Name(s)
Lyron O. Co Ting Keh

Project Number
38696

Project Title
A Novel Hierarchical Machine Learning Model for Non-Invasive Cancer of Unknown Primary Classification

Abstract

Objectives/Goals
Cancer of unknown primary (CUP) is the 4th leading cause of cancer-related deaths worldwide. Patient prognoses can be significantly improved with site-specific therapy. Thus, the objective of this project is to design and train a reliable and cost-effective model to carry out non-invasive tissue-of-origin classification.

Methods/Materials
The chimeric nature of cell-free DNA (cfDNA) in the bloodstream disrupts intraclass distributions and interclass independence assumptions and is subject to low signal-to-noise ratios. This model design is focused on alleviating these challenges with a hierarchical ensemble framework consisting of 3 Support Vector Classifiers (SVCs) and 4 Epsilon Support Vector Regression (SVR) predictors. Each model was trained with interpolated data from 3598 solid tumor profiles and 299 whole blood profiles at various tumor fractions. I conducted Recursive Feature Elimination (RFE) with a variety of random seeds and subsamples to produce a stable reduced feature set and a Grid Search with 5-Fold Cross Validation to tune hyperparameters. These models were integrated into a pipeline that employs the SVRs to infer the tumor fraction of a sample and then feeds the data into the SVCs to restore previously violated assumptions and produce a tissue-of-origin prediction.

Results
This multi-level model predicted tissue-of-origin with 96% accuracy on a withheld test set (n=525), corresponding to a 20% improvement from current non-invasive methods. Furthermore, this pipeline is able to achieve 82% accuracy in the context of early-detection and classification between 6 primary sites, whereas current screening methods only target single cancer types. This performance is maintained when the model is restricted to only 0.8% of the raw feature set, drastically reducing the computational and monetary costs of the test.

Conclusions/Discussion
The results demonstrate the capabilities of this hierarchical approach over standard single-level models when dealing with highly convoluted data with underlying structures, such as cfDNA profiles. This has important implications in treatment decision-making in CUP and in the development of early detection assays. In addition, the novel design and training technique presented in this study can be applied to other problems involving cfDNA such as monitoring cancer progression and observing treatment response to enhance the non-invasive examination of tumors.

Summary Statement
I designed a hierarchical machine learning model that utilizes a simple blood sample to affordably classify cancer of unknown primary with higher efficacy than published methods.

Help Received
Joanne Soo, Dr. David Kurtz, and Dr. Ash Alizadeh from Stanford provided me with guidance on the biological aspects of the project and best-practices when validating my methodology.