Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230577

Predicting Overall Survival in METABRIC Cohort Using Machine Learning

Afroz BANU[a,$], Rayyan AHMED[b,$], Saleh MUSLEH[b] , Zubair SHAH[b] , Mowafa HOUSEH[b] and Tanvir ALAM[b,1]
[a]*College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar*
[b]*College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar*

Abstract. Triple-negative breast cancer (TNBC) is an aggressive form of breast cancer that presents very high relapse and mortality. However, due to differences in the genetic architecture associated with TNBC, patients have different outcomes and respond differently to available treatments. In this study, we predicted the overall survival of TNBC patients in the METABRIC cohort employing supervised machine learning to identify important clinical and genetic features that are associated with better survival. We achieved a slightly higher Concordance index than the state of art and identified biological pathways related to the top genes considered important by our model.

Keywords. Breast Cancer, Machine Learning.

1. Introduction

Triple-negative breast cancer (TNBC) is the most fatal form of breast cancer world wide. TNBC is mainly characterized by the signature feature of the absence of progesterone, estrogen and HER2 receptors. It is highly aggressive with high relapse rates and a tendency to metastasize which is associated with poor prognosis [1]. TNBC is caused by genetic mutations and the BRCA1 gene is highly associated with TNBC. However, with the development of next-generation sequencing, several genetic mutations affecting the expression of multiple susceptibility genes have been identified that are associated with TNBC, elucidating its highly heterogeneous genetic nature [2]. Different genes enrich distinct biological pathways that reflect differences in clinical outcomes [3,4]. Many models have been applied to cancer cohorts to classify patients and predict survival based on clinical data [5]. One advantage of these models is that individuals can be stratified into risk groups paving way for personalized medicine and the accuracy with which it could be achieved can be improved by integrating genetic data with the clinical profile of patients [6,7]. Considering the high dimensionality of the gene expression data [8], *Kumar et al.* reduced the dimensions of genetic data and predicted survival based on different features on Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) data and achieved the highest Concordance Index (CI) with the multi-task logistic regression (MTLR) model (Table 1) [7]. In this study, we predicted the survival based on clinical and genetic attributes of the METABRIC dataset to identify genes and

[1] Corresponding Author: Tanvir Alam, E-mail: talam@hbku.edu.qa.
$: Contributed Equally.

underlying biological pathways that can be targeted to improve the overall survival of the breast cancer patients.

Table 1. Existing results for survival prediction on METABRIC dataset using different features. PCA- Principle component analysis; LDA- discretized Latent Dirichlet Allocation; PAM50- Prediction Analysis of Microarray 50

Feature groups	CI
Clinical + PCA + dLDA+PAM50	0.7202
Clinical + PCA + dLDA	0.7079
Clinical + dLDA	0.7139
Clinical + PCA	0.6999
Clinical	0.6820

2. Methods

2.1 Dataset Collection and Preprocessing

We used the METABRIC [9] cohort which is comprised of 31 clinical attributes, gene expression profiles of 331 genes, mutations of 175 genes from 1904 breast cancer samples. The clinical and genomic data was downloaded from cBioPortal. The dataset is imbalanced with 103 patients who survived and 801 died. We first checked if the dataset contains any missing data. To deal with such missing data, we replaced them with zeros in order to obtain our baseline results. Next, as the dataset contains multiple categorical features, we applied one-hot encoding scheme to encode categorical features into a zero-one numeric array. Then we normalized the features using min-max normalization.

2.2 Supervised machine learning algorithm

We applied four machine learning models, k-nearest neighbor (kNN), Logistic Regression, Random Forest, and Support Vector Machine (SVM) and split the dataset into training and test at 80-20 ratio. To evaluate our model's performance, we use the Area Under the Curve of Receiver Operating Characteristic curve (AUC-ROC), Concordance Index (CI), and Matthews Correlation Coefficient (MCC).

2.3 Functional Analysis of Genes

For functional analysis of genes, we identified statistically significant (adjusted p-value < 0.05 after false discovery rate (FDR) correction) gene ontology (GO) terms using gProfiler [12].

