Dive into Prostate Cancer:
Influence of Gene Expression Levels on Clinical Diagnosis Results

Minglei Cai, Qixuan Guo, Zhonghao Xue

1. Introduction

Prostate cancer is the most common non-dermatological cancer in the United States. Early diagnosis provides an opportunity for curative surgery. However, up to 30% of men undergoing radical prostatectomy will relapse. The challenge is to identify those patients at risk for relapse and to better understand the molecular abnormalities that define tumors at risk for relapse.

Recently, genomic methodologies have been used to discover consistent gene expression patterns associated with a give histological or clinical phenotype. [1]

2. Background

2.1 About the Data

The dataset is composed of information of 102 patients. The predictors are levels of 6033 gene expressions, which are all continuous variables; the response is the clinical prostate cancer diagnosis result, which is a binary variable. Since features is much more than samples, the famous 'curse of high dimensionality' does exist for this dataset.

2.2 Objectives

We focus on the following two problems:

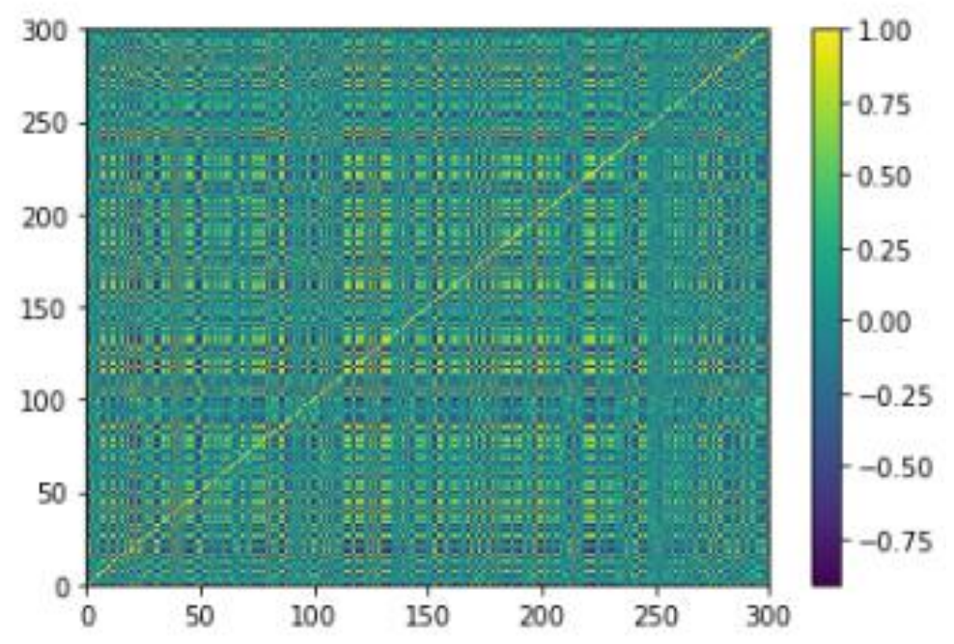
(1) Build models using selected predictors with great performance.

(2) Find the significant genes whose expression levels are influence the diagnosis result.

3. EDA

3.1 Correlations

In the very beginning, it's good to notice that correlations between gene expressions are common.

