Machine Learning On Gene Expression Data For Pathologic Stage Classification Of Breast Cancer

Chen Wang, Jianyi Zhang

HIDS-506 final project
2020/05/07

01
Background

Overall context

Topic introduction

The analysis is based on classification, by detecting the differences of gene expression among various genes simultaneously, thus predict pathologic stages of breast cancer based on those features.

Background

“Breast cancer” is increasingly understood as an umbrella designation for various tumor subtypes and stages that differ in their prognoses and responses to therapy *(Marchionni, L., 2008).* We sought to investigate if gene expression in breast cancers contains information about pathologic staging features of the disease in patients.

Challenges

In the past, breast cancer diagnosis, prognosis, and treatment decisions were based on clinical-pathological analysis of the breast cancer tissue and axillary lymph nodes *(Reis-Filho, 2011)* Those features could sometimes imperfectly predict clinical outcomes and could result in excessive treatment of many patients with chemotherapy for marginal benefit.

Data from gene expression arrays hold an enormous amount of biological information *(Tarca, A. L., 2006).* Quantitative measurement of gene expression can potentially provide clues about the mechanisms of gene regulation and interaction, and at the abstract level about biochemical pathways and the cellular function of a cell *(Callari, M., 2018).*

So, methods based on gene expression difference between individuals are needed to further understand of breast cancer in order to optimize and individualize breast cancer treatment.

Significance

In diagnosis and treatment

Comparison between genes expressed in different stages of breast cancer will further our understanding in the disease pathology, and also help to identify genes or groups of genes as targets for potential therapeutic intervention.

In wider healthcare scope

Our project, along with larger data scales and clinical technologies, could be applied to the application of early cancer detection and personalized treatment with precision on gene level. We can also use our model to look back to see what genes have most influence on stages of breast cancer, thus validate the biomedical findings. This would help to better understand the gene expression difference of the disease in molecular level.

02

Method: Dataset

Dataset Source

Source

The data is generated from Invasive breast cancer (BRCA) of Cancer Genome Atlas (TCGA)

Count

The total number of observations is 1231

Attributes

Gene expression of 14409 genes and clinical information include gender, race, ethnicity, age at initial pathologic diagnosis, breast carcinoma estrogen receptor status, histological type.

Dataset Labels

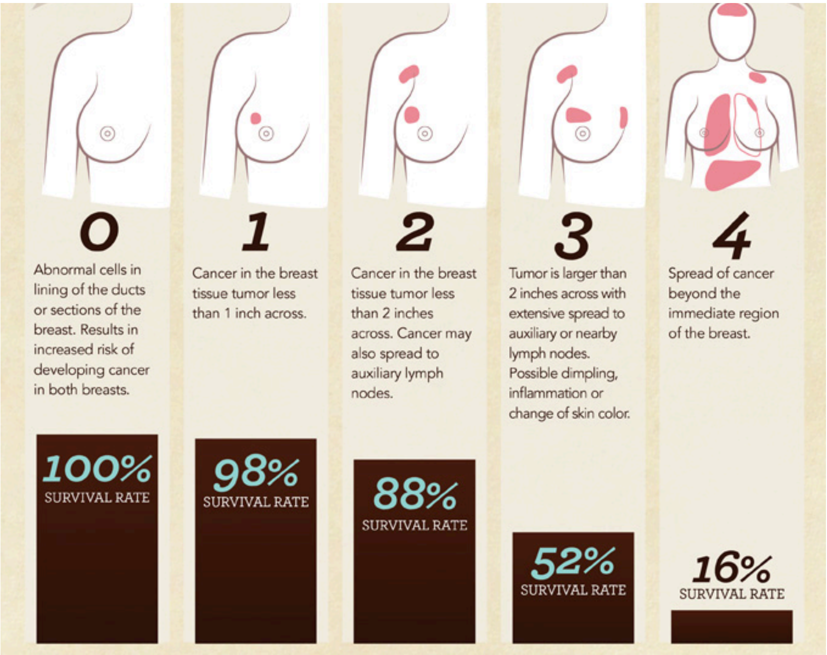
We’ ve divided the pathological stages of breast cancer (The cancer stage grouping system, rather than the TMN system) in two labeled groups:

Localized/Early stages
(stage I & II)
Label: 0

Cancer cells have not yet spread to other parts of the body with few lymph nodes involved. Tumors mostly less than 50mm

Spread/Late stages
(stage III & IV)
Label: 1

Cancer has spread through lymph nodes to other areas, metastasis occurs. Tumors generally larger than 50mm.



Abnormal cells in lining of the ducts or sections of the breast. Results in increased risk of developing cancer in both breasts.

Cancer in the breast tissue tumor less than 1 inch across.

