Johns Hopkins University, AMS Department

Risk genes to residual tumor of breast cancer patients

CMM final project

Meng Chen

Introduction

After skin cancer, breast cancer is the most common cancer diagnosed in women in the United States.[1] Breast cancer happens when cells in your breast grow and divide in an uncontrolled way, creating a mass of tissue called a tumor.[2] Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized.[3] Studies have shown that your risk for breast cancer is due to a combination of factors. The main factors that influence your risk include being a woman and getting older. Most breast cancers are found in women who are 50 years old or older.[4] If you have a strong family history of breast cancer or inherited changes in your BRCA1 and BRCA2 genes, you may have a high risk of getting breast cancer.[5]

Residual cancer burden is estimated from routine pathologic sections of the primary breast tumor site and the regional lymph nodes.[6] Residual cancer burden after neoadjuvant chemotherapy can accurately predict disease recurrence and survival across all breast cancer subtypes.[7]

Fine needle aspiration

Fine needle aspiration (FNA) uses a very thin, hollow needle attached to a syringe to take out a small amount of fluid and very small pieces of tissue from the tumor. The doctor can aim the needle while feeling the tumor, if it's near the surface of the body. If the tumor is deeper inside the body and can't be felt, the needle can be guided while being watched on an imaging test such as an ultrasound or CT scan.[8]

Core biopsy

Needles used in a core biopsy are slightly larger than those used in FNA. They remove a small cylinder of tissue (about 1/16 inch in diameter and 1/2 inch long). The core needle biopsy is done with local anesthesia (drugs are used to make the area numb) in the doctor's office or clinic. Like FNA, a core biopsy can sample tumors that the doctor can feel as well as smaller ones that must be seen using imaging tests.[9]

Gene samples of patients with breast cancer were taken by either FNA(Fine needle aspiration) or core biopsy from primary tumor before they were taking chemotherapy. Patients with no residual tumor in the breast and the axillary lymph node are labeled as pathological complete response (pCR). While patients with a residual disease in either the breast, axilla or both are labeled as residual disease (RD).

Our dataset includes 139 patients and 13299 genes. Values of genes were normalized and log2-scaled. There is no missing value in the dataset. The ratio of RD over pCR is approximately 11:3 (figure 1). Our goal is to find a compact set of genes which is related to the response and is able to

---

[1] https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[2] https://my.clevelandclinic.org/health/diseases/3986-breast-cancer

[3] https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

[4] https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm

[5] https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm

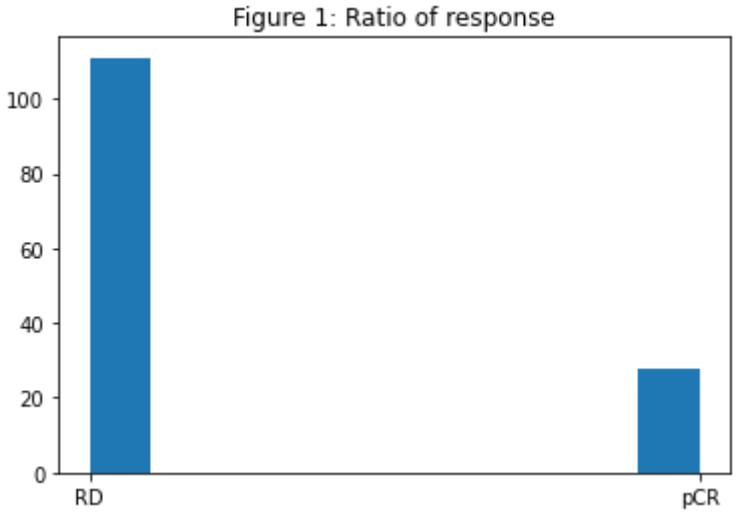
[6] https://ascopubs.org/doi/10.1200/jco.2007.10.6823

[7] https://ascopost.com/issues/january-25-2020/residual-cancer-burden-is-prognostic-of-outcomes-across-breast-cancer-subtypes/

[8] https://www.cancer.org/treatment/understanding-your-diagnosis/tests/testing-biopsy-and-cytology-specimens-for-cancer/biopsy-types.html

[9] https://www.cancer.org/treatment/understanding-your-diagnosis/tests/testing-biopsy-and-cytology-specimens-for-cancer/biopsy-types.html

predict the residual disease.

