is used in addition with *imbalanced-learn* 0.8.0 library [12] for the resampling and BRF.

4. Results

In order to find best suitable hyperparameters, the parameters search was repeated several times. It was noticed that the best parameters' vicinity is similar for the same model type between the classification tasks. Also, change in classification performance as a function of the employed thresholds was noted. Table 1 contains best parameters configuration for each of the pipelines. Results of three-class classification is shown in the Figure 2.

Table 1. Parameters configuration for the three pipelines in three-class classification

Parameters	SKB_BRF	SMOTE_SKB_RF	SKB_RF
RF max depth	15	3	3
RF # of estimators	150	200	200
RF class weight	None	None	Balanced subsample
SKB *k*	500	300	300
*th*1	0.45	0.6	0.55
*th*2	0.5	0.53	0.5

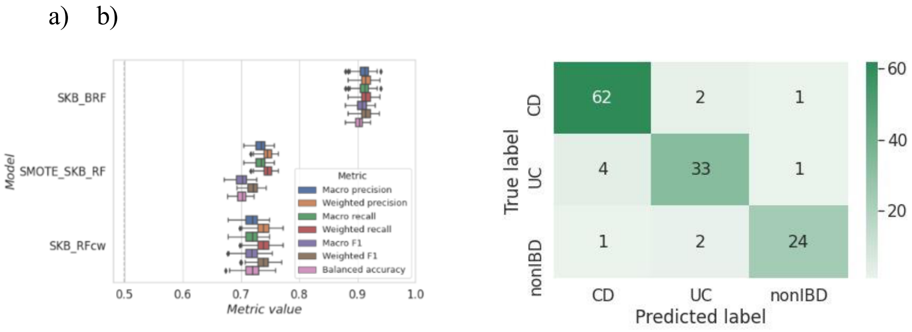

Figure 2. Three-class classification results: (a) metric values in 100 repeats and (b) confusion matrix on patient level for the best model (SKB_BRF).

Each model performance metric can be seen on the left (Figure 2a) and average number (in repeats) of well classified and misclassified patients for the pipeline containing BRF classifier on the right (Figure 2b). For the best evaluated model (SKB_BRF) all metrics are more or less similar (balanced accuracy is slightly worse in average) with an average score 91%. Other two model pipelines achieved significantly worse results with the weighted metrics slightly better than unweighted (macro).

Additionally, importance of each feature for correct classification of the best performing model was calculated. The *Bacteroidetes* genus has shown to be the most important. The rest of the features are mainly on species taxonomic level, but also some strains show and two orders *Coriobacteriales* and *Lactobacillales*. Top 20 selected features and their importance are presented in the Figure 3. Feature order is constant in repeats. Since OTU name is too large, we show only last taxonomic level of particular OTU. The first letter in the feature name indicates taxonomic level: o - order; g - genus; s - species and t - strain.

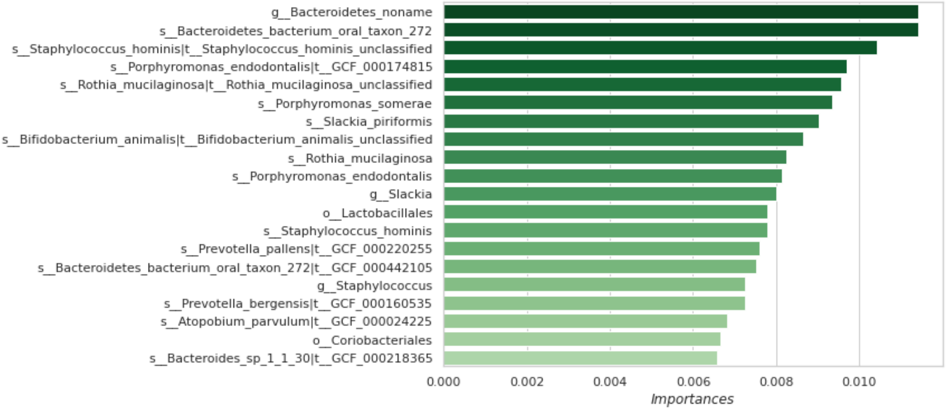

Figure 3. Feature importance for the best model in three-class classification.

In binary classification tasks, SKB_BRF model pipeline again significantly outperformed the rest. In CD versus UC case after 100 repeats, both average AUC and balanced accuracy reached 90% and average F1 score (unweighted and weighted) was 91%. In nonIBD versus IBD case, these values were 92%, 91% (unweighted) and 94% (weighted), respectively. Confusion matrices are in the Figure 4.