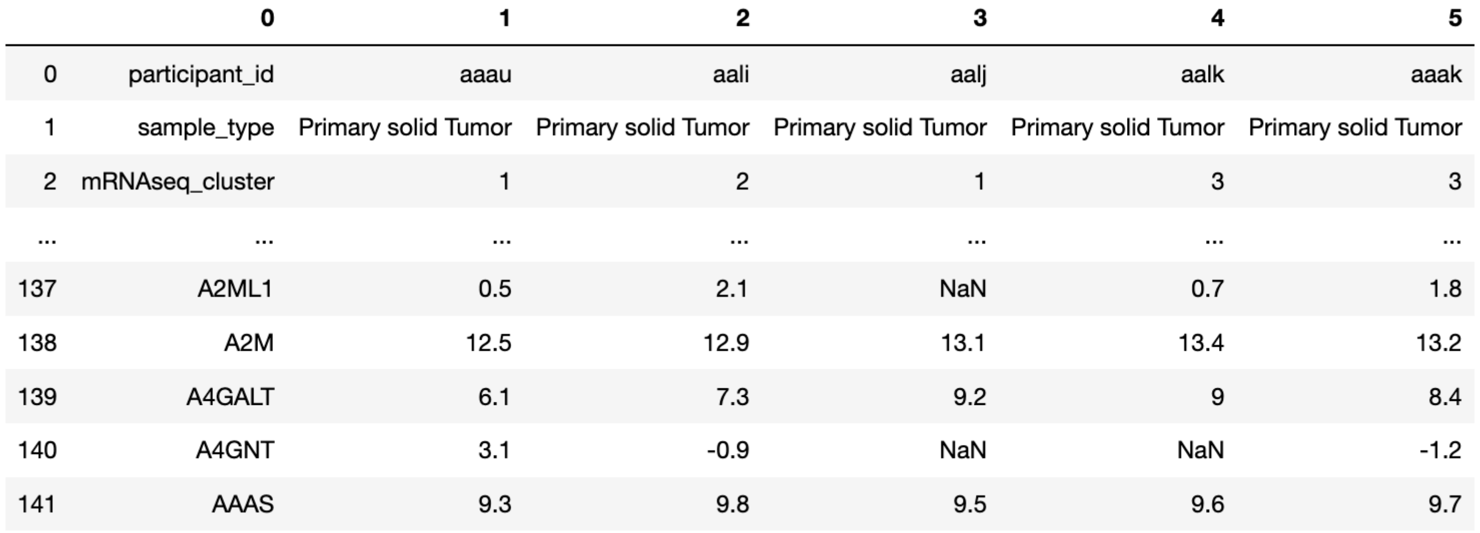
Cancer in the breast tissue tumor less than 2 inches across. Cancer may also spread to auxiliary lymph nodes.

Tumor is larger than 2 inches across with extensive spread to auxiliary or nearby lymph nodes. Possible dimpling, inflammation or change of skin color.

Spread of cancer beyond the immediate region of the breast.

