PH240C Final Project
The Influence of Gene Expression on Survival Time of Patients with Breast Cancer

Shuyao Wang, Yanbo Wang, Sergio Mares

December 10, 2022

Abstract

Break cancer is one of the most common cancers that influences women's survival nowadays. Despite several diagnosed methods that have been applied to predict the survival rate of breast cancer patients applied with variant molecular subtypes, the individual difference still indicates that it is critical for us to identify survival and mortality-associated genes in breast cancer. Several researchers have used the random forest method to indicate that gene expression is very likely to be related to the longevity of breast cancer patients. We would like to further investigate these findings including both variables of survival status and survival time using a dataset from cBioPortal as well as exploring pairs of genes that are likely to function together to influence the survival time of breast cancer patients with the iRF method. Our findings indicated that survival status does not get significantly predicted results compared to survival time, but we did find four pairs of important co-functioning genes in breast cancer which strongly suggest researchers further explore. Further study would also consider utilizing an ensemble gradient boost algorithm or deep learning algorithm for the analysis.

1 Introduction

1. Motivation

Breast cancer accounts for the most frequent malignancy in women in the world. Aside from lung cancer, breast cancer is the leading cause of women's deaths each year (Sung et al., 2021). Every year there are over 1.5 million diagnoses of breast cancer and 570,000 deaths had been reported in 2015 with associations with this cancer (Stewart et al., 2015) worldwide. Thus, the need for the development of diagnostic tools for early prognosis in these patients is critical for the improvement of decision therapies and risk prediction (Scully et al., 2012).

Diagnosis methods such as mammograms, nuclear medicine, and biomarkers, have been used to predict the survival rate of breast cancer patients applied with variant molecular subtypes including Luminal A, Luminal B, HER2-enriched, and basal-like (Sergiusz et al., 2021). Patients within the similar molecular subtypes group also indicated different prediction results, which shows the need to make a strong accurate conclusion. Thus, it seems to be important to develop a model to evaluate the expression profiles of patients and predict the survival length and mortality of the patient since it can be beneficial as a target for robust therapies, as well as an early detector for higher-order therapies (Kruppa et al., 2012).

Moreover, previous studies have also shown the importance of identifying survival and mortality-associated genes in breast cancer. In Ning et al. (2020), they use expression profiles, methylation, and genomic mutational information in association with survival time. The model reports 10 genes related to the overall survival time of breast cancer, as well as its association with recurrence-free survival patients. However, the mutational and methylation information may play a negative in the generation of this hypothesis for candidate genes involved in cancer progression. This may be due to the fact that methylation information in both histone-modification levels and DNA methylation levels does not contain enough information longitudinally to evaluate the importance of an effect of activation or repression of a gene promoter. DNA methylation has been a focus of cancer research, due to its effect on the increased transcription of oncogenes by recruiting transcription factors within intron sequences and repressing transcription by recruitment of de-methyltransferases (MET-1) to the promoter of tumor-suppressor genes. Transcriptomic data includes sufficient information in the active genes of the cells and may be used as a better proxy to relate to cancer progression and ultimately patient mortality. Later studies may benefit to evaluate the expression profiles separately from the methylation and genomic mutational information to generate novel targets for therapies due to the redundancy of including methylation information as the expression is a direct result of the active promoter sites of the gene.

The evidence above facilizes our motivation to further evaluate the expression of genes involved in mortality and survival time in breast cancer patients as well as potential pairing effects that contribute to these predictions.

2. Dataset

cBioPortal (Gao et al., 2013) for Cancer Genomics is a resource for the interactive exploration of multidimensional cancer genomics data sets hosted by the Center for Molecular Oncology at Memorial Sloan Kettering Cancer Center (MSK). It aims to aid and expand accessibility to molecular profiles and clinical attributes and translate their rich datasets into biological insights and clinical applications (cBioPortal, 2022). The dataset we found specialized to our topic of breast cancer includes targeted sequencing of 2,509 primary breast tumors with 548 matched normal. Samples were collected by an integrated analysis of copy number and gene expression as long as the long-term clinical follow-up. All participants included in the data aged ranging from 21 and 96 years including both those with chemotherapy and hormone and the one without. In the reported information, other physiological information such as tumor size, vital status, ER status, HER2 status, menopausal status, PR status etc. is also recorded as a reference. However, in this paper, we are primarily going to focus on patients' survived time, survived status, and the gene profile expression to fit our goals.

