Comparative Study of Machine Learning Algorithms for Estrogen Receptor Status Prediction Using Breast Cancer Metabolomics Data

Roshini Balasubramanian

Abstract

Metabolomics, the comprehensive analysis of metabolites in a biological sample, is emerging as an important tool for detecting altered cellular processes in diseased individuals. In this study, metabolomics data are used to test the accuracy of various machine learning and deep learning models on an estrogen receptor status classification task. The Deep Learning framework was the optimal model in terms of area under the receiver-operating characteristic (AUC), with support vector machines and random forests performing as a close second and third, respectively. Further analysis of predictor importance reveals that biological significance of these approaches and areas for future investigation.

Keywords: machine/deep learning, metabolomics, breast cancer, estrogen receptor status

BACKGROUND

Breast cancer manifests as a metabolic disease, and many studies use metabolites as biomarkers of receptor status (e.g. [Tang et al., 2014]). Metabolomics data from breast cancer tissue samples are used in this paper to classification as estrogen receptor positive (ER+) or estrogen receptor negative (ER-). An accurate ER status, along with human epidermal growth factor receptor 2 (HER2) and progesterone receptor (PR), supports segregation of the molecular subtypes of breast cancer: Luminal A (ER+, PR+, HER2-), Luminal B (ER+, PR+, HER2±), Her2-enriched (ER-, PR-, HER2+), and triple-negative/basal-like (ER-, PR-, HER2-), and normal-like (ER+, PR+, HER2-) [Al-thoubaity, 2020]. Identification of molecular subtype is important for determining approaches in therapy and prognosis of the patient.

[Alakwaa et al., 2018] includes two discussions: a comparative analysis of machine learning models for classification of estrogen receptor status and the validation of important metabolite features detected by the deep learning model through gene expression data. This paper addresses the first aim, testing the suitability of machine learning models on breast cancer metabolomics data, particularly for medium size samples (i.e. several hundred). The metabolomics data used in this study are available through the supplementary material of [Budczies et al., 2013], which investigates the relationship between metabolomic alterations on breast cancer subtypes using GC-TOMFS based metabolomics. I applied the following methods: support vector machines (SVM), recursive partitioning and regression trees (RPART), random forest (RF), linear discriminant analysis (LDA), generalized boosted models (GBM), and feed-forward artificial neural networks (DL). The ER status predictive accuracy was examined and

compared across the models and demonstrated that the DL, SVM, and RF models have the best performance in terms of AUC. An assessment of predictor importance for the models reveals biological interpretability and potential advantages of DL.

RESULTS

Predictive Accuracy

This study aims to compare the ability of models to classify patients by ER status. Metrics are calculated on each model for 10 independently split and shuffled training and testing subsets. The average AUC over the 10 subsets is used to compare the overall performance of the 6 classification methods. *Figure 1* below displays the results of DL (0.8808077), SVM (0.8802900), RF (0.8687700), GBM (0.8565100), RPART (0.7831400), and LDA (0.7190400). DL has the highest accuracy, with a statistically significant advantage over RPART and LDA (Wilcoxon signed-rank test $p < 0.05$).

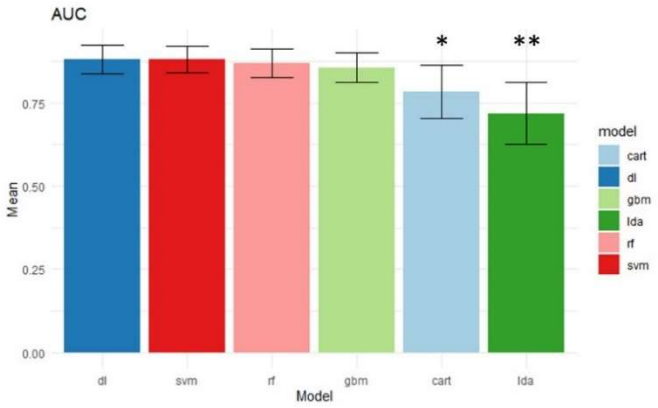

Figure 1: Average AUC of 10 independently split subsets for ER status classification. Wilcoxon signed-rank tests were used to assess the statistical significance of performance difference between the DL and the other methods (** $p < 0.001$, * $p < 0.01$).