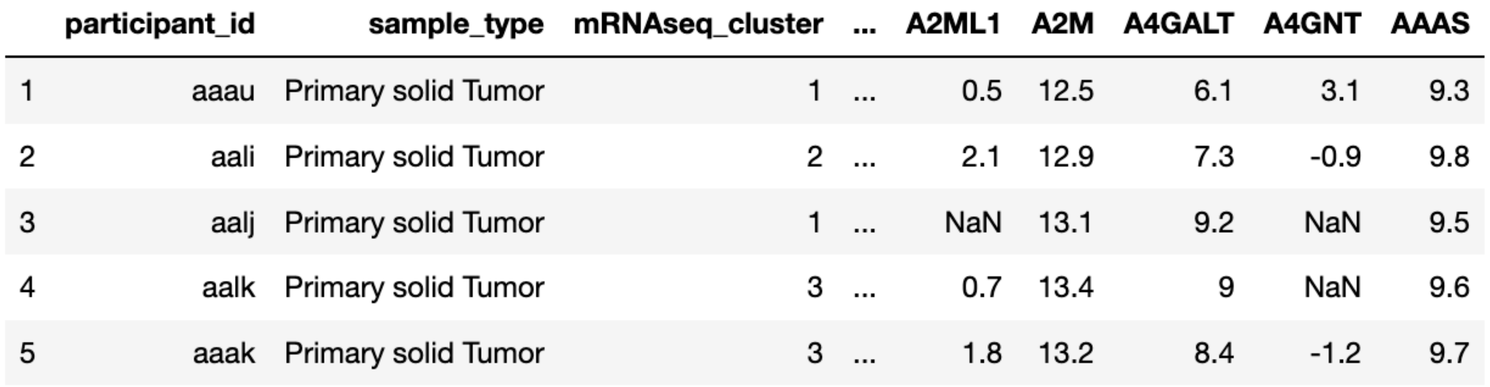
Dataset features

original dataset



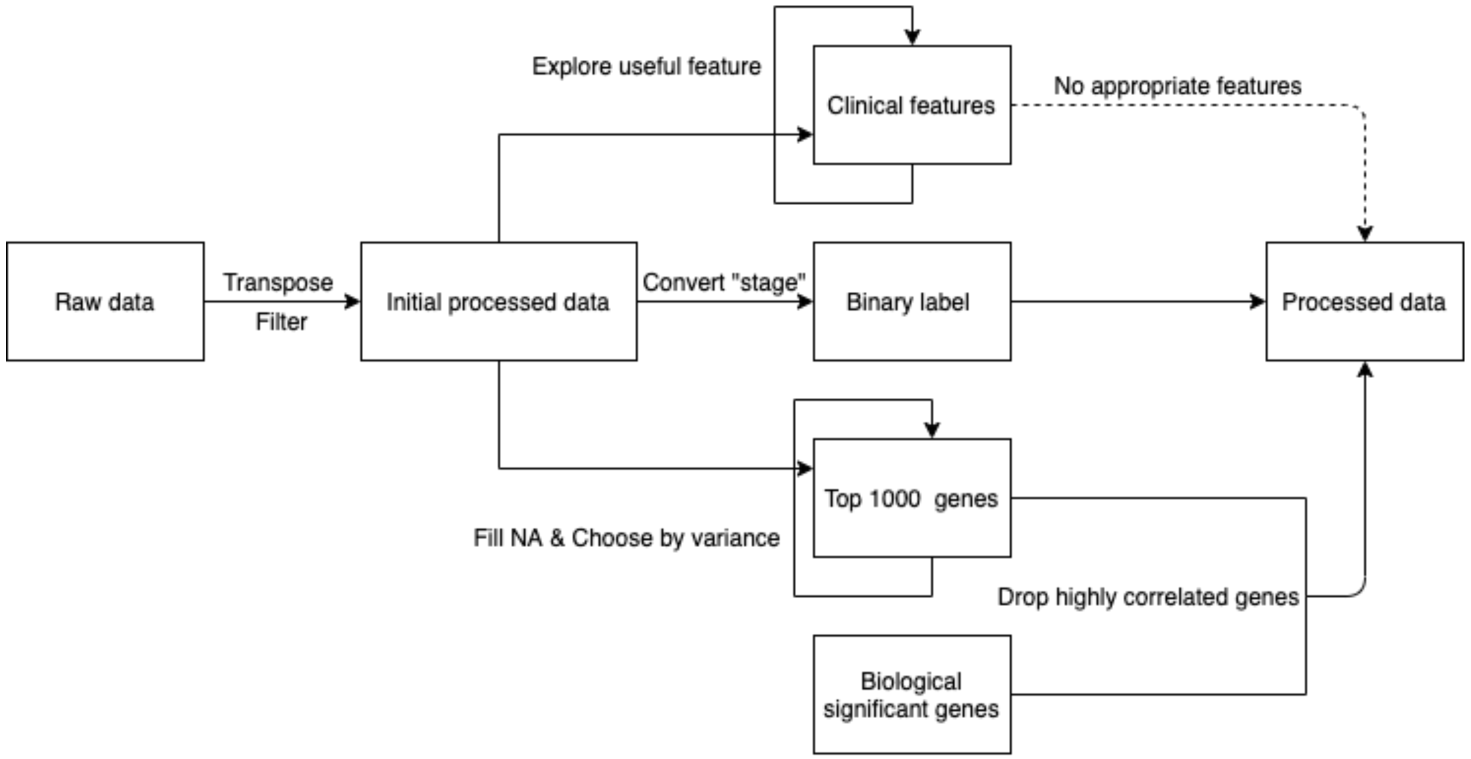
	0	1	2	3	4	5
0	participant_id	aaau	aali	aalj	aalk	aaak
1	sample_type	Primary solid Tumor	Primary solid Tumor	Primary solid Tumor	Primary solid Tumor	Primary solid Tumor
2	mRNAseq_cluster	1	2	1	3	3
...	...	...	...	...	...	...
137	A2ML1	0.5	2.1	NaN	0.7	1.8
138	A2M	12.5	12.9	13.1	13.4	13.2
139	A4GALT	6.1	7.3	9.2	9	8.4
140	A4GNT	3.1	-0.9	NaN	NaN	-1.2
141	AAAS	9.3	9.8	9.5	9.6	9.7

transposed



	participant_id	sample_type	mRNAseq_cluster	...	A2ML1	A2M	A4GALT	A4GNT	AAAS
1	aaau	Primary solid Tumor	1	...	0.5	12.5	6.1	3.1	9.3
2	aali	Primary solid Tumor	2	...	2.1	12.9	7.3	-0.9	9.8
3	aalj	Primary solid Tumor	1	...	NaN	13.1	9.2	NaN	9.5
4	aalk	Primary solid Tumor	3	...	0.7	13.4	9	NaN	9.6
5	aaak	Primary solid Tumor	3	...	1.8	13.2	8.4	-1.2	9.7