3. Methodology

Admittedly, there are some studies that focus on the generation of linear models to predict the association between expression profiles and mortality among patients. Linear models that associate genomic impact, splicing variables, and prior knowledge of kinase genes as well as cancer genes, may benefit from non-linear models due to the simplicity of the construction of the model. Most recent studies introduce non-linear to aid in the non-linear relationships such as gene promoter variability and expression profiles (Dey et al., 2020). These non-linear models allow finding relationships among co-variate variables that otherwise be limited in linear models. Thus, we argue that evaluating the expression profiles' relationship with mortality and survival length may benefit a non-linear model architecture.

We are also going to introduce a random forest ensemble model and iterative random forest due to the simplicity, robustness, and high interpretation of the model. These models have shown to be useful in classification problems, and as a model that attempts to integrate additional data, it made sense to use a random forest-based model.

In addition, we implement an iterative random forest to get pairs of genes that are predictive of the result. According to a previous study (Basu et al., 2018), an iterative random forest is proposed to find high-order interactions of genes in predicting problems. The main difference between iRF and the original random forest is that there is a weight assigned to each leaf node, which is the probability of that feature being chosen at each split. In contrast, in the original random forest model, the weight of each feature is the same. In iRF, in each iteration, the weight of each feature is calculated by its feature importance (for example, Gini impurity), and in the next iteration, the probability that each feature is chosen at each split would be its weight. The initial weight of each feature is equal. After several iterations, we take the random intersection of decision paths, and the more often one pair of genes appears in the intersections, the more important it is.

4. Overview

In this paper, we investigate the patient's mortality and survival length with breast cancer with different types of genes included. The first aim of this study is to evaluate the ability of non-linear prediction models to predict the survival time of cancer patients given the transcriptome information of the sample. Our second aim focuses on evaluating paired expressions for the same task. To find out the most influenced genes with longevity, we run both the random forest model and the iterative random forest model with patients' binarized survival time.

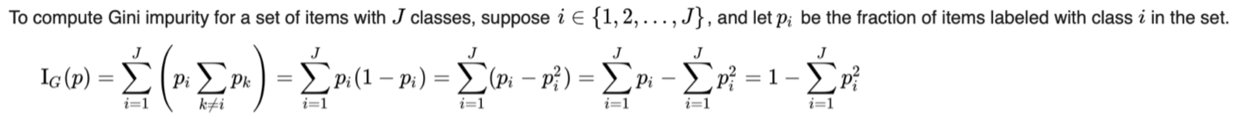
This study is organized as follows: Section 2) We introduce our dataset and the ways we clean it more elaborately. Section 3) We describe the general method, including random forest and iterative random forest used in the project followed by a complete literature review. Section 4) We report the results run by the code in python and verify it with our hypothesis and make the conclusion. Section 5) We further discuss the limitation and future direction that may help other researchers.

2 Dataset description

The transcriptome data set of breast cancer was one of the subsets from the cBioPortal for Cancer Genomics by the Center for Molecular Oncology at Memorial Sloan Kettering Cancer Center (MSK). For this study, it is pivotal to have complete information on patient mortality and survival time. The cBio dataset METABRIC (Curtis et al., 2012) provided gene expression of 2,509 primary breast tumors with 548 matched normal samples. The dataset also contains mutation molecular profiles and promoter methylation for 2,433 and 1,418 samples respectively. The clinical information table contains 39 physiological features from the patient. The expression profiles of the samples were reported for 20,603 genes with correspondent gene symbols. The age range of the reported patients is from 21 to 96 years of age with variant size tumor, vital, ER, HER, and menopausal status.

