Machine Learning to Aid Diagnosis of Inflammatory Bowel Disease Using Metagenomic Data

July 17, 2023

Milla Kosonen

1 Introduction

Machine learning models and omics data have been widely used to characterize and detect IBD. During recent years there have been major advancements in developing and using machine learning techniques to aid diagnosis of IBD as well as to identify its two subtypes to determine correct treatment types. Random forest classifiers have been particularly popular and have produced promising results with high classification accuracies. Other common machine learning models that have been used for diagnostic purposes in IBD include boosting methods, neural networks, regression, support vector machines and hierarchical clustering. [11]

For example, in [7] Mihajlović et al. used operational taxonomic units (OTUs) as features in classification of IBD/healthy. They found that the Balanced Random Forest model with the select k-best feature selection method produced the best results and achieved a balanced accuracy score and average AUC score of 0.92. In [6] Manandhar et al. also found that random forest classifier achieved best performance on the IBD/healthy classification task. In this study [6] the researchers used bacterial taxonomic features and OTUs for the classification. In their study, random forest with the bacterial taxonomic features achieved an average AUC score of 0.80 and accuracy 0.72 and both SVM (with rbf kernel) and elastic net achieved AUC score 0.73 and accuracy 0.66.

There are also studies that investigate the use of machine learning and feature selection methods to identify IBD related biomarkers and the set of gut microbiota that is mostly associated with the disease. Manandhar et al. [6] found that at the phylum level IBD is characterized by increased levels of Firmicutes and decreased levels of Bacteroidetes, which is consistent with previously reported results [1] [2] [10]. In addition, they observed increased levels of phyla Verrucomicrobia and Actinobacteria in IBD patients.

The aim of this project is to compare performances of two data types in binary classification of IBD versus healthy based on metagenomic data. Metagenomics refers to the direct analysis of genetic content of environmental samples [12]. In this case the samples are collected from the human gut from both healthy individuals and patients diagnosed with IBD. Metagenomic data can be characterized in different ways. This project focuses on using two data types, namely relative abundance of microbes in the human gut and abundance of clade-specific marker genes. The goal is to compare the predictive abilities of these two data types.

2 Methods

This section introduces the datasets, data pre-processing steps and the machine learning model that were used in this project. SVM was chosen as the machine learning model for this project.

2.1 Dataset description

The datasets were obtained from the curatedMetagenomicData R-package [8]. The package includes data from multiple studies relating to various diseases including e.g. IBD, colorectal cancer and diabetes. The datasets we used were obtained from [3], [5] and [13].

The curatedMetagenomicData package provides 6 types of data for each study. We chose two of these data types for comparison. The first data type is relative abundance which tells how many percentages of the microbiome in a sample is made up of a specific microbe. The second data type is clade-specific marker gene abundance. Clade-specific markers are genes that are common within a clade and a clade is a group that consists of a common ancestor and its direct descendants.

2.2 Pre-processing

To increase sample size we combined data from different sources. We first examined data from 6 different studies containing samples labelled either healthy or IBD. Each study had found slightly different features i.e. different microbes or marker genes. Because of this combining these 6 studies by their common features resulted in loss of data. To overcome the problem of losing data we chose 3 studies [3] [5] [13] that had originally about the same number of features and sufficient number of samples. By combining data from these 3 studies each sample was left with at least 70% of the original data for both data types. Before combining the datasets, features that were zero for all samples were removed.

The final datasets contain 901 samples with 258 samples from healthy subjects and 643 from patients diagnosed with IBD. The relative abundance dataset has 454 features and the marker abundance dataset has 49463 features. The datasets were divided into train and test sets such that 80% of samples were used for training. The data was split into train and test sets in a stratified fashion based on the labels to ensure that both sets contain the same proportion of control and IBD samples as the original dataset.

2.3 Machine Learning Model

In this section the basic principles of the machine learning model and its parameters are described. The best model parameters were determined by grid search using scikit-learn GridSearchCV [9]. Grid search performs an exhaustive search over a specified parameter space. For each set of parameters stratified cross-validation with 5 folds was performed, mean accuracy over all folds was calculated, and the parameter combination with highest cross-validation accuracy was chosen. The stratification was based on the labels and the train set was used for the parameter search.

2.3.1 Support Vector Machine

SVM classifier was implemented in Python with sklearn.svm.SVC [9]. The parameters to be optimized for this implementation include kernel, C and γ parameters. C regulates the misclassification penalty and γ is a kernel parameter. γ parameter describes how far the influence of a training data point reaches. Low values of γ mean long-distance influence and high values mean short range influence. Too large γ values lead to overfitting and too small γ values fail to capture the complexity of the dataset. Common kernel functions that are used in SVM are the polynomial, linear and sigmoid kernels and radial basis function (RBF).