Preprocessing



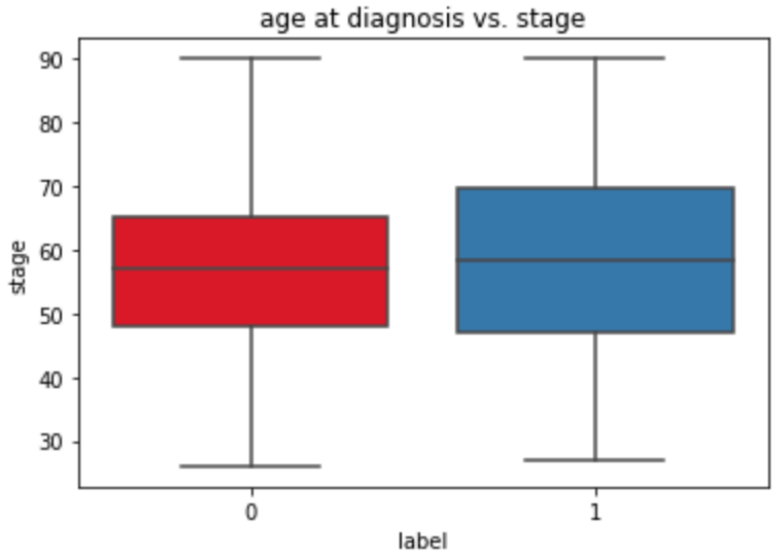
- Transpose
- Filter to unified subtypes
- Drop & fill NA
- Clinical feature screening
- Gene selection by variance
- Drop highly-correlated genes
- Assign labels

Clinical Feature Screening

Most of the clinical features are irrelevant, missing, or vague.

We checked with **Age** and **Race**, it seems like they are not associated with **Stage**.

Boxplot of **Age**



Confusion matrix of **Race**

		Early Stage	Late Stage
Race	Asian	37	10
	Black or African	114	36
	White	434	139
	American indian	0	1

Gene Feature Selection

We screened out the top 1000 genes showing the most variability in expression levels (largest variance in values), and 7 additional genes that are widely proved to have strong impact on breast cancer biologically, though those are not within the top 100 variant genes., including:

- **MYC** as the most frequent CNA cancer gene (Generate from BRC data of TCGA, Firehose Legacy, PanCancer Atlas)
- **PIK3CA** and **TP53** as the most frequent mutated cancer genes (Generate from BRC data of TCGA, Firehose Legacy, PanCancer Atlas)
- **BRCA1, BRCA2, CDH1, PTEN** as highly to moderately penetrant mutations of genes (Walsh, Michael F., et al., 2016), also frequently used as breast cancer biomarkers (National Comprehensive Cancer Network, Inc., 2018).

We will then test **50 vs 100 vs 1000 most variant genes**, plus **with vs. without 7 biologically important genes** as input of the model to compare and investigate.