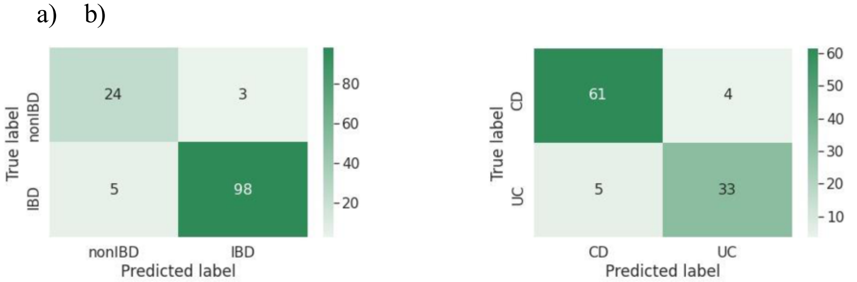

Figure 4. Confusion matrices for the best model (SKB_BRF) on patient level: (a) nonIBD versus IBD case and (b) CD versus UC case.

Based on feature importance, the top 20 features differ slightly in these two cases. In nonIBD versus IBD case, the top features comprise mainly *Alistipes* and *Bacteroides* genus, with *A. putredinisi* as most important, and *A. shahii* and *B. ovatus* among most common species. In CD versus UC case, the highlighted features are Clostridium, *Odoribacter*, *Dorea* and *Alistipes* genus, with species *C. clostridioforme*, *D. formicigenerans*, and strain *O. splanchnicus* GCF_000190535.

5. Conclusion and discussion

We addressed the classification problem in IBD dataset, taking in consideration the associated problems: class imbalance, number of samples per subject imbalance and an overall lack of data, in an attempt to avoid overfitting, a usual pitfall in applying ML algorithms. Moreover, we managed to reduce considerably large hypeparameters searching space emerged in dealing with all these problems. Among the evaluated classifiers, Balanced Random Forest showed the best performance and achieved the balance among different metrics. An additional advantage of the model used is the information on feature importance. However, to investigate the level of the OTUs (features) presence in different subject groups further analysis on a dataset is needed including the domain experts. The development of standardized ML pipeline would benefit from more data and an enhanced model explainability.

6. Acknowledgement

The financial support of the Ministry of Education, Science and Technological Development of the Republic of Serbia is acknowledged (Grant No. 451-03-9/2021-14/200358).

References

[1] Glassner K.L, Abraham B.P, Quigley E.M.M. The microbiome and inflammatory bowel disease. *J Allergy Clin Immunol*. 2020;145,1:16-27.

[2] Azimi T, Nasiri M.J, Chirani A.S, Pouriran R, Dabiri H. The role of bacteria in the inflammatory bowel disease development: a narrative review. *APMIS*. 2018;126,4:275-283.

[3] Hassouneh S.A.D, Loftus M, Yooseph S. Linking Inflammatory Bowel Disease Symptoms to Changes in the Gut Microbiome Structure and Function. *Frontiers in Microbiology*. 2021;12.

[4] Marcos-Zambrano L. J, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, Magali B, Gruca A, Hasic J, Hron K, Klammsteiner T, Kolev M, Lahti L, Lopes M.B, Moreno V, Naskinova I, Org E, Pacienca I, Papoutsoglou G, Shigdel R, Stres B, Vilne B, Yousef M, Zdravevski E, Tsamardinos I, de Santa Pau E.C, Claesson M, Moreno-Indias I, Truu J. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology*. 2021;12,313.

[5] Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe*. 2014;16,3:276-89.

[6] Chawla, N, Bowyer, K, Hall, L, Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res*. 2002;16: 321-357.

[7] Breiman L. Random Forests. *Machine Learning*. 2001;45:5–32.

[8] Chen C, Breiman L. Using Random Forest to Learn Imbalanced Data. Berkeley: *University of California*. 2004.

[9] Huynh-Thu V.A, Saeys Y, Wehenkel L, Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery, Bioinformatics. 2012;28,13:1766–1774.

[10] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* (Oxford, England). 2007;23:2507-2517.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thriton B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau, D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12,85:2825-2830

[12] Lemaitre, G, Nogueira F, Aridas C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*. 2017;18,17:1-5.