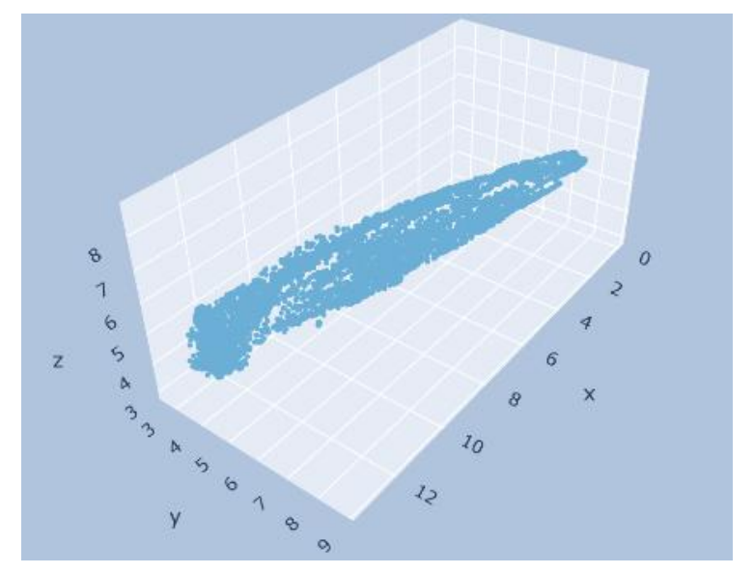


Correlation Matrix of 300 Randomly Picked Genes

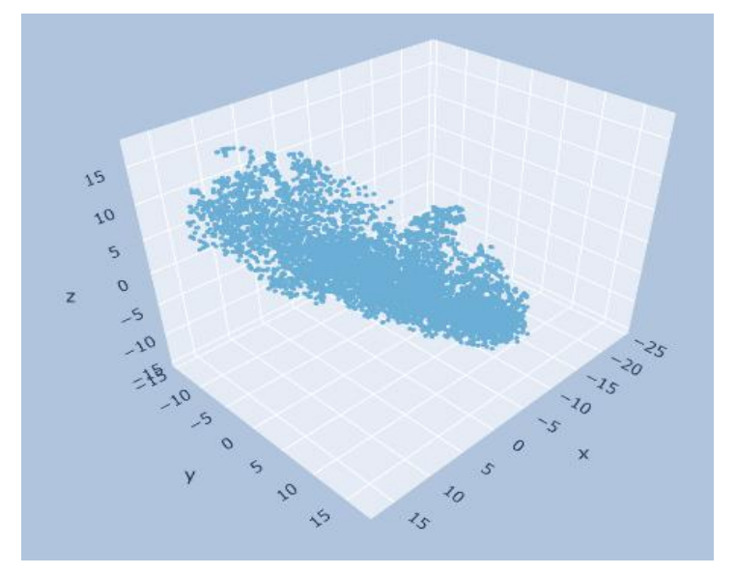
3.2 UMAP & TSNE

UMAP & TSNE could be used to transform the original data close to a manifold in a low dimension.

UMAP



TSNE



4. Analysis

We divide our further analysis into 3 parts: variable selection, model creation and model evaluation.

4.1 Variable Selection

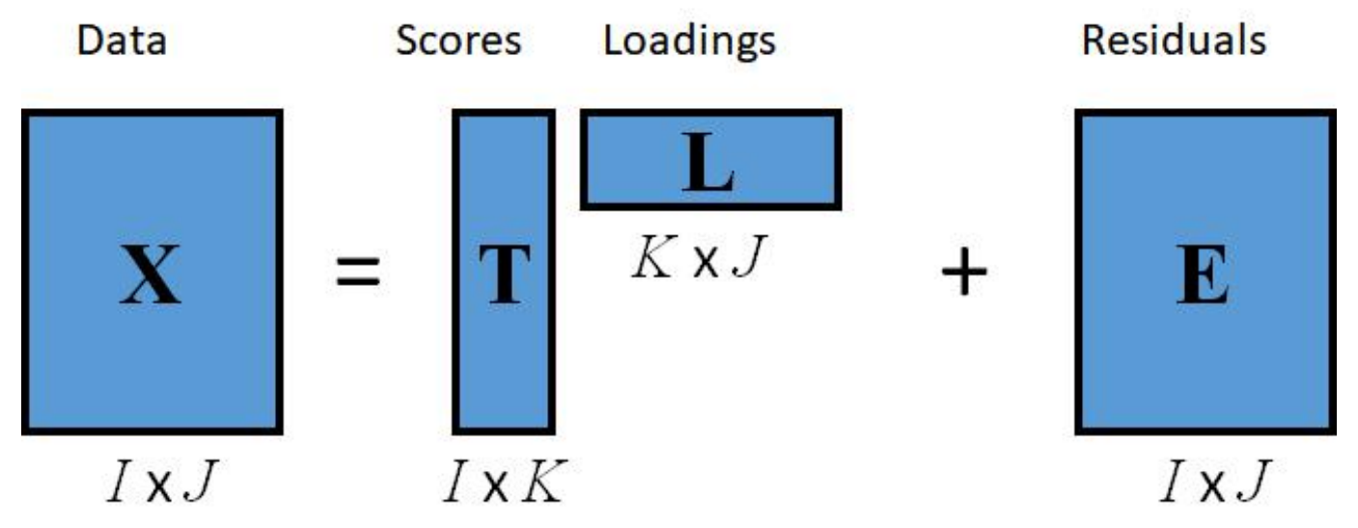
In this part, we try different methods to reduce the number of predictors.