03

Method: Model

Selection of models

1 **Logistic Regression**

2 **Random Forest**

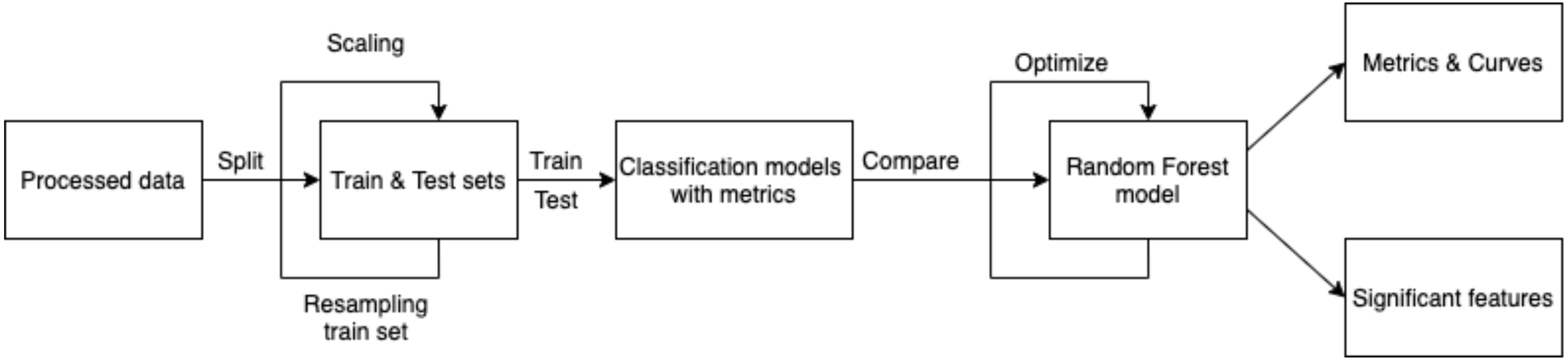
3 **Naive Bayes**

4 **Multilayer Perceptron (MLP)**

5 **SVM**

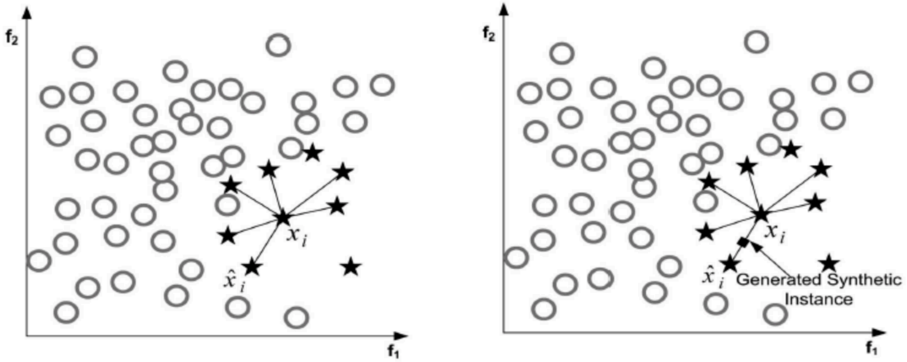
We’ ve tested with 5 models in all. Among them, **Random Forest** showed the best performance.

Modeling Strategy



Resampling Method

To deal with imbalance of the data, we used SMOTE for oversampling and undersampling.



Approach to optimizing the model performance

Adjust the number of gene selection

After modeling with 100 genes, we then chose to select the top 50 and 1000 genes with largest variance. Except for those, we tried to see the difference with/ without adding the 7 additional genes.

Adjust the parameters of models

We changed the estimator number and threshold of Random Forest to choose the best performance among them.

04
Results

Performance Metrics

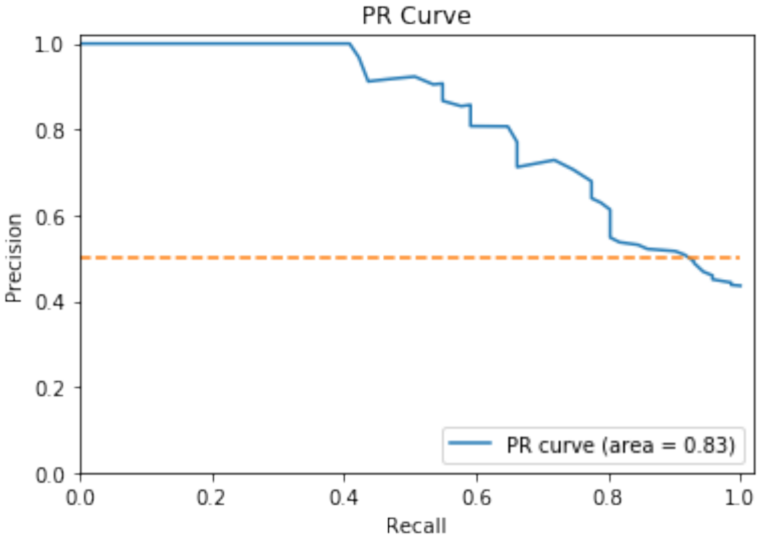
F1 score of 5 models with adjustment of gene number

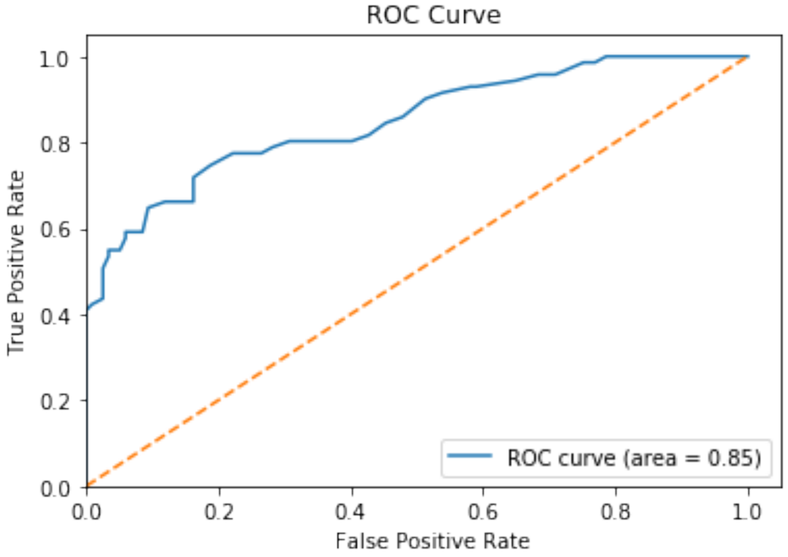
F1 score		Number of genes				
		1007	1000	7	107	57
Model	LR	0.69	0.65	0.43	0.62	0.51
	RF	0.71	0.59	0.47	0.71	0.63
	NB	0.60	0.64	0.45	0.64	0.57
	MLP	0.70	0.75	0.41	0.64	0.46
	SVM	0.63	0.63	0.32	0.41	0.25