Predictor Importance

The predictor importance of the models is related to the importance of each metabolite in determining ER status of the sample. In this way, machine learning models are biologically interpretable. Of the 162 metabolites in the dataset, 61 were ranked in the top 20 most important features for at least one of the models. *Table 1* in the Appendix lists all these metabolites and the number of models each one is included in. *Table 2* below displays a subset of these, namely the metabolites present in the top 20 most important features for at least 4 of the 6 models, and *Figure 2* presents boxplots for the features included in all 6 models.

Metabolite	Number of Models
beta-alanine	6
xanthine	6
3-phosphoglycerate	6
uracil	6
nicotinamide	5
threonic acid	5
glutamic acid	4
glyceric acid	4

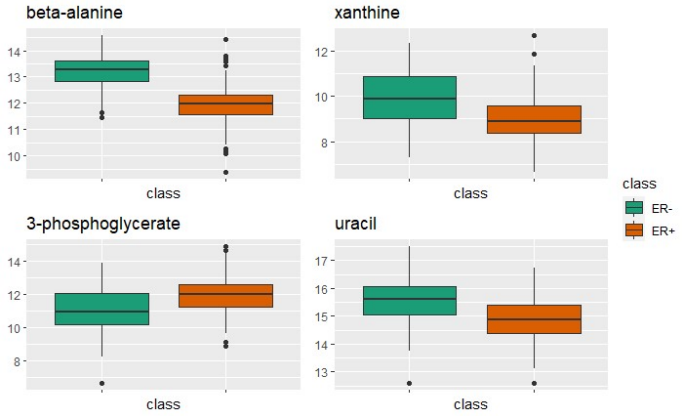

Table 2: (Left) Table displaying selected metabolites and the number of times the metabolite appears in the 20 most important features for the 6 tested models. **Figure 2:** (Right) 4 metabolites were important in all 6 models: beta-alanine, xanthine, 3-phosphoglycerate, and uracil. Boxplots of these metabolites visualize the distributions for ER+ and ER- samples.

Many of the top metabolites were identified as important in other breast cancer studies, which demonstrates the significance of these models. For instance, [Budczies et al., 2013] notes a strong change in beta-alanine between ER- and ER+ breast cancer samples and reports on the regulation of beta-alanine catabolism in cancer. A comparative metabolic profiling of human epithelial breast cancer cell lines found that xanthine and uracil (purines) and glyceric acid (sugar) are prognostic markers of breast cancer metastasis [Kim et al. 2016]. Glutamic acid was one of the metabolites found to mirror tumor burden and a potential screening tool for breast cancer patients in [Wang et al., 2018].