The expression data table contains 20,603 genes across 1,982 patients with both Hugo_symbol and Entrez_gene_Id annotations. After filtering the redundant gene identifiers, the dataset contains the expression profiles for 1,479 patients containing correspondent information for mortality and survival time. We also extract those patients' survival time from the clinical dataset and remove the empty vectors and replace NA with zero values to calculate the results more conveniently.

In order to use the random forest method for further calculation. We can propose one of the objective functions of tree models to be:



To compute Gini impurity for a set of items with J classes, suppose $i \in \{1, 2, \ldots, J\}$, and let p_i be the fraction of items labeled with class i in the set.

$$\mathrm{I}_G(p) = \sum_{i=1}^{J} \left(p_i \sum_{k \neq i} p_k \right) = \sum_{i=1}^{J} p_i (1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$

And the objective function overall is the sum of each individual tree.

3 Methodology

In order to evaluate the expression of genes involved in mortality and survival time in breast cancer patients, we can briefly compare their overall survival status and survival times. Thus, we come up with three models which fit to ask about our motivations.

First, we are interested in exploring the impact of different breast patients' genes expression on their overall survival status (diseased or living). In order to make the classification clear to see, we ran a classification random forest model on the overall survival status (diseased or living) by changing the patients' survival status into the binary encoding of diseased or living and taking out the 'Died of other causes and NA objects.

Our second model is to explore the impact of different breast patients' genes expression on their overall survival time. In this model, we also ran a classification random forest on overall survival time. But since the survival time is a continuous variable, we chose to binarize the patients' survival time into long and short by abandoning the data points between 90 and 160 months to draw a clear distinction between long and short survival time, which is a different strategy from Ning and his colleague's research (2020). We tuned the hyperparameters in the model by cross-validation strategy and derived the optimal model instead.

In both models, we calculated the feature importance to find predictive genes of patients' longevity. In the first model, we perform an AUC curve to evaluate the model's performance on novel data, and obtained the feature importance for every gene, while in the latter model, we performed two kinds of feature importance measures: permutation importance and MDI (mean decrease in impurity) and chose genes that tend to be important under both measures.

Another big difference between our work from Ning et al.'s study in 2020 is that by applying iRF, we could not only find individual genes that are important to survival time but also find important gene pairs that matter by applying iterative random forest to explore our third model, which wants to find pairs of predictive genes that impact the overall survival times. The reason for this significance is that genes tend to co-express and co-active, pairs of important genes could tell us more about the underlying cause of different survival times (Sipko et al., 2018).

4 Real Data Analysis

The pre-processing of the survival status data within the samples obtained is limited to Patient ID sorting, since only a limited number of patients contained both mortality and survival rate time, as well as its correspondent biopsy of the tumor sample.

The packages used in the generation of the model focused on sklearn, pandas and NumPy as a way to generate clean and simple code to be replicable within different systems. When choosing our parameters for the random forest, we expanded the number of estimators to 20,603 genes to cover the whole transcriptome given in the dataset.

First, we performed to data-dimensionality reduction algorithm principal component analysis (PCA), which allowed for the clustering of groups in the case of a linear association. In the following figures, we show that the PCA plots strengthen our hypothesis of the need for a non-linear model.