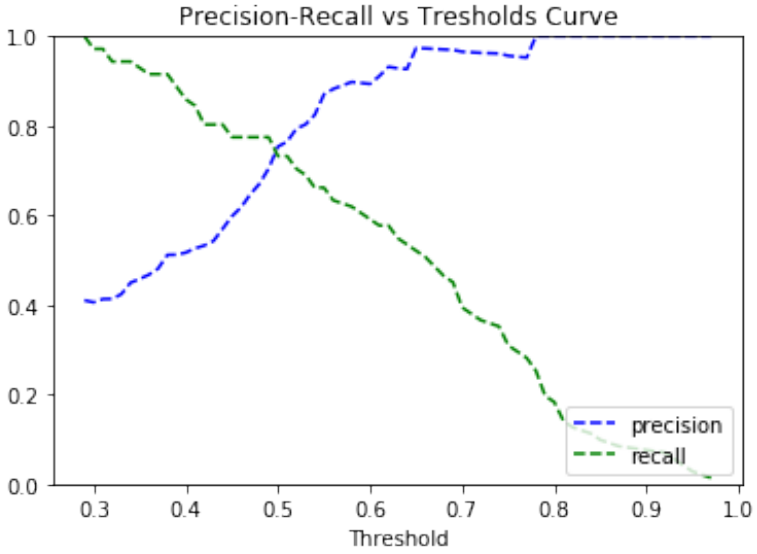
Performance metrics of Random Forest (107 genes) with adjustment of estimator number

		precision	recall	specificity	sensitivity	f1	auroc	accuracy
Estimator number	50	0.75	0.69	0.86	0.69	0.72	0.85	0.8
	100	0.77	0.75	0.86	0.75	0.76	0.87	0.82
	500	0.7	0.7	0.84	0.7	0.71	0.85	0.78

Performance Plot of Random Forest







Feature Analysis

50 most important genes

SLC30A8,SLC9A2,CNTNAP2,PIP,ABCA12,GRB14,MUC6,LRP2,LPPR3,GSTM1,COL2A1,RPS28, PPP2R2C,TAT,FAM3B,ONECUT2,CLCA2,DHRS2,LOC728606,AQP5,BRCA1,DKK1,IL20,HS6ST3,SYT1,CEACAM6,ATRNL1,CDH1,TP53,C16orf89,GLDC,TFAP2B,CRISP3,SORCS1,CLIC6,NBPF4,BRCA2,FOLR1,MUC15, SLC6A4,MSMB,FOXI1,TMPRSS4,BMPR1B,PIK3CA,TNNT1,CBLN2,PON3,AKR1B10

Most relevant pathways sorted by p-value (by Reactome)

Pathway name	Related genes found
Regulation of TP53 Expression	TP53
TP53 Regulates Transcription of DNA Repair Genes	BRCA1;TP53
SUMOylation of transcription factors	TFAP2B;TP53
Degranulation	CALML5;TCN1;CEACAM6;CRISP3;OLFM4;S100A8;LTF
Termination of O-glycan biosynthesis	MUC16;MUC5B
Resolution of D-loop Structures through Synthesis-Dependent Strand Annealing (SDSA)	BRCA1;BRCA2

05

Conclusions

Results Summary

- In this study, we try to predict early / late pathologic stage of breast cancer with gene expression data using the classification algorism.