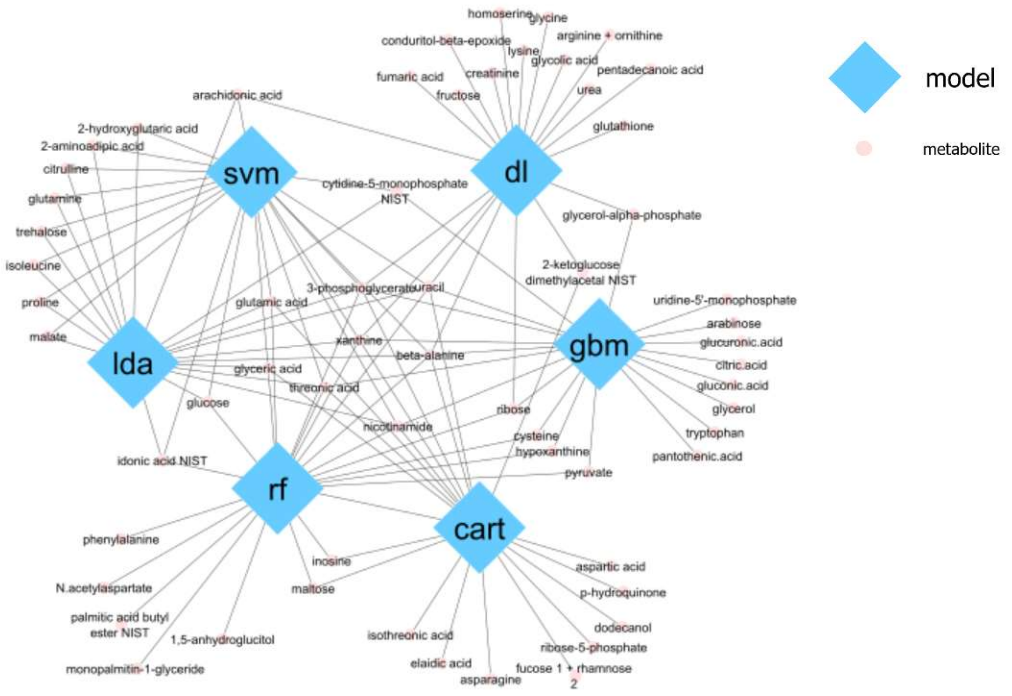

Figure 3: Bipartite network including 6 models (large, blue diamonds) and their top 20 metabolites (small, pink circles). An edge between a model and a metabolite means that this metabolite is one of the top features for this model.

Table 3 in the Appendix reports the top 20 metabolites for each model. *Figure 3*, generated in Cytoscape [Shannon et al. 2003], presents a graph of the same information, as well as the overlap between models. 4 metabolites are shared by all 6 models: beta-alanine, xanthine, 3-phosphoglycerate, and uracil. DL, GBM, RF, and RPART identified unique metabolites that were not shared with any other model.

DISCUSSION

While DL in genomics has been widely studied, applications in metabolomics are less common. This paper shows that DL, namely feed-forward neural networks, can be used for classification on metabolomics data. The performance of the DL model varies dramatically depending on the complexity and tuning of the hyperparameters, making it difficult to form broad generalizations. Increasing the complexity of the model yields higher predictive accuracy but is computationally expensive. A relatively small number of parameters were tuned, which is a limitation in this study. Implementations with different packages should be tested as well to increase the scope of findings. Considerations such as computational expense, interpretability, and size of the dataset should be weighed when using DL models. Future research should consider the effect of data size and collinearity on DL performance, which was not explored in this paper, and compare predictive accuracy with unsupervised machine learning algorithms.

Predictor importance is one way to extract biologically significant information from these models. As shown in the results, many of the metabolites identified as important through DL were not discovered by other models. [Alakwaa et al., 2018] further finds that the biological interpretation of the first hidden layer of the DL framework reveals significant metabolomics pathways, which cannot be obtained from the other machine learning algorithms. This suggests that DL is advantageous for reasons beyond predictive accuracy and requires further investigation.

CONCLUSION

Current metabolomics analysis is mainly focused on traditional machine learning methods, and this study demonstrates the usability of DL as well. DL outperforms other models on a classification task in terms of predictive accuracy, but the difference in performance compared to SVM, RF, and GBM is statistically insignificant. The metabolites selected as essential by DL varied from those of the other models, suggesting that the features extracted as most important could lead to the discovery of new biomarkers.

METHODS

Data and Pre-processing

The metabolomics data used for this analysis contain 271 breast cancer tissue samples (204 ER+, 67 ER-) and 162 metabolites. Data pre-processing followed the steps from [Alakwaa et al., 2018], including the imputation of missing value through K-Nearest Neighbors, addition of the ER statuses from [Budczies et al., 2013], a log transformation, mean centering, autoscaling by the standard deviation, and quantile normalization. The quantile normalization procedure followed [Zhao et al., 2020]: rank each sample by magnitude, calculate the average value for samples occupying the same rank, and substitute the values of that rank with the calculated average value. *Figure 4* depicts the results of normalization.