For both datasets the RBF was chosen as the kernel function. RBF has been demonstrated to be a good choice and it has been used for example in [6] in a similar IBD related classification task. The parameters to optimize then include C and γ. The best C parameter was searched from the range $C = 2^{-5}, 2^{-3}, ..., 2^{10}$ and γ from $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$ [4]. For the relative abundance dataset the best C parameter was 32 and best γ was 0.000122. For the marker abundance dataset the best C parameter was 2 and best γ was 0.0000305. The features were first scaled with sklearn.preprocessing.StandardScaler [9]. Scaling is required for the SVM classifier to work properly.

3 Results

This section presents the results from the two datasets. The results were evaluated with common performance measures including precision, recall, specificity, f1-score, balanced accuracy (average of recall obtained on each class) and AUC-score. Precision is also called positive predictive value and it tells what proportion of all positive predictions (model predicted IBD) were true positives. Recall or true positive rate describes how well the model identifies IBD and specificity or true negative rate how well it identifies negatives i.e. healthy. In addition, confusion matrix and ROC curve were used for evaluation and visualization of the prediction performance.

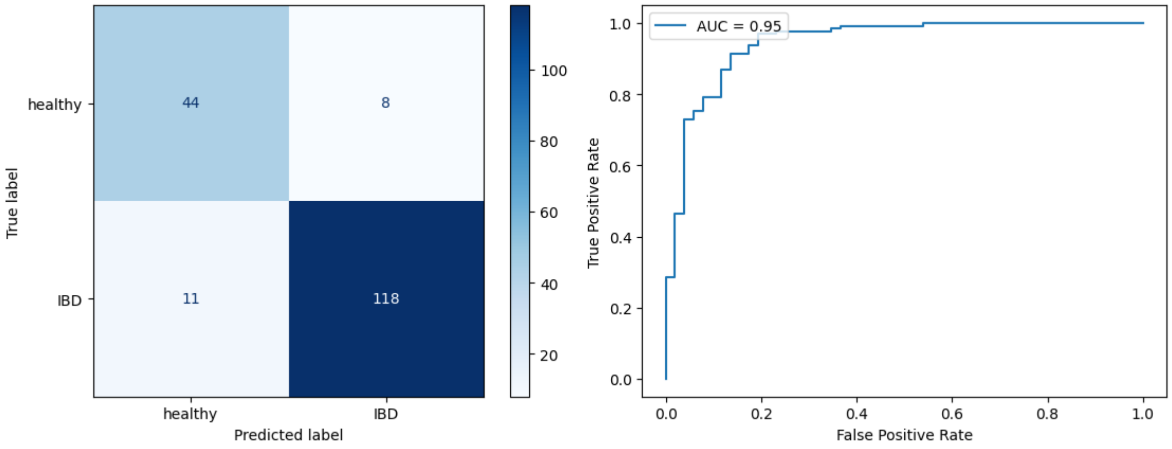
3.1 Comparison of The Data Types

Table 1 summarizes the performance of SVM classifier on both datasets. SVM and relative abundance dataset achieved higher scores for every metric except recall which was 0.94 for the marker abundance and 0.91 for the relative abundance. One important observation is that relative abundance achieved significantly higher specificity than marker abundance data. Specificity is important metric in our use case because the datasets are imbalanced: there is a lot more samples from IBD patients than healthy individuals. Therefore the classifier can achieve decent accuracy by predicting all cases as IBD. Specificity describes how accurately the model identifies healthy cases and higher specificity leads to less false alarms. Although both datasets scored high accuracy and there is no major differences in the scores, achieving higher specificity makes relative abundance dataset more accurate in identifying healthy individuals.

SVM	**Accuracy**	**Recall**	**Specificity**	**Precision**	**f1-score**	**AUC**
relative	0.88	0.91	0.85	0.94	0.93	0.95
marker	0.85	0.94	0.77	0.91	0.92	0.93

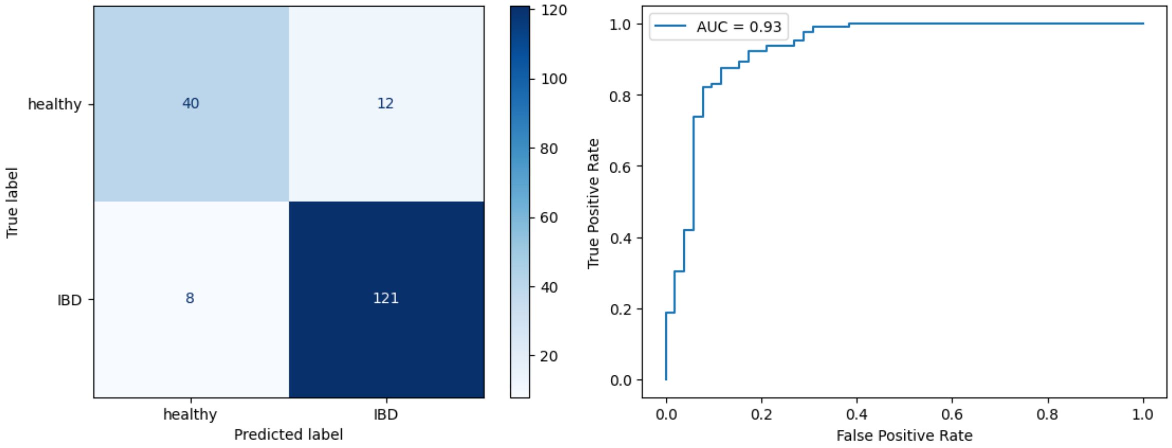
Table 1: Comparison of performance metrics for the relative and marker abundance datasets for the SVM model.