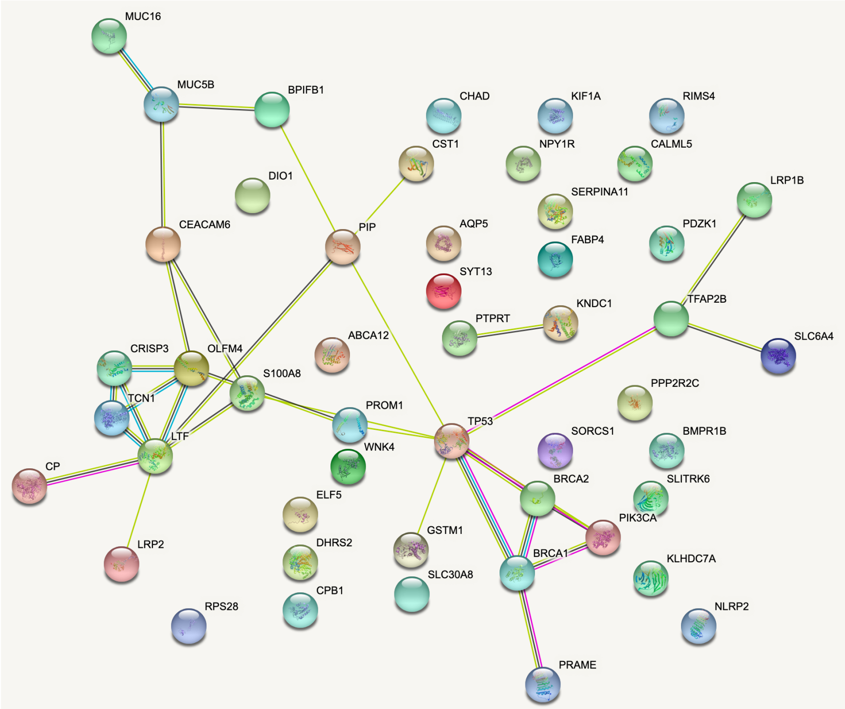
- Random forest shows the best performance in this case when we set the number of genes as 107 and the estimator number of 100, with recall of 0.75, precision of 0.77 , F1 score of 0.76.

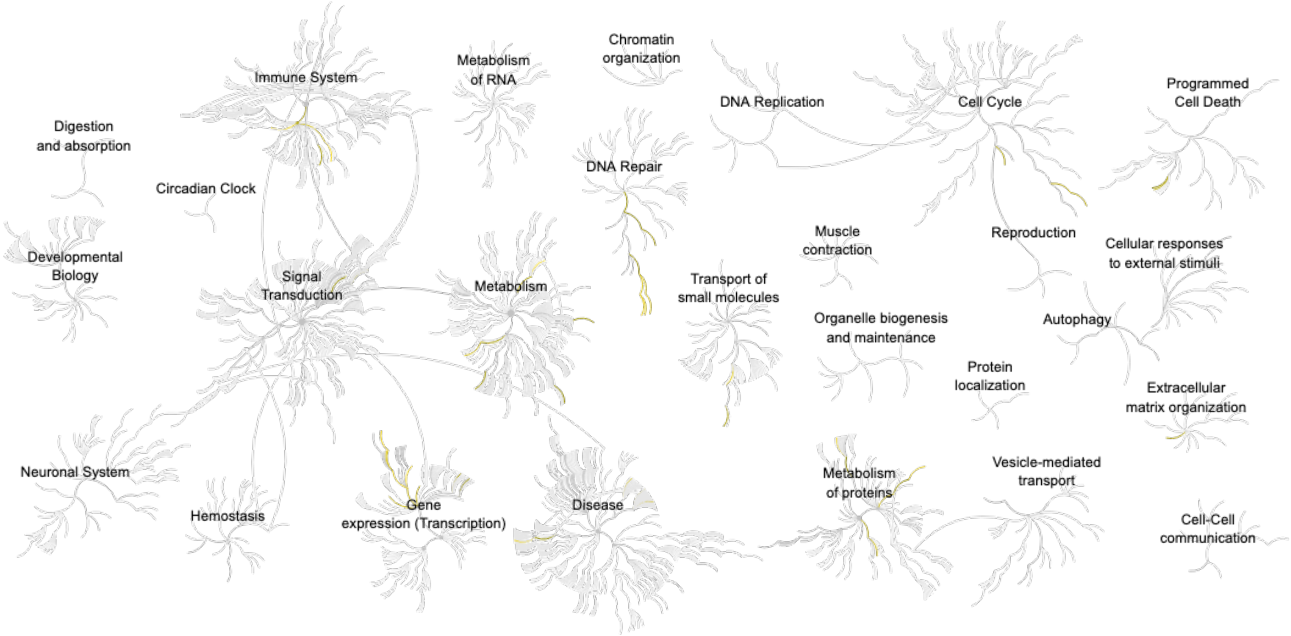
- Through the model, we find top 50 genes showing most importance to the model, we then link them with different pathways ,and map them with "hot" genes from biomedical findings which may be proved to have significant influence in breast cancer development.