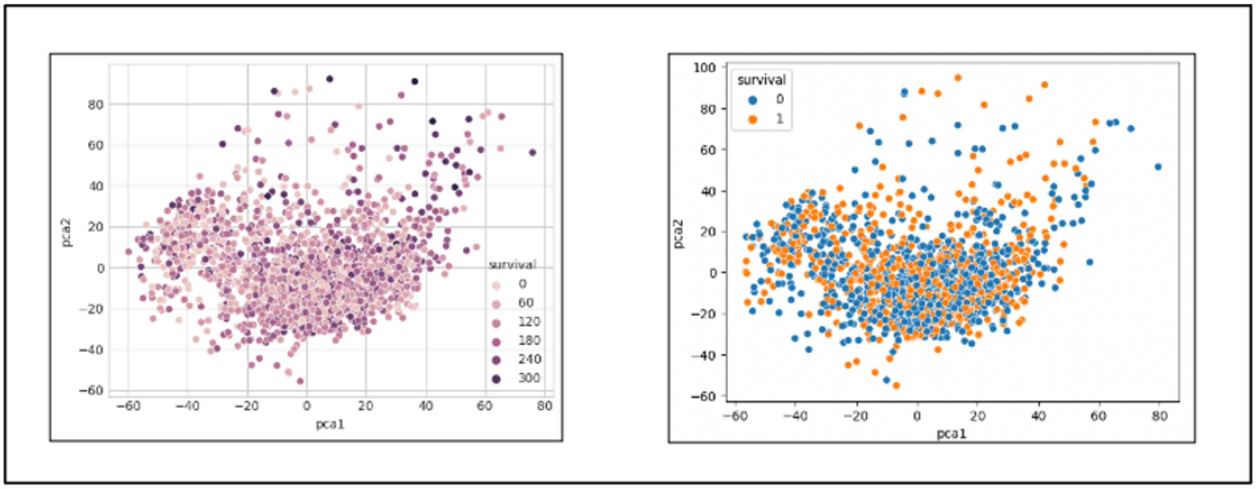


Figure 1: Figure 1. A) PCA Component system of the expression profile of the 20,603 genes colored by a binarized survival time. B) PCA Component system of the expression profile of the 20,603 genes colored by binarized survival status.

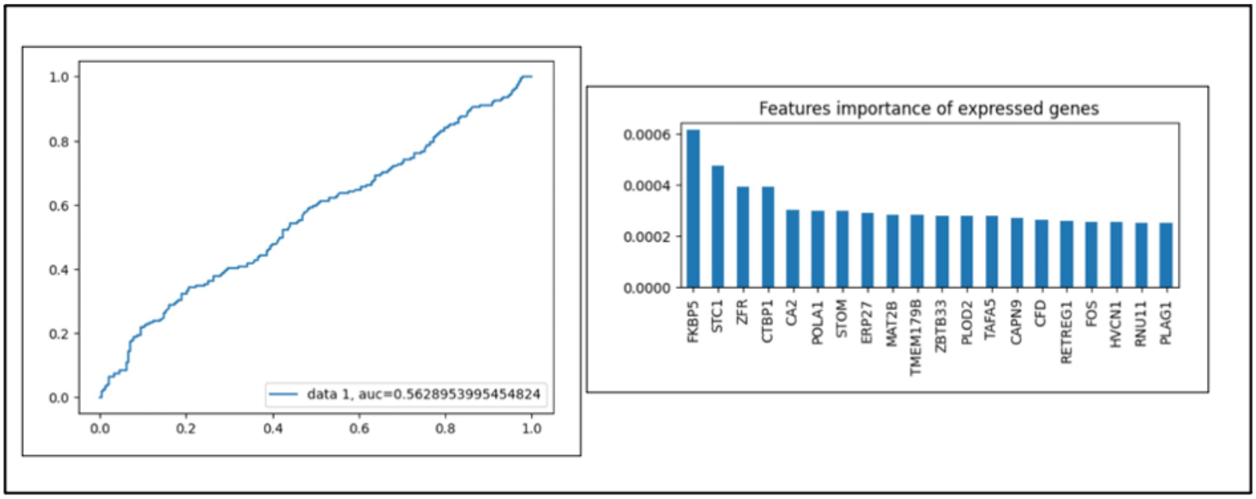


Figure 2: Figure 2. A) ROC curve of Random Forest model trained on the expression profile evaluated on new patients' transcriptomes. B) Feature selection of the most relevant genes in Random Forest model.

The trained model predicted AUC = 0.56 above a random state AUC = 0.50 and can be improved by expanding the dataset to contain information about the ancestry of the patient, heterogeneity of the samples, primary or recurrent biopsy, and other factors discussed in the conclusion.