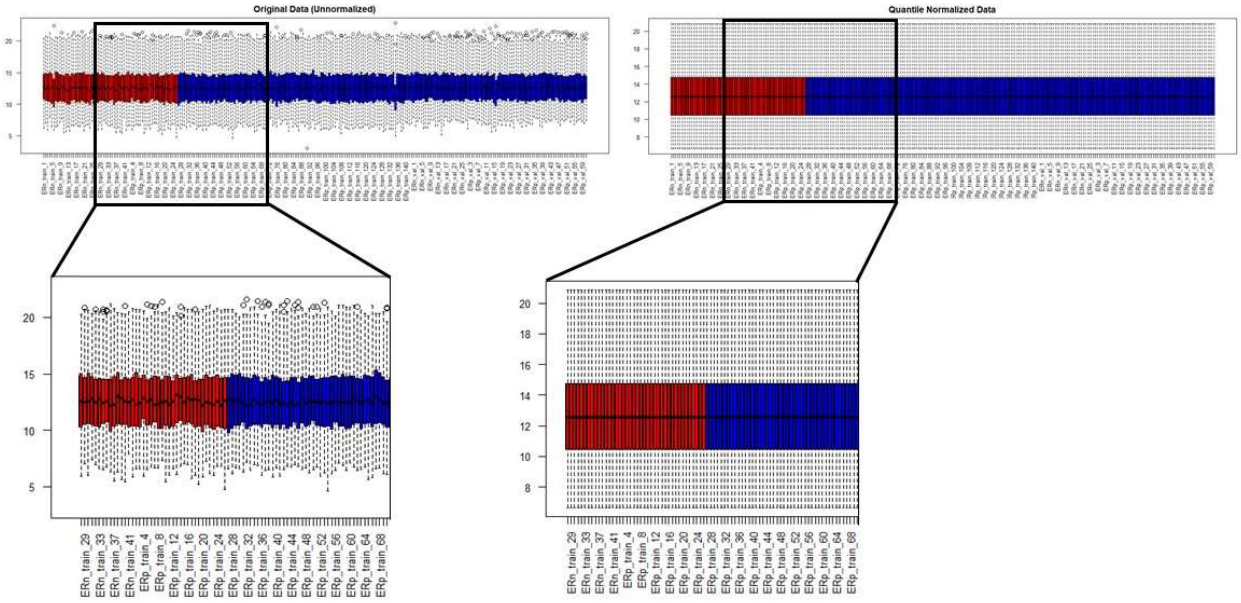

Figure 4: Box plots of expression data before and after quantile normalization. Unlike the unnormalized data (left), the normalized data (right) box plot is distributed across the same interval, indicating successful preprocessing. The normalized data was used for further analysis.

The metabolite order is shuffled, and the dataset is separated into the training set, a subset to train the model, and the testing set, a subset to test the model, in order to determine the predictive power of the model on new data. The data are split such that 80% is for training and the remaining 20% is for testing.

Machine Learning Models

A representative set of models were included in the comparison. The *caret* Package in R was used for model training and tuning on SVM, RPART, RF, GBM, and LDA [Kuhn, 2008]. Predictor importance was evaluated through model specific metrics such that a feature's importance is tied to model performance. Further details are available in [Kuhn, 2008]. The variable importance function varImp was used to extract rankings.

SVM is a supervised classification algorithm. For binary classification, SVM constructs a hyperplane, a boundary between the two classes (ER+ and ER-), by maximizing the margin between the classes without capturing any samples. Given training data *X*, the samples are split in such a way that the distance of the support vectors to the hyperplane, $W \cdot X + b = 0$, is maximized. A radial kernel function was used to build a non-linear SVM classifier. Optimal values were chosen for the model tuning parameters to maximize model accuracy.