4.1.1 Forward Selection

Basically forward selection could be used to select the predictor set using logistic regression penalty, with an AIC/BIC term or a 1st/2nd-norm regularized term.

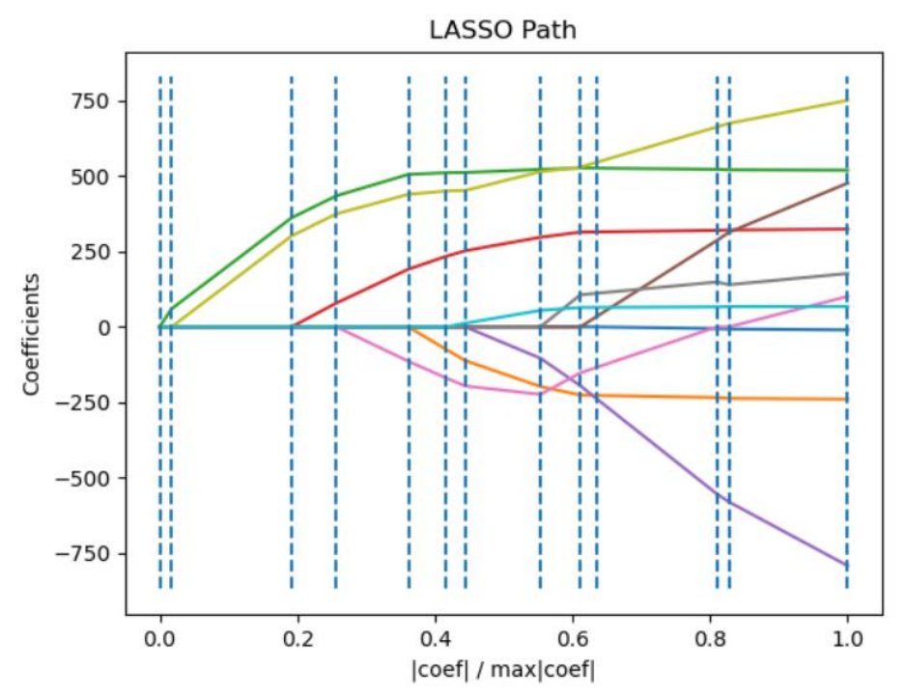
4.1.2 PCA

By operating SVD for the design matrix and ignoring the small singular values, we can achieve a PCA model which reduces the features of original data.



4.1.3 LASSO Path

Adjustments of the penalized coefficient could lead to changes of number of predictors, which could be used in model selection.



4.1.4 Chatterjee Correlation

Chatterjee Correlation is defined between two vectors of the same length, aiming to detect whether there is a functional relationship between the two vectors. See more details about Chatterjee correlation in the paper[2].

In this context, Chatterjee correlations could be computed between the diagnosis result column and each gene expression column. We select the most response-correlated genes as predictors according to the Chatterjee correlations for model creation.

$$\xi_n(X,Y) := 1 - \frac{n\sum_{i=1}^{n-1}|r_{i+1}-r_i|}{2\sum_{i=1}^{n} l_i(n-l_i)}.$$

$$\xi(X,Y) := \frac{\int \mathrm{Var}(\mathbb{E}(1_{\{Y\geq t\}}|X))d\mu(t)}{\int \mathrm{Var}(1_{\{Y\geq t\}})d\mu(t)},$$

4.2 Model Creation

In this part, we try different models using the selected variables.