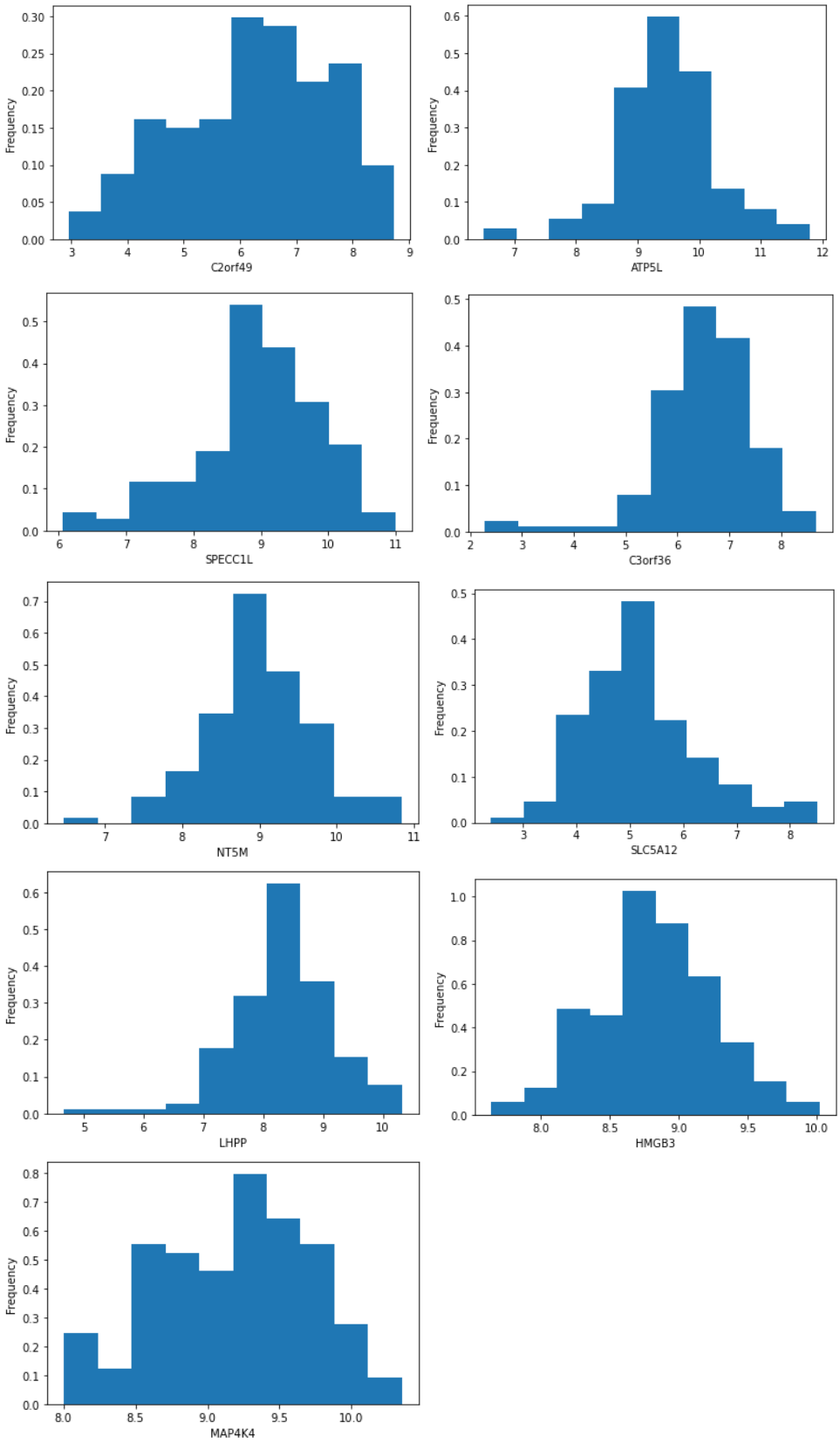


Methods

Our first step is to reduce the dimension of genes to find sets of genes which are most likely to predict the response. To achieve this, we use three methods which are joint mutual information, Pearson Correlation and Wilcoxson Rank-Sum Test. Specifically, we want to leave 20 most informative genes, so we find three sets of genes A, B and C among 13299 genes. Genes in A have 20 highest joint mutual information of genes and the response. Genes in B have 20 highest Pearson Correlation to the response. Genes in C have 20 smallest P-values of Wilcoxson Rank-Sum test with the null hypothesis values of a gene related to two labels are equally spread. In the second step, we split our data into training and test set with proportion 8:2. Notice that our data is greatly imbalanced, we keep the proportion same for two labels of response in training and test set. In addition, we can also implement weighted methods to eliminate the effect of imbalance. Next step is to use different machine learning methods, particularly, Logistics Regression, SVM, LDA, QDA, Random Forest and Artificial Neural Network, to train and predict using A, B, C three sets. We try to find the best methods and gene set based on ROC curve and AUC (Area under curve), since the data is imbalanced AUC instead of accuracy score is an appropriate criterion for the prediction.

