# A Machine Learning Based Diagnostic Tool for the Early Detection of Colorectal Cancer

Kothapalli, Abihith (School: Blue Valley West High School)

Colorectal cancer (CRC) is not only one of the most common non-sex-specific cancers, but also one of the most deadly. Early diagnosis and treatment, however, can significantly improve patient prognosis, prompting the need to develop more comprehensive and less invasive methods to diagnose CRC in its early stages. Previous studies have established the role of the human gut microbiome in CRC carcinogenesis and progression. In this study, the efficacy of gut microbiome data in detecting CRC was investigated. This was done using six publicly available datasets, containing microbial abundance data from the gut microbiomes of 609 patients. Utilizing robust feature selection methods, a total of 121 potential biomarkers for CRC were identified and were subsequently used to develop machine learning models for the detection of CRC. To evaluate the predictive capabilities of these models, the area under the receiver operating characteristic curve (AUC) and the accuracy on testing data was calculated. The top performing model in this study was a random forest model, obtaining an AUC of 0.9238 and an accuracy of 90.16%. Ultimately, this paper demonstrates the viability of metagenomics data in machine learning to enable the development of a diagnostic tool for the early detection of CRC, facilitating improvements in both treatments and patient prognosis. Moreover, this study represents one of the largest meta-analyses of metagenomic data in the context of CRC performed to date. In the future, further investigation of the relationships between the biomarkers identified in this study and CRC pathogenesis could aid in understanding the etiology of CRC and gaining insight into potential therapeutic targets for CRC.