**Evaluation of classification pipelines that predict colorectal cancer progression with microbiota-associated biomarkers**

Begüm D. Topçuoğlu1, Jenna Wiens2, Patrick D. Schloss1

As gut microbiome field continues to grow, there will be an ever-increasing demand for reproducible machine learning methods to determine association of the microbiome with a continuous or categorical phenotype of interest. Currently, the use of machine learning in microbiome literature lack clarity over the training, validation and testing of the models. There is a need to properly implement good machine learning practices to generate reproducible and robust models.

Recently, there has been an interest in using machine learning to predict colorectal cancer progression with microbiota-associated biomarkers. Colorectal cancer is one of the leading cause of death among cancers in the United States. Colonoscopy as a screening tool is effective, however it is invasive, expensive and have a low rate of patient adherence. Previous studies have shown that bacterial population abundances in the stool can predict screen relevant growth in the colon and be used as a non-invasive screening tool. However, the prediction performance of these previous models vary greatly, with areas under the receiver operating characteristic curve (AUC) of 0.7-0.9 (1–4). The variation in classification performance is based in part on differences in the task definition, in part on differences in the study populations, and in part on the learning pipeline.

In this study, hemoglobin levels and bacterial population abundances in the stool were used to predict colorectal disease status of 490 patients. The colorectal disease status was defined as showing screen-relevant colonic growth or not. Training, validation and testing pipelines were established for L2-regularized Logistic Regression, L1 and L2 Linear Suppor Vector Machines (SVM), Radial Basis Function SVM, Decision Tree, Random Forest and XGBoost classifiers. L2-regularized Logistic Regression had a mean AUC of 0.68 ± 0.04, L1 Linear SVM had a mean AUC of 0.76 ± 0.05, L2 Linear SVM had a mean AUC of 0.68 ± 0.05 and Radial Basis Function SVM had a mean AUC of 0.69 ± 0.05. Decision Tree had a mean AUC of 0.7 ± 0.05, Random Forest had a mean AUC of 0.76 ± 0.06 and XGBoost had a mean AUC of 0.76 ± 0.04. Tree-based models were less susceptible to overfitting and in general had higher sensitivity and specificity for colonic screen-relevant growth. Aside from evaluating generalization and classification performance of each classifier, this study established standards for training, validation and testing of the microbiome-associated models.

## References

1. **Sze MA**, **Schloss PD**. 2018. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio **9**:e00630–18. doi:[10.1128/mBio.00630-18](https://doi.org/10.1128/mBio.00630-18).

2. **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**:37. doi:[10.1186/s13073-016-0290-3](https://doi.org/10.1186/s13073-016-0290-3).

3. **Baxter NT**, **Koumpouras CC**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2016. DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. Microbiome **4**. doi:[10.1186/s40168-016-0205-y](https://doi.org/10.1186/s40168-016-0205-y).

4. **Zackular JP**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2014. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res **7**:1112–1121. doi:[10.1158/1940-6207.CAPR-14-0129](https://doi.org/10.1158/1940-6207.CAPR-14-0129).