4.2.1 Logistic Regression

A basic method to predict the classes.

$$\min_{w,c} \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1)$$

4.2.2 Penalized Logistic Regression

Improved version of logistic regression with a regularized term.

L1 $$\min_{w,c} \|w\|_1 + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

L2 $$\min_{w,c} \frac{1}{2}w^T w + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

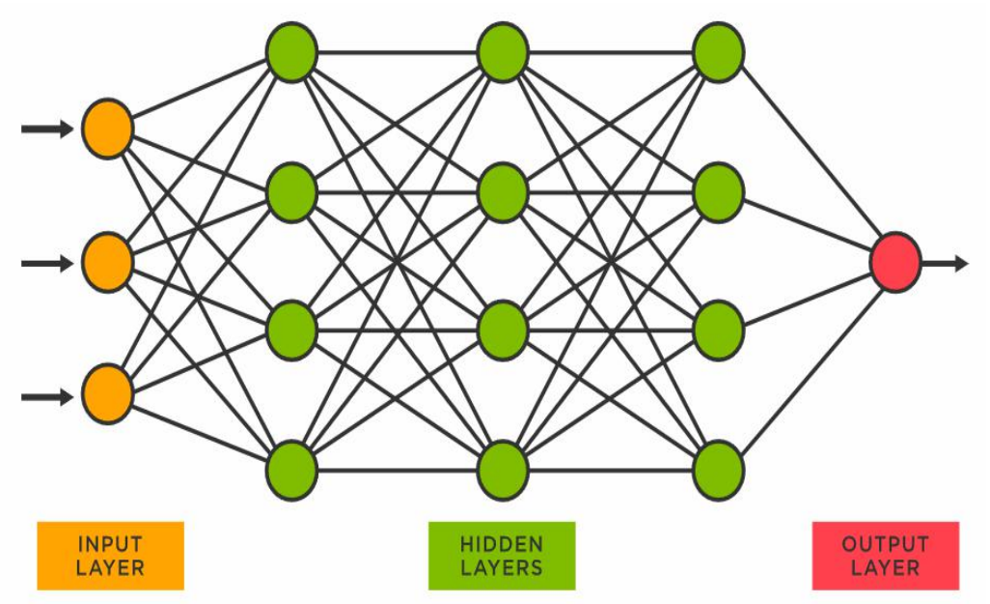
4.2.3 Graphical LASSO & LDA/QDA

While LDA/QDA could be used to predict the class, large errors emerge with high-dimensional data if we compute the precision matrix directly as the inverse of covariance matrix. We use Graphical LASSO to predict the precision matrix.

$$\begin{aligned}\log P(y=k|x) &= \log P(x|y=k) + \log P(y=k) + Cst \\ &= -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k) + \log P(y=k) + Cst,\end{aligned}$$

4.2.4 Neural Network

Neural network is proved to be a good approximation to a large group of functions when the number of layers is large enough. We use sequential neural networks here for prediction.