RPART fits classification and regression tree (CART) models. The classification tree is built by searching through predictors to find the best value for a single predictor in order to minimize misclassifications. The data is split into two groups based on the identified value, and the process is repeated, creating a hierarchical structure. Splitting continues until criterion for stopping is met, and then, trees go through 10-fold cross-validation for pruning, in which terminal nodes are iteratively removed to avoid overfitting. The specific tree is chosen through *caret*'s default "one SE" rule. *Figure 5* displays a possible tree output for the metabolomics data

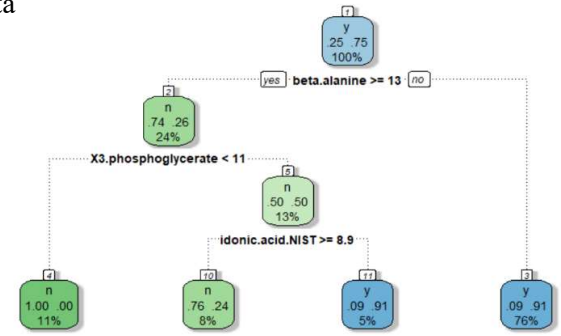

Figure 5: A tree dendrogram representing an example RPART classification on the training data. The "n" and "y" classifications represent ER- and ER+ status, respectively.

RF is a combination of these tree-structured predictors, where each tree is grown through CART without pruning. With ensemble learning, a model is fit with multiple trees, created using bootstrapped data from the original dataset, to improve performance. By adding more trees, the risk of error from an individual tree is reduced. The tree number of the forest was tuned.

GBM boosts weak learning algorithms, individual trees, by considering past fits. After a tree is created, the weight of each tree is determined based on its quality and subsequent iterations, to create more trees of roughly the same size. The final prediction is a weighted average of each tree as follows:

$f(x) = \frac{1}{M}\sum_{j=1}^{M} \beta_j f_j(x)$, where M is the total number of trees, f_j is the j^{th} tree, β_j is the weight of the tree, and $f(x)$ produces a classification.

LDA adopts a linear combination of variables as a predictor, $F_{LDA} = WX + E$, to maximize the ratio of the variance between the classes and the variance within classes. Predictions are made by estimating the probability that the sample belongs to each class through Bayes Theorem, with the sample being assigned to the class it receives the highest probability for.

Deep Learning Model

The *h2o* Package in R was used to model and tune the parameters for DL [Candel, 2017]. Parameters tuned in this study include the activation function, number of hidden layers, L1 and L2 penalties, and Epochs. The variable importance function varimp in *h2o* was used to extract rankings.

The feed-forward artificial neural network connects the input x and the output z by multiple hidden layers y with a weight matrix W. The shape of W is determined by the number of units in the layers mapped from and to. For activation function f, $y = f(Wx + b)$. *Figure 6* depicts an extremely simplified version of the DL model used in this study to demonstrate the general architecture of a neural network.

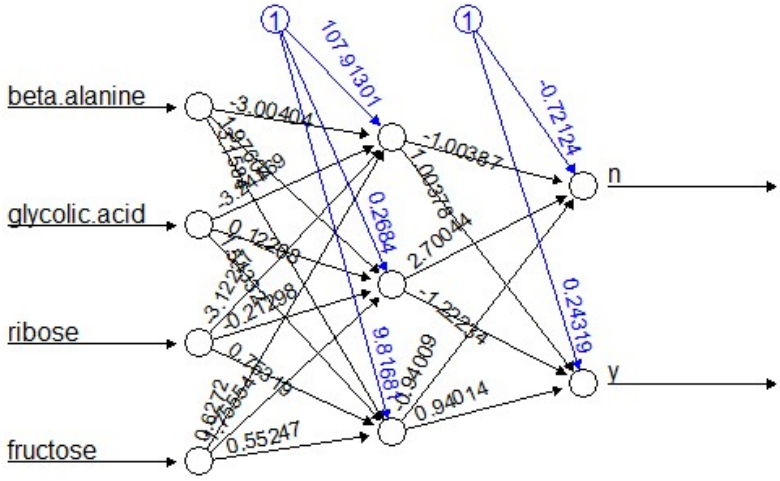

Figure 6: Visualization of a neural network including only 4 metabolites for clarity. The neural network used for prediction in this study included all 162 predictors and was much more complex. For instance, this neural network pictured here only has a single hidden layer.