Preprocess

We randomly select 9 genes in the total set and plot the distribution of them. We can see that they are all approximately normal shaped and roughly range in the interval from about 5 to 10. This could be the result of log-scale.



The genes in A, B and C three sets are showed bellowed. We can see every set is entirely exclusive to each other, which means there is no agreement made by three criterions. Although, joint mutual information, Pearson Correlation and Wilcoxson Rank-Sum Test all measure the relationship

between genes and the response, the imbalanced proportion of sample size and genes size results from the vast amount of genes leads to great deviation of results measured by them. To examine which set of genes best separate and predict the response, we need to resort to other methods.

12414	FRMD1
18	PLD1
39	FAU
42	RPL18
383	HIF1A
386	IPO7
570	SDC1
762	TMX4
769	ANXA2
1098	MX1
1159	VEZF1
1285	TRIM2
1389	AFG3L2
1586	CASP3
1813	MYO1E
1891	TFAM
1926	CLK2
2087	EML3
2102	VPS16
2204	DYNC1LI2

Gene set A

4046	SPIB
3369	CYP2J2
13108	TJAP1
3498	SURF2
10733	RNF25
3902	SDS
10386	ENY2
10836	COL5A3
713	PFDN1
3058	ATP2B2
245	IQGAP1
2967	ABCG1
3502	RBMS2
1794	TTC37
2100	STX7
436	SNU13
1658	CNPY2
10295	CARHSP1
7951	SETD2
1407	TNFAIP2

Gene set B

9777	TRIM44
5060	LOC101928635///ALDH1A2
8188	MISP
9778	ADI1
8203	NCAPH
556	NUDCD3
564	LOC101926921///DAB2
8190	TSPYL4
8260	NTAN1
12028	PAQR5
8180	THAP11
11823	MCUR1
8175	LRRC47
8227	MZT2A///MZT2B///PHGDH
9788	NDUFA4
8194	MCF2L
2312	BAK1
9746	EIF3L
9767	COPG1
568	APEH

Gene set C

Next we will implement them separately to 6 machine learning methods. But first we can obtain some rough idea based on the scatter plot for the first 2 genes in each set with respect to two labels(Fig 2,3,4). We notice that clear separation can only be observed in set B, while scatters in other two plots are similarly spread out, which implies that Pearson Correlation might be the best separation criteria for this gene sets.