4.3 Model Evaluation

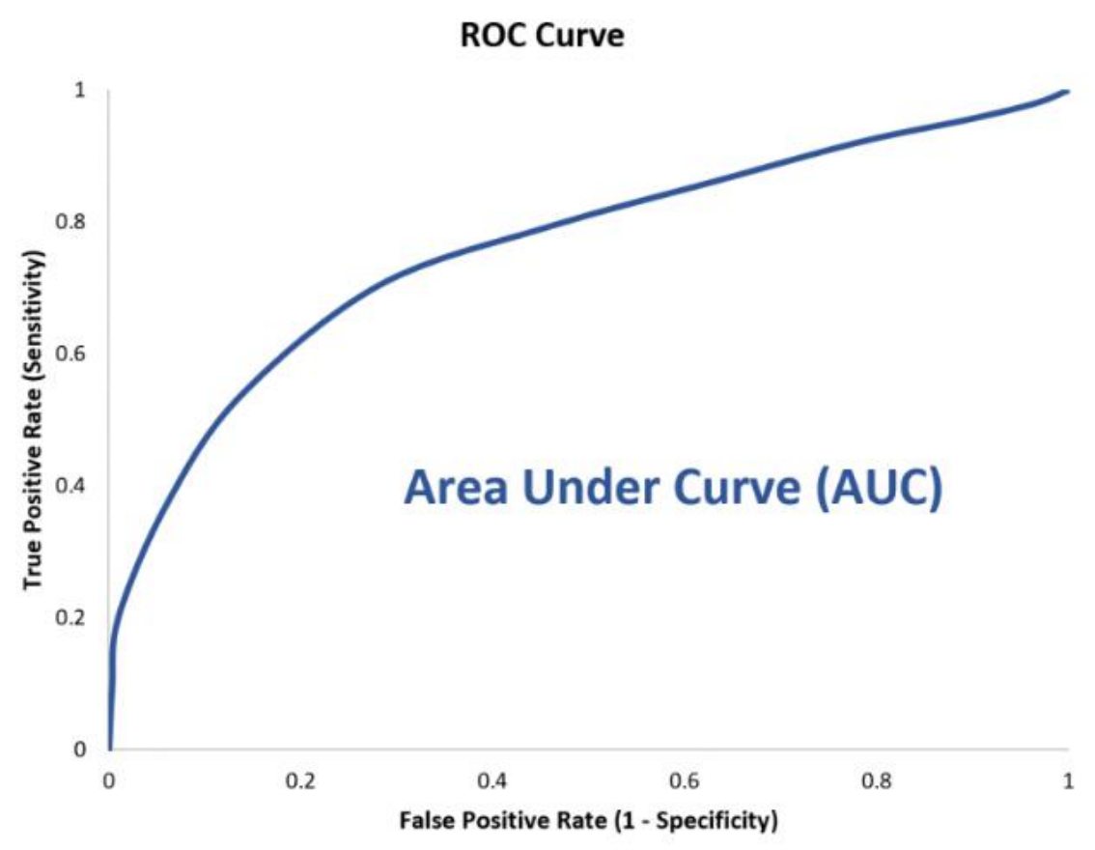
In this part, we use cross-validation and ROC curve to evaluate our models.

4.3.1 Cross-Validation

We use 5-fold cross-validation to compute the probability of both classes for the dataset.

4.3.2 ROC Curve

With the TPR-FPR curve, we have a comprehensive evaluation for the models.



5. Results

5.1 ROC Curves

We built different models after reducing variables to 20 using different methods.