Although our model was very limited, the prediction of FKBP5, as a gene involved in tumorigenic activity within the prediction of mortality, it has been mentioned to modulate intracellular signaling it is known to play a role in stress response. STC1 gene was also predicted to be involved in the proliferation and migration of breast cancer cells, which may elucidate the tumorigenic activity within mortality prediction. This consistency with the literature gives hope for the further improvement of the model and the addition of features for the prediction task.

Meanwhile, the pre-processing with the survival time problem is done in the following steps. First, classify survival time (which is a continuous variable) into long and short, in order to run a classification tree. Then, drop the data points with NA's; Finally, do dimensionality reduction of gene expression data based on the variances. We chose the genes that have a bigger variance of expression levels across individuals since they tend to behave differently among the patients. We didn't choose PCA here because PCA would give us the linear combination of gene expressions, while we are interested in investigating individual genes rather than their linear combinations.

We used ranger and iRF packages in R in order to perform random forest and iterative random forest algorithms. For the random forest model, we tuned mtry, which is the number of variables to possibly split at in each node; and we tried different split rules, including Gini, Hellinger, and extra trees.

We separated the data into 80% training set and 20% test set, and on the training set, we used a k-fold cross-validation strategy for tuning hyperparameters. For each possible hyperparameter, we run an 8-fold cross-validation process. We then chose the best parameter by comparing the CV loss of each model. The best choice of mtry is 90, and the best choice of the splitting rule is Gini impurity.

After that, we apply the model with optimal hyperparameters on the 20% hold-out set in order to see its performance on the unseen data. The gene expression information is predictive of the overall survival time of breast cancer patients with an accuracy of 67% on the validation set. Four important genes are selected: AURKA, PTTG1, CDCA5, and TPX2.

We first did the pre-processing of data, and run random forest models with the ranger package in R. By applying cross-validation, we found the optimal hyperparameters for the model. Then we calculated the MDI and permutation importance of the genes and picked out four genes that are important under both measures.

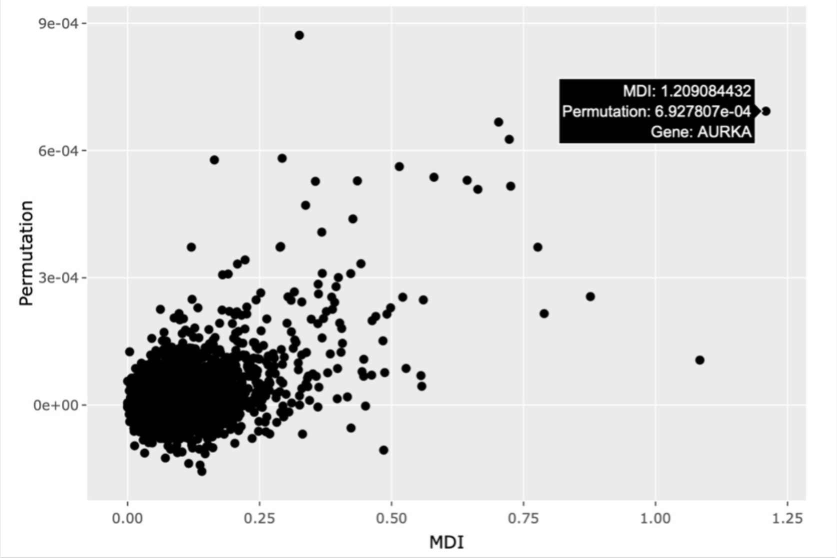
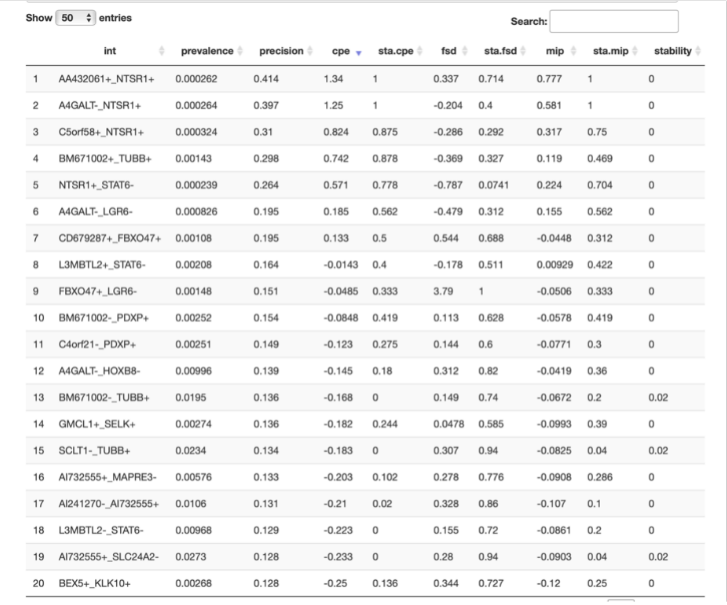


