A Novel Machine Learning Based Identification Tool (ELECT) for Early Colorectal Cancer Detection through Advanced Microbiome Composition Analysis

Ma, Alan (School: Jesuit High School)

Colorectal cancer(CRC) ranks third in occurrence and second in mortality among all cancers. Current CRC identification methods are often ineffective due to the invasiveness of such procedures and long waiting times for test results. Most CRC cases are identified in late stage which has a drastically low 14% 5-year survival rate(5ySR). However, if found at an early stage, the 5ySR of CRC cases is around 90%. Thus, detecting cancer early is crucial to preventing CRC deaths. This project goal is to accurately detect CRC early on and identify high-correlation cancer biomarkers. It utilizes Elastic-Net regression machine learning model to predict cancerous patients based on microbiome bacterial samples. The model was trained, validated, and tested on a set of over 1.5 million unique gut bacterial samples for robustness. Incremental hyperparameter tuning and feature selection were simultaneously run to select the best performing model and increase correlation score. After cross-validation, the final model was able to predict CRC with an accuracy greater than the current industry best by over 9%. Model results in statistical cluster plots and heatmaps further demonstrate precise and accurate predictions. In addition, this model significantly reduced the dataset size needed by 99% - shrinking the initial pool of 5207 bacteria types to 43 of the most critical bacteria biomarkers. This research allows oncologists to quickly and accurately identify CRC patients early on in a less invasive manner yielding a greater survivability. Project future work will focus on CRC recurrence identification and piloting in clinical trials.

Awards Won: