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

Begüm D. Topçuoğlu[1], Jenna Wiens[2], Patrick D. Schloss[1†]

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 used. There is a need to properly implement good machine learning practices to generate reproducible and robust models.

Recently, there is 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. Each person in the industrialized world has on average a one-in-twenty chance of developing colorectal cancer (CRC) (1–3). Colonoscopy as a screening tool is effective, however it is invasive, expensive and have a low rate of patient adherence. Therefore, gut microbiome-based biomarkers emerged as a non-invasive screening method.

In this study, classification models that use human hemoglobin levels and bacterial population abundances in the stool were used to predict colorectal disease status as 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. The generalization and prediction performance of these classifiers were evaluated and each classifier was examined for its reproducibility, robustness and susceptibility to overfitting. 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 overfiting and in general had higher sensitivity and specificity for colonic screen-relevant growth.

## References

1. **Howlader N KM** Noone AM. SEER cancer statistics review, 1975-2013, (national cancer institute. bethesda, md).

2. **Street W**. Colorectal cancer facts & figures 2017-2019 40.

3. **Weir HK**, **Thompson TD**, **Soman A**, **MÃžller B**, **Leadbetter S**. 2015. The past, present, and future of cancer incidence in the united states: 1975 through 2020. Cancer **121**:1827–1837. doi:10.1002/cncr.29258.