Availability of Data and Materials

The data and data profiles that support the findings of this study are available through the supplementary material of [Budczies et al., 2013]. Generated data and R code for analysis can be found in the following repository: https://github.com/roshinib3/breast_cancer_metabolomics.

DECLARATIONS

APPENDIX

Table 1: Table of all 61 metabolites included in any of the model's 20 most important features and the number of models each metabolite is in the 20 most important for (i.e. 1 means the metabolite is in the top 20 for one model and 6 means it is the top 20 for all 6 models).

Metabolite	Number of Models
beta-alanine	6
xanthine	6
3-phosphoglycerate	6
uracil	6
nicotinamide	5
threonic acid	5
glutamic acid	4
glyceric acid	4
ribose	3
glucose	3
idonic acid NIST	3
cytidine-5-monophosphate NIST	3
arachidonic acid	3
hypoxanthine	2
inosine	2
cysteine	2
pyruvate	2
maltose	2
2-hydroxyglutaric acid	2
glutamine	2
2-aminoadipic acid	2
isoleucine	2
proline	2
citrulline	2
trehalose	2
malate	2
2-ketoglucose dimethylacetal NIST	2
glycerol-alpha-phosphate	2
monopalmitin-1-glyceride	1
phenylalanine	1
1,5-anhydroglucitol	1
palmitic acid butyl ester NIST	1

N-acetylaspartate	1
ribose-5-phosphate	1
aspartic acid	1
dodecanol	1
fucose 1 + rhamnose 2	1
p-hydroquinone	1
elaidic acid	1
isothreonic acid	1
asparagine	1
citric acid	1
gluconic acid	1
tryptophan	1
arabinose	1
uridine-5'-monophosphate	1
glucuronic acid	1
pantothenic acid	1
glycerol	1
glycolic acid	1
fructose	1
creatinine	1
glutathione	1
glycine	1
conduritol-beta-epoxide	1
pentadecanoic acid	1
fumaric acid	1
lysine	1
arginine + ornithine	1
urea	1
homoserine	1

Table 2: Each column lists the top 20 features, from most to least important, extracted from the respective model.

RF	SVM	RPART	LDA	GBM	DL
beta-alanine	beta-alanine	beta-alanine	beta-alanine	beta-alanine	beta-alanine
xanthine	xanthine	xanthine	xanthine	xanthine	glycolic acid
3-phosphoglycerate	glutamic acid	glutamic acid	glutamic acid	3-phosphoglycerate	ribose
monopalmitin-1-glyceride	uracil	uracil	uracil	nicotinamide	fructose
hypoxanthine	2-hydroxyglutaric acid	threonic acid	2-hydroxyglutaric acid	citric acid	creatinine
phenylalanine	threonic acid	3-phosphoglycerate	threonic acid	gluconic acid	3-phosphoglycerate
uracil	idonic acid NIST	idonic acid NIST	idonic acid NIST	hypoxanthine	uracil
1,5-anhydroglucitol	glyceric acid	ribose-5-phosphate	glyceric acid	tryptophan	xanthine
glutamic acid	3-phosphoglycerate	aspartic acid	3-phosphoglycerate	arabinose	glutathione
inosine	glutamine	glyceric acid	glutamine	uridine-5'-monophosphate	glycine
nicotinamide	2-aminoadipic acid	dodecanol	2-aminoadipic acid	ribose	arachidonic acid
glyceric acid	cytidine-5-monophosphate NIST	fucose 1 + rhamnose 2	cytidine-5-monophosphate NIST	pyruvate	2-ketoglucose dimethylacetal NIST
palmitic acid butyl ester NIST	isoleucine	maltose	isoleucine	cytidine-5-monophosphate NIST	conduritol-beta-epoxide
ribose	proline	inosine	proline	glucuronic.acid	glycerol-alpha-phosphate
cysteine	nicotinamide	2-ketoglucose dimethylacetal NIST	nicotinamide	pantothenic.acid	pentadecanoic acid
threonic acid	arachidonic acid	p-hydroquinone	arachidonic acid	threonic acid	fumaric acid
glucose	citrulline	elaidic acid	citrulline	cysteine	lysine
pyruvate	trehalose	isothreonic acid	trehalose	glycerol	arginine + ornithine
N-acetylaspartate	glucose	nicotinamide	glucose	glycerol-alpha-phosphate	urea
maltose	malate	asparagine	malate	uracil	homoserine