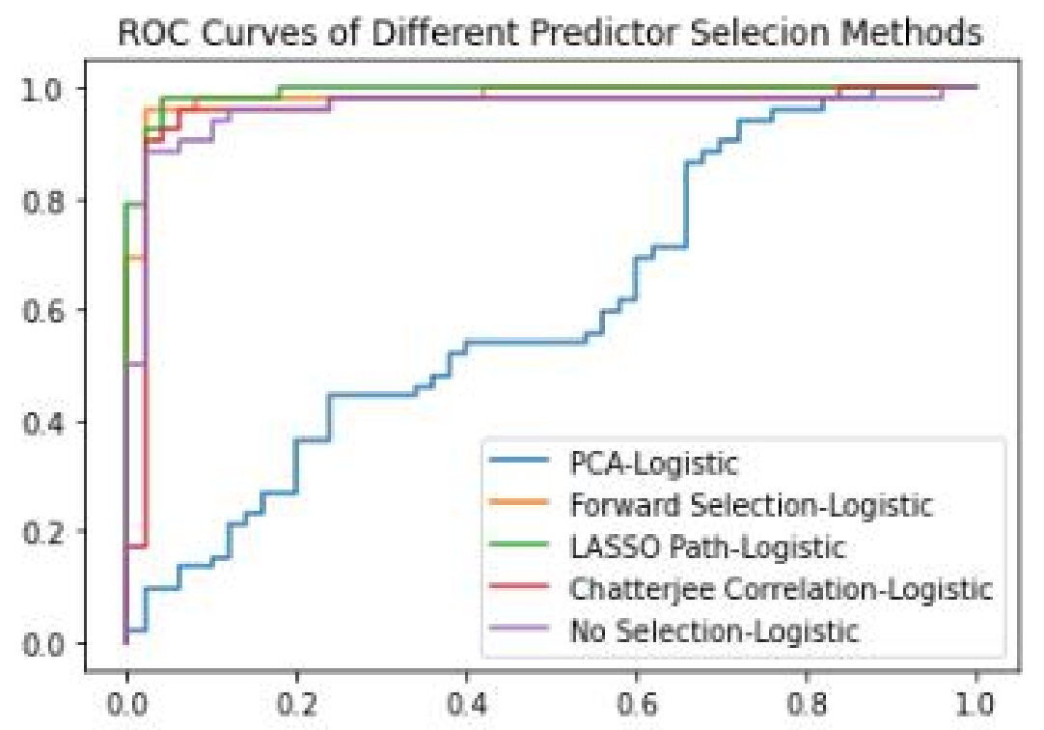
5.1.1 Different Ways of Variable Selection

AUC Performance:

Forward-Selection≈LASSO Path≈Chatterjee Correlation>>PCA

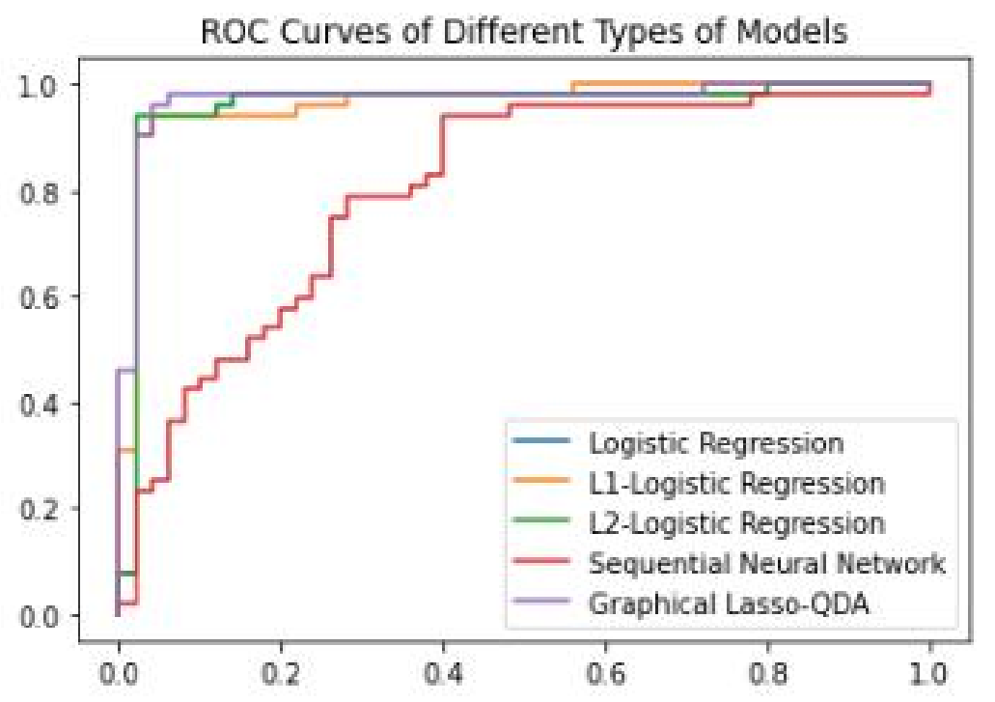
Time Complexity Sequence:

Forward Selection>>LASSO Path>Chatterjee Correlation



5.1.2 Different Types of Model

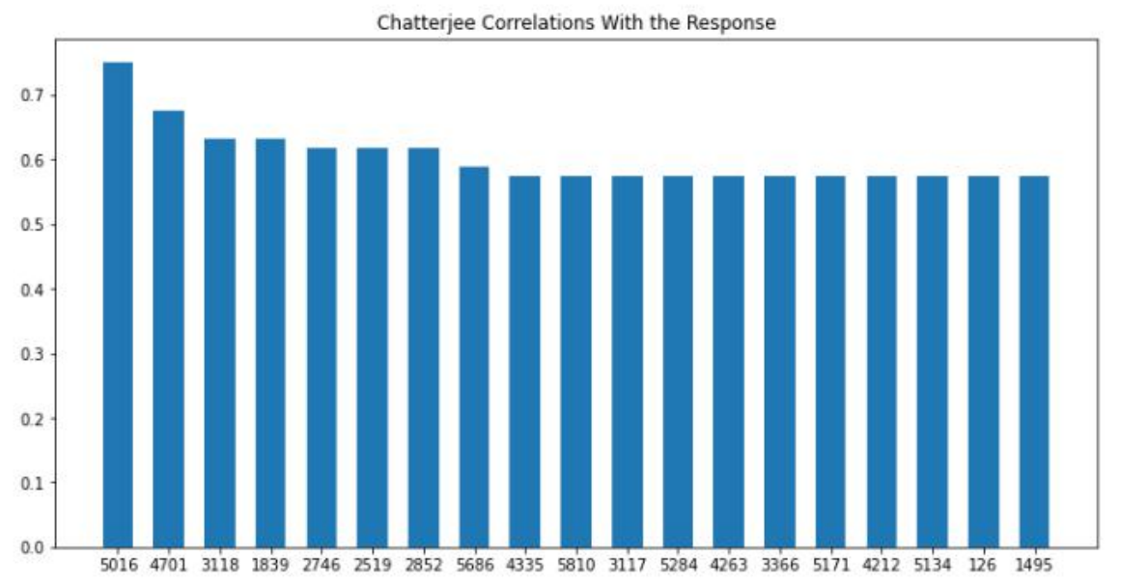
While scarcity of samples becomes an issue for the neural network. Other methods performs similarly.



5.2 Significant Features

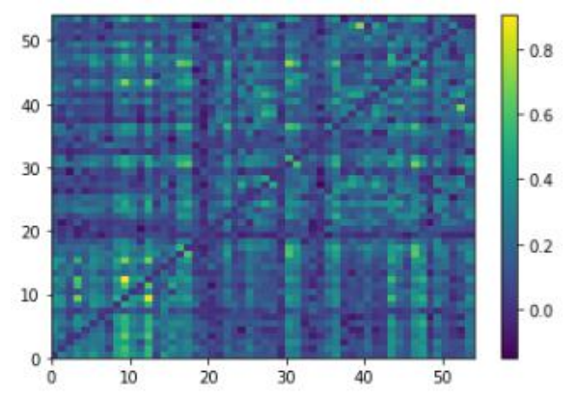
5.2.1 The Most Significant Gene

The intersection of three variable sets computed by Forward-Selection, Lasso-path and Chatterjee Correlation only contains one element, the 5016th gene.



5.2.1 Correlations in Significant Gene Expressions

The reason for the difference of three predictor sets may be the correlations between significant variables.



6. Conclusions

(1) The most significant gene is the 5016th gene.

(2) We can predict prostate cancer according to combinations of gene expression levels.

7. Codes & References

7.1 Codes (Github Link)

https://github.com/kianakaslana648/ANLY512-FINAL-PROJECT

7.2 References

[1] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002 Mar;1(2):203-9. doi: 10.1016/s1535-6108(02)00030-2. PMID: 12086878.

[2] Chatterjee, S. (2019). A new coefficient of correlation. arXiv e-prints, page. arXiv preprint arXiv:1909.10140, 711.