Interpretability - Gene Features

- **TP53** mutation is associated with more aggressive disease and worse overall survival in breast cancer. Abrogation of the negative growth regulatory functions of p53 occurs in almost all tumors.
- **PIK3CA** mutations represent one of the most common genetic aberrations in breast cancer. Phosphatidylinositol 3-kinases are important regulators of cellular growth, transformation, adhesion, apoptosis, survival and motility.
- **BRCA1 and BRCA2** with specific inherited mutations notably increase the risk of female breast and ovarian cancers. They are essential in activating DNA repair in response to cellular stress.
- **PROM1** is involved in the stemness maintenance of cancer cells can have differential expression patterns, the encoding protein, CD133, was identified as a transmembrane protein for human hematopoietic stem cells.
- **MUC16** is usually overexpressed in the epithelial female cancer subtypes and regarded as the biomarker of cancer cell stemness, it plays an important role in progression and metastasis.
- **GSTM1** encodes for a glutathione S-transferase (GST) which play a role in the detoxification of metabolites of environmental carcinogens.

Interpretability - Network





Generated by STRING

Generated by Reactome

Potential Uses and Limitations

Potential uses

- Presents a method for identifying informative genes with differential expression, can be future used for supervised learning.
- Provide insights for pathway mechanisms of breast cancer in different stages.
- Predict the development of cancer in molecular level.

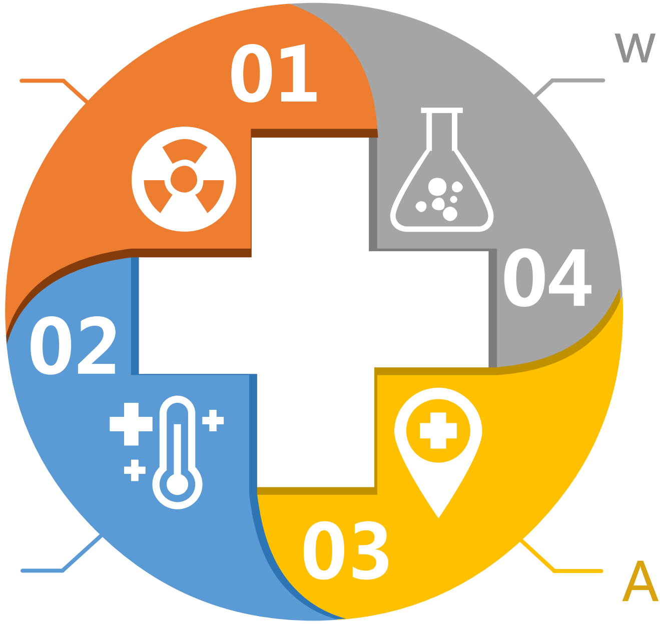
Limitation

- Comparably small sample size.
- The gene selection is merely based on variance, there could be better methods rather then unsupervised selection.
- Lack of good choices of clinical features. Most are vague or meaningless, some contains a lot of missing values.

Future Improvements

More observations. We only have 1213 samples for the dataset.

Further optimizing the model, with more parameters adjusting.



Considering adding some clinical features.

Adopting better ways for gene selection.

Reference

[1] Tan, A. C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification.

[2] Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., ... & West, M. (2003). Gene expression predictors of breast cancer outcomes. The Lancet, 361(9369), 1590-1596.

[3] Marchionni, L., Wilson, R. F., Wolff, A. C., Marinopoulos, S., Parmigiani, G., Bass, E. B., & Goodman, S. N. (2008). Systematic review: gene expression profiling assays in early-stage breast cancer. Annals of internal medicine, 148(5), 358-369.

[4] Glaab, E., Bacardit, J., Garibaldi, J. M., & Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PloS one, 7(7).

[5] Tecza, K., Pamula-Pilat, J., Lanuszewska, J., Butkiewicz, D., & Grzybowska, E. (2018). Pharmacogenetics of toxicity of 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer patients. Oncotarget, 9(10), 9114.

[6] Wang, D., Li, J. R., Zhang, Y. H., Chen, L., Huang, T., & Cai, Y. D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. Genes, 9(3), 155.

[7] Zardavas, D., Phillips, W. A., & Loi, S. (2014). PIK3CA mutations in breast cancer: reconciling findings from preclinical and clinical data. Breast cancer research, 16(1), 201.

[8] Karakas, B., Bachman, K. E., & Park, B. H. (2006). Mutation of the PIK3CA oncogene in human cancers. British journal of cancer, 94(4), 455-459.

[9] Liu, Y. (2004). Active learning with support vector machine applied to gene expression data for cancer classification. Journal of chemical information and computer sciences, 44(6), 1936-1941.

[10] Lee, A., Moon, B. I., & Kim, T. H. (2020). BRCA1/BRCA2 pathogenic variant breast Cancer: treatment and prevention strategies. Annals of laboratory medicine, 40(2), 114-121.

Thank you!

Q&A