ACKNOWLEDGMENTS

This study was performed for QCB 455: Introduction to Genomics and Computational Molecular Biology at Princeton University. I thank the Instructors and Teaching Assistants—Prof. Josh Akey, Prof. Mona Singh, Prof. Claire McWhite, Riley Skeen-Gaar, Antonio Muscarella, and Tyler Park—for leading an engaging course and providing valuable advice on this paper. I thank Dr. Fadhl Alakwaa for assistance with accessing the data and Dr. Carsten Denkert and Dr. Oliver Fiehn for making the metabolomics data available.

REFERENCES

[1] Tang X., Lin, C. C., Spasojevic, I., Iversen, E. S., Chi, J. T., Marks, J. R. (2014). A joint analysis of metabolomics and genetics of breast cancer. *Breast Cancer Res*.,16(4):415. 5. https://doi:10.1186/s13058-014-0415-9

[2] Al-Thoubaity F. K. (2019). Molecular classification of breast cancer: A retrospective cohort study. *Annals of medicine and surgery, 49*, 44–48. https://doi.org/10.1016/j.amsu.2019.11.021

[3] Alakwaa, F. M., Chaudhary, K., & Garmire, L. X. (2018). Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. *Journal of proteome research*, *17*(1), 337–347. https://doi.org/10.1021/acs.jproteome.7b00595

[4] Budczies, J., Brockmöller, S. F., Müller, B. M., Barupal, D. K., Richter-Ehrenstein, C., Kleine-Tebbe, A., Griffin, J. L., Orešič, M., Dietel, M., Denkert, C., Fiehn, O.. (2013). Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism, *Journal of Proteomics*, *94*, 279-288. https://doi.org/10.1016/j.jprot.2013.10.002

[5] Kim, H. Y., Lee, K. M., Kim, S. H., Kwon, Y. J., Chun, Y. J., & Choi, H. K. (2016). Comparative metabolic and lipidomic profiling of human breast cancer cells with different metastatic potentials. *Oncotarget*, *7*(41), 67111–67128. https://doi.org/10.18632/oncotarget.11560

[6] Wang, X., Zhao, X., Chou, J., Yu, J., Yang, T., Liu, L., Zhang, F. Taurine, glutamic acid and ethylmalonic acid as important metabolites for detecting human breast cancer based on the targeted metabolomics. *Cancer Biomark*, *23*(2), 255-268. https://doi: 10.3233/CBM-181500.

[7] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498-2504. https://genome.cshlp.org/content/13/11/2498.full.pdf+html.

[8] Zhao, Y., Wong, L., Goh, W.W.B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Sci Rep, 10***,** 15534. https://doi.org/10.1038/s41598-020-72664-6

[9] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*. http://www.math.chalmers.se/Stat/Grundutb/GU/MSA220/S18/caret-JSS.pdf

[10] Candel, A., Parmar, V., LeDell, E., Arora, A. (2016). Deep Learning With H2O, 5th ed. Mountain View, CA, USA: H2O.ai Inc. http://h2o-release.s3.amazonaws.com/h2o/rel-turnbull/2/docs-website/h2o-docs/booklets/DeepLearningBooklet.pdf