Figures 1 and 2 show confusion matrices and ROC curves for both datasets.



(a) (b)

Figure 1: (a) Confusion matrix and (b) ROC-curve and AUC-score for the SVM model and relative abundance test set.



(a) (b)

Figure 2: (a) Confusion matrix and (b) ROC-curve and AUC-score for the SVM model and marker abundance test set.

4 Conclusions

Early and accurate diagnosis of IBD is crucial for effective treatment of the disease. By analyzing changes in the composition of the gut microbiome we can observe early signs of gut dysbiosis and IBD. Our aim was to study how well machine learning models could predict IBD and what kind of data collected from stool samples would be most accurate in characterizing the disease. To achieve this goal we compared two data types using SVM classifier. Our first data type was relative abundance of microbes and the second was abundance of clade-specific marker genes.

Although there were no major differences in accuracies between the two datasets relative abundance data achieved higher specificity. In addition, the relative abundance dataset is considerably smaller and thus requires less resources and time for training. For these reasons, the conclusion for this project is that the relative abundance dataset is better for classifying healthy people and IBD patients.

In this project only SVM classifier was used in predictions. To further test the prediction performances of different metagenomic datasets more classifiers could be used and tested. Many studies using machine learning and metagenomic data have used random forests and discovered that they produce the best results. Random forest and different metagenomic data types (in addition to marker genes and relative abundance) could be used to further investigate what type of data would best characterize IBD.

References

[1] Mohammad Tauqeer Alam et al. “Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels”. In: *Gut pathogens* 12.1 (2020), pp. 1–8.

[2] Daniel N Frank et al. “Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases”. In: *Proceedings of the national academy of sciences* 104.34 (2007), pp. 13780–13785.

[3] Andrew Brantley Hall et al. “A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients”. In: *Genome medicine* 9 (2017), pp. 1–12.

[4] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. *A practical guide to support vector classification*. 2003.

[5] Junhua Li et al. “An integrated catalog of reference genes in the human gut microbiome”. In: *Nature biotechnology* 32.8 (2014), pp. 834–841.

[6] Ishan Manandhar et al. “Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases”. In: *American Journal of Physiology-Gastrointestinal and Liver Physiology* 320.3 (2021), G328–G337.

[7] Andrea Mihajlović et al. “Machine learning based metagenomic prediction of inflammatory bowel disease”. In: *PHealth 2021: Proceedings of the 18th International Conference on Wearable Micro and Nano Technologies for Personalized Health*. Vol. 285. IOS Press. 2021, p. 165.

[8] Edoardo Pasolli et al. “Accessible, curated metagenomic data through ExperimentHub”. en. In: *Nat. Methods* 14.11 (Oct. 2017), pp. 1023–1024. ISSN: 1548-7091, 1548-7105. DOI: `10.1038/nmeth.4468`.

[9] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[10] Sumei Sha et al. “The biodiversity and composition of the dominant fecal microbiota in patients with inflammatory bowel disease”. In: *Diagnostic microbiology and infectious disease* 75.3 (2013), pp. 245–251.

[11] Imogen S Stafford et al. “A Systematic Review of Artificial Intelligence and Machine Learning Applications to Inflammatory Bowel Disease, with Practical Guidelines for Interpretation”. In: *Inflammatory Bowel Diseases* 28.10 (June 2022), pp. 1573–1583. ISSN: 1078-0998. DOI: `10.1093/ibd/izac115`. eprint: `https://academic.oup.com/ibdjournal/article-pdf/28/10/1573/46284491/izac115.pdf`. URL: `https://doi.org/10.1093/ibd/izac115`.

[12] Torsten Thomas, Jack Gilbert, and Folker Meyer. "Metagenomics-a guide from sampling to data analysis". In: *Microbial informatics and experimentation* 2 (2012), pp. 1–12.

[13] Arnau Vich Vila et al. "Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome". In: *Science translational medicine* 10.472 (2018), eaap8914.