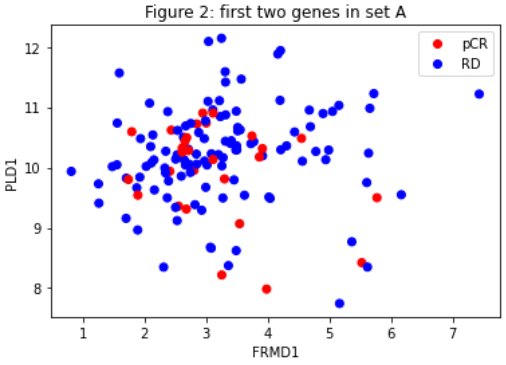


Figure 2: first two genes in set A

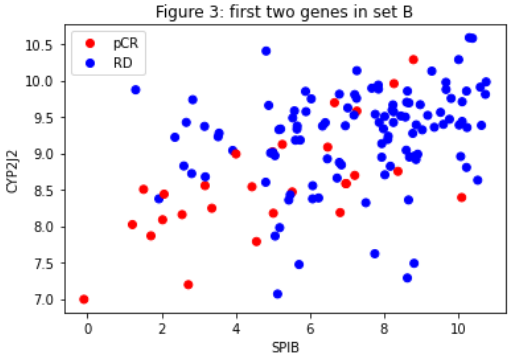
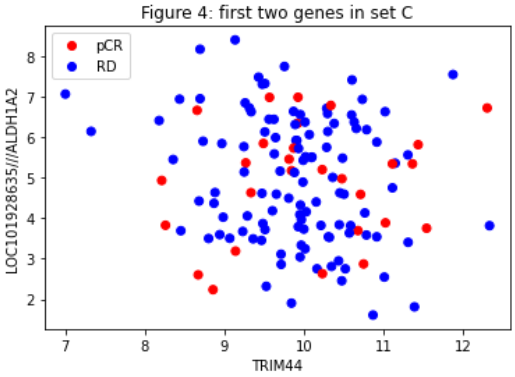
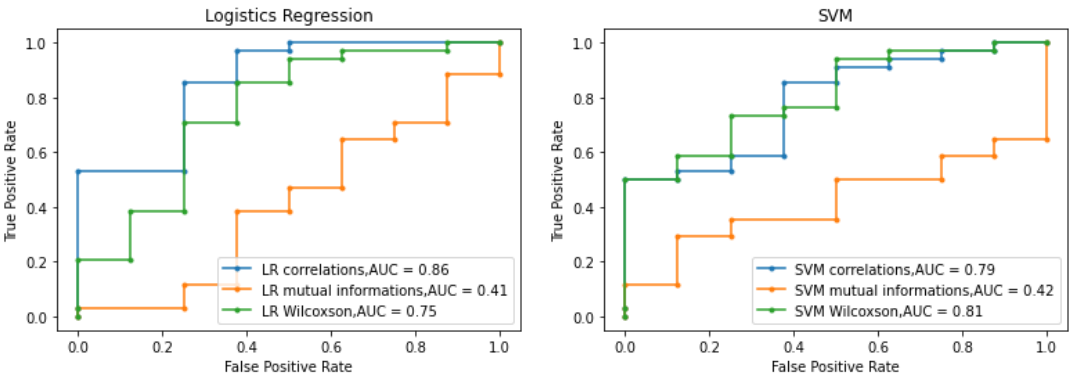