Figure 3: Figure3. 2d-feature importance of gene expression level on survival time in RF model.

For the iRF model, we first split the data into 80% training and 20% testing, and then run an iRF model on the training data. To our surprise, the accuracy on the test data reaches 84%. Here we chose the default parameters in the packages, since the performance is already very good, and tuning parameters in the iterative random forest is computationally expensive. We found important gene pairs as follows: BEX5+_MMP25+, DYRK2+_ MMP25+, MMP25+_TMEM144+ and BGN-_HOXB8-.



	int	prevalence	precision	cpe	sta.cpe	fsd	sta.fsd	mip	sta.mip	stability
1	AA432061+_NTSR1+	0.000262	0.414	1.34	1	0.337	0.714	0.777	1	0
2	A4GALT-_NTSR1+	0.000264	0.397	1.25	1	-0.204	0.4	0.581	1	0
3	C5orf58+_NTSR1+	0.000324	0.31	0.824	0.875	-0.286	0.292	0.317	0.75	0
4	BM671002+_TUBB+	0.00143	0.298	0.742	0.878	-0.369	0.327	0.119	0.469	0
5	NTSR1+_STAT6-	0.000239	0.264	0.571	0.778	-0.787	0.0741	0.224	0.704	0
6	A4GALT-_LGR6-	0.000826	0.195	0.185	0.562	-0.479	0.312	0.155	0.562	0
7	CD679287+_FBXO47+	0.00108	0.195	0.133	0.5	0.544	0.688	-0.0448	0.312	0
8	L3MBTL2+_STAT6-	0.00208	0.164	-0.0143	0.4	-0.178	0.511	0.00929	0.422	0
9	FBXO47+_LGR6-	0.00148	0.151	-0.0485	0.333	3.79	1	-0.0506	0.333	0
10	BM671002-_PDXP+	0.00252	0.154	-0.0848	0.419	0.113	0.628	-0.0578	0.419	0
11	C4orf21-_PDXP+	0.00251	0.149	-0.123	0.275	0.144	0.6	-0.0771	0.3	0
12	A4GALT-_HOXB8-	0.00996	0.139	-0.145	0.18	0.312	0.82	-0.0419	0.36	0
13	BM671002-_TUBB+	0.0195	0.136	-0.168	0	0.149	0.74	-0.0672	0.2	0.02
14	GMCL1+_SELK+	0.00274	0.136	-0.182	0.244	0.0478	0.585	-0.0993	0.39	0
15	SCLT1-_TUBB+	0.0234	0.134	-0.183	0	0.307	0.94	-0.0825	0.04	0.02
16	AI732555+_MAPRE3-	0.00576	0.133	-0.203	0.102	0.278	0.776	-0.0908	0.286	0
17	AI241270-_AI732555+	0.0106	0.131	-0.21	0.02	0.328	0.86	-0.107	0.1	0
18	L3MBTL2-_STAT6-	0.00968	0.129	-0.223	0	0.155	0.72	-0.0861	0.2	0
19	AI732555+_SLC24A2-	0.0273	0.128	-0.233	0	0.28	0.94	-0.0903	0.04	0.02
20	BEX5+_KLK10+	0.00268	0.128	-0.25	0.136	0.344	0.727	-0.12	0.25	0

Figure 4: Figure 4. important gene pairs in iRF model.

