Name(s)
Cynthia Chen

Project Number
S0805

Project Title
Type 2 Diabetes Prediction Using Longitudinal Machine Learning Analyses and Integrative Personal Omics Profiling

Abstract

Objectives/Goals
Type 2 Diabetes (T2D) affects more than 200 million people worldwide. Steady-state plasma glucose (SSPG) values are crucial to determining T2D as they indicate a patient's insulin resistance. The goal of our project was to analyze large longitudinal omics datasets using machine learning and statistical analyses in order to accurately predict SSPG values for pre-diabetic patients.

Methods/Materials
Our dataset consists of SSPG values, BMI, and 3500 omics features for 23 patients at 4 different time-points. We preprocessed our data in three stages: 1) PCA dimensionality reduction for feature selection, 2) taking derivatives between consecutive time-points to preserve the longitudinal time sequence, and 3) data normalization and standardization.

For SSPG value prediction using machine learning, we developed 10 different classification algorithms and 5 regression algorithms in Python, and tested these algorithms for optimal performance. We also generated correlation matrices among the omics datasets and determined the most correlated feature pairs as well as the optimal microbiome taxonomy depth levels. Using these feature correlation analyses, we improved the prediction performance of the regression and classification models.

Results
AdaBoost classification achieved an accuracy rate of 87.5%, and LASSO regression performed the best with a root mean square error (RMSE) of 24.765. We improved these results to 90.0% accuracy and 22.455 RMSE by using the feature correlation analyses described above.

Conclusions/Discussion
We concluded that our computational model was successful in accurately predicting T2D for pre-diabetic patients. Our project has major implications in the medical field, as our novel longitudinal time sequencing and feature correlation methods can provide new, improved pathways for disease prediction. In the future, we would like to expand upon our project by analyzing more datasets, such as genomics and RNA sequencing.

Summary Statement
We created an accurate computational model using machine learning methods and statistical analyses to predict SSPG values for pre-diabetic patients.

Help Received
I worked at the Stanford Laboratory of Quantitative Imaging under the guidance of Dr. Imon Banerjee and Prof. Daniel Rubin. I developed the methods and algorithms independently and received help from my mentors.