CMM Project Proposal

The objective of this project is to develop a predictor for distinguishing patients with superficial T1 cancers who later progress to a more lethal, muscle invasive stage from those doesn't based on gene expression (information on sex and age is not incorporated in this project). The general steps are outlined as follows.

1. Data processing: There are three datasets containing gene expression values and corresponding phenotype labels across two platforms. The gene expression values are normalized to z-score by distracting the global mean and dividing the global standard deviation for each gene. This project will only use the information on the 9,584 genes that all three datasets have in common. Those common genes are extracted from three datasets and ordered alphabetically, then combined into one new dataset. All the samples will be randomly split into pieces of training and testing sets (k-fold cross validation, k is initially set to be 5, might need adjustment later as only a small portion of progression class present in the dataset).

2. Filter out the most differentially expressed genes based on the Wilcoxon-rank sum test using the training dataset. 10 vs. 100 vs. 1000 genes will be obtained to evaluate the effect of the level of filtering. For each order of magnitude, train a predictor.

3. Quadratic discriminant analysis (QDA) is chosen for this project (might subject to change later on). Other techniques such as kNN and random forest will be tested out as well. Performances will be compared (overall accuracy, sensitivity and specificity).

4. The best performed approach will be used for constructing a ROC using the testing set. Find the t_{80} point.