5 Conclusion

We present a non-linear model that evaluates the expression profiles and predicts the mortality and survival length of a breast cancer patient. The model was trained on 20,603 genes across 1,479 patients. We trained the model on 1,036 transcriptomes and test the model on 443 transcriptomes. The AUC curve for the test data was 0.56. This model is limited by a variety of factors from both molecular and clinical information perspectives. We think the model could have benefited from healthy samples that were associated with cancer, as these may be used as a positive control for the transcriptome counts. The model could have also benefited from including blood samples of the patients, as these may include information for specific patient bias. Also, we think tumor heterogeneity may influence the extraction of the transcripts obtained in the tumor sample. Lastly, there are known projects that have found limitations in genomic studies to the ancestry of the patients, which is known that ancestry plays a big role in immune responses. The random forest method is limited to predicting an average of previously observed labels, as it has a high and lower bound limit to the data presented. In the case novel data is shown to the model with a different range or distribution from ours, there may be a covariate shift, and would be difficult for the model to predict since it cannot extrapolate.

Although the focus of this project was to introduce a non-linear model trained on expression profiles, the addition of clinical physiological data of patients may be beneficial. Also, the binarization of the survival outcome of the patient is another limitation, as the PCA of the mortality was not shown to be any clustering for either the mortality outcome or length of survival. Future work may investigate implementing clinical information as well as obtaining a more comprehensive longitude sample from patients with a high risk of breast cancer, and first and recurrent biopsies. It is crucial to note that the larger the cohort the great the scope for gene variability and expression of the transcriptome. In addition, an ensemble gradient boost algorithm or deep learning algorithm could be used as a replacement for future models. However, for the deep learning architecture, it is important to note that interpretability may be difficult to obtain from this non-linear model.

Reference

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S; METABRIC Group, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature,* 486(7403):346-52, 2012. doi: 10.1038/nature10983. PMID: 22522925; PMCID: PMC3440846.

Dey KK, van de Geijn B, Kim SS, Hormozdiari F, Kelley DR, Price AL. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat Commun,* 11(1):4703, 2017. doi: 10.1038/s41467-020-18515-4. PMID: 32943643; PMCID: PMC7499261.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal*.* *Sci Signal*, 6(269):pl1, 2013. doi: 10.1126/scisignal.2004088. PMID: 23550210; PMCID: PMC4160307.

Kruppa J, Ziegler A, König IR. Risk estimation and risk prediction using machine-learning methods. *Hum Genet*, 131(10):1639-54, 2012. doi: 10.1007/s00439-012-1194-y. Epub 2012 Jul 3. PMID: 22752090; PMCID: PMC3432206.

Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. *Cancers (Basel),* 13(17):4287, 2021. doi: 10.3390/cancers13174287. PMID: 34503097; PMCID: PMC8428369.

Ning S, Li H, Qiao K, Wang Q, Shen M, Kang Y, Yin Y, Liu J, Liu L, Hou S, Wang J, Xu S, Pang D. Identification of long-term survival-associated gene in breast cancer. *Aging (Albany NY),* 12(20):20332-20349, 2020. doi: 10.18632/aging.103807. Epub 2020 Oct 20. PMID: 33080569; PMCID: PMC7655188.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics

2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin,* 71(3):209-249, 2021. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338.

Stewart ML, Tamayo P, Wilson AJ, Wang S, Chang YM, Kim JW, Khabele D, Shamji AF, Schreiber SL. KRAS Genomic Status Predicts the Sensitivity of Ovarian Cancer Cells to Decitabine. *Cancer Res*,75(14):2897-906, 2015. doi: 10.1158/0008-5472.CAN-14-2860. Epub 2015 May 12. PMID: 25968887; PMCID: PMC4506246.

Scully OJ, Bay BH, Yip G, Yu Y. Breast cancer metastasis. *Cancer Genomics Proteomics*, 9(5):311-20, 2012. PMID: 22990110.