Figure 3: first two genes in set B



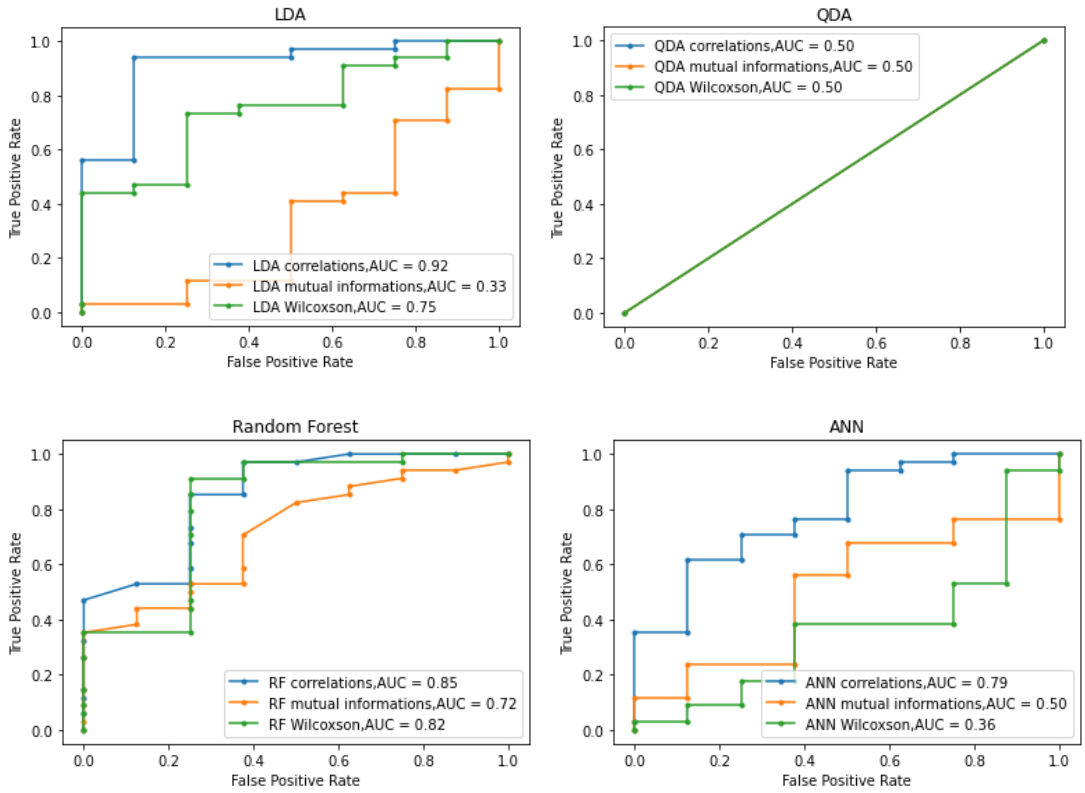
Results

The corresponding ROC curve and AUC are showed below. We find that the average AUC for prediction with set A is 0.476, set B is 0.842 and set C is 0.698. Therefor the best set for prediction is set B which is 20 genes with highest Pearson Correlation, which conforms to our speculation. We notice that joint mutual information doesn't seem to be a good criteria since mutual information might lead to selection of redundant and irrelevant features.[10] Among all machine learning methods, excluding set A, Random Forest and LDA has highest average AUC 0.835, then is Logistics Regression 0.805, SVM 0.8, ANN 0.575. We notice that QDA fails to predict because of some genes are collinearly. The computation in QDA involves the inverse of design matrix like linear regression. Thus the collinearity leads to an inaccuracy in computing inverse of design matrix. The inaccuracy of ANN to the prediction with set C might attribute to the inappropriate design of network structures which is unavailable.[11]

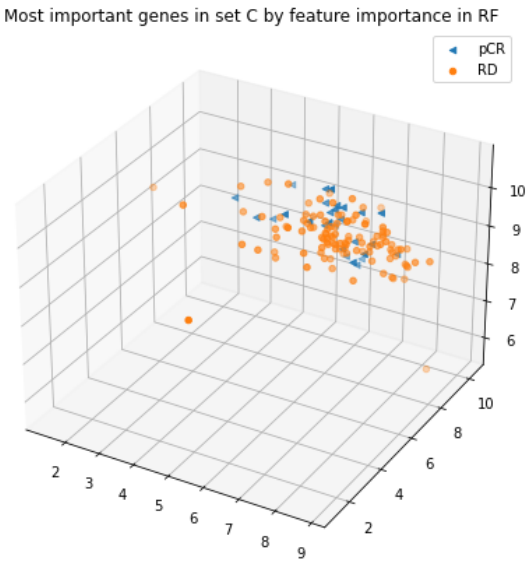


[10] https://www.sciencedirect.com/science/article/pii/S0957417415004674

[11] https://www.asquero.com/article/advantages-and-disadvantages-of-artificial-neural-networks/



We can also notice from the above result that Random Forest is the only way with fair results in all three sets. There is a criterion in Random Forest called future importance which measures the percentage importance of features. For the three gene sets A, B and C, the top 3 most important genes in set A are: IPO7, MYO1E and CLK2, with importance 0.154, 0.106 and 0.064. Top 3 most important genes in set B are: ATP2B2, RNF25 and SDS, with importance 0.096, 0.090 and 0.084. Top 3 most important genes in set C are: THAP11, NUDCD3 and ADI1, with importance 0.111, 0.104 and 0.068. Those top 3 genes are not intersecting with the pairs of genes we obtained from sorting in A, B and C. One thing we notice is that importance for genes in set B are more equally spread, however, in other two sets are converging to the top 2 genes. We also plot three similar but in 3-D scatter plots. We can see that still, genes in set B are most separable respect to labels in response. But we can see a little separation in set A as well as dots labeled as 'RD' are smaller in one axis.



Conclusion

From our results, we conclude that LDA trained by 20 genes with highest Pearson Correlations to residual disease can do the best prediction whose AUC reaches 0.92. This could be the result of normality of genes and they also have approximately same variance as showed in the Chapter Preprocess, which perfectly fit the assumptions or LDA. There is also one contradiction with respect to the inaccuracy prediction by QDA. The warning of Python indicates "Variables are collinear", however, LDA and Logistics Regression which also involve inverse of design matrix don't pop out this warning. This might be the result of underlying functions of Python in those methods, however no exact explanation was found. There is also limitation in this analysis. CNN instead of ANN could be implemented to this data, where CNN tends to be a more powerful and accurate way of solving classification problems.[12]In addition, by using pretrained CNNs, we don't need to adjust the structure of network by trials and they are more sophisticated built.

[12] https://viso.ai/deep-learning/ann-and-cnn-analyzing-differences-and-similarities/