3. Results and Discussion

3.1 Model performance

In our study, we have shown the predictive ability of ML and prioritized the survival-related genes based on their importance determined by the model. Out of four models we tried, LR based model achieved the best performance (Table 2). We observed that incorporating more features led to increased complexity and the high dimensionality of

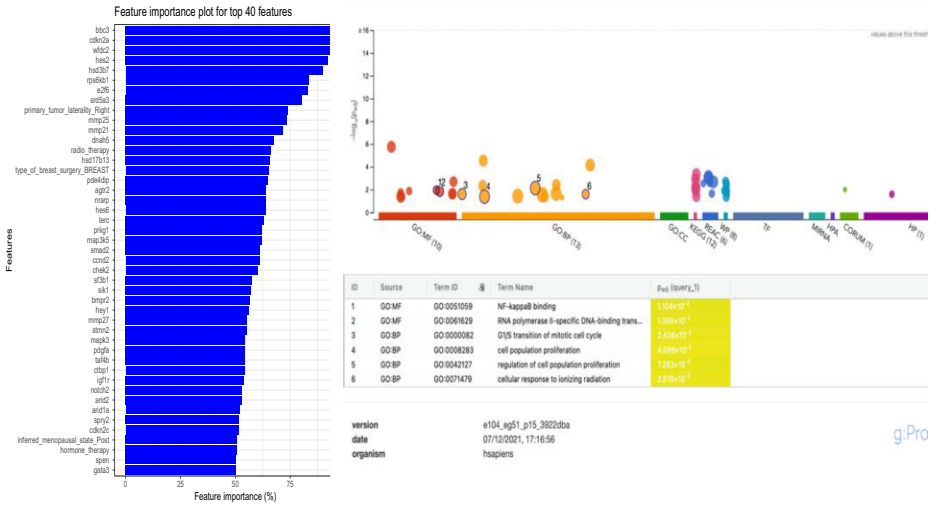
data increases the burden of multiple hypothesis testing and the likelihood of false-positive genetic associations [10].

Table 2. Machine Learning models performance on METABRIC dataset.

Model	AUC	MCC	CI
Random Forest Classifier	0.62	0.27	0.623
SVC Classifier	0.70	0.41	0.705
KNN	0.72	0.43	0.716
Logistic Regression	**0.73**	**0.45**	**0.726**

3.2 Functional enrichment analysis.

We selected the features that are at least 50% important in predicting overall survival and observed that majority of the features were genes. Among clinical attributes, radiotherapy, primary tumour laterality and breast-conserving surgery were identified as important (Figure 1 A). Machine learning algorithms [11] can recognize patterns in data and can predict the outcome but it is difficult to determine the causal effect of these feature. For example, we have been able to identify survival-related clinical attributes and genes employing ML but why they are important cannot be determined. Therefore, we queried the top genes for enriched gene ontology (GO) in gProfiler which is a web-based tool for finding biological pathways enriched in gene lists. We observed that most of the identified GO terms were cancer-related such as regulation of cell proliferation and cell cycle-related pathways (Figure 1 B).



(A) (B)

Figure 1. A. Feature importance plot for features that are at least 50% important in predicting overall survival. **B.** Plot showing enriched pathways for the 37 genes which were important in predicting overall survival in TNBC patients.

Then we did the analysis for the GO terms they could be involved in. We noticed that BBC3, CCND2, CHEK2 and GATA3 were enriched cellular response to ionizing radiation suggesting that radiotherapy is an indicator of survival probably due to the involvement of these genes. Functional validation of these genes in a suitable model could identify novel pathways and drug targets that can be exploited for improving overall survival in TNBC patients. Since TNBC is a heterogeneous disease with multiple

involved genes, not all TNBC patients will have the same gene expression pattern. Therefore, a more realistic approach would be identifying the genes that are survival-related at the individual level. This could be done by calculating the contribution of each gene in each participant towards survival followed by clustering which will be more informative in terms of genes driving the clusters providing a more specific genetic landscape of TNBC. Full list of genes and related to GO are shared in GitHub: https://github.com/tanviralambd/Metabric .

4. Conclusion

Our analysis showed that ML model incorporating clinical and genetic data can predict the overall survival of TNBC patients. This approach can identify the survival-related genes and investigating these genes could increase our understanding of why TNBC patients have different clinical outcomes and therapeutic responses. In clinical practice, the implementation of such an approach could identify high-risk patients and guide appropriate management. However, ML algorithms behave differently from humans in that they are literal which means they understand what has been told explicitly and can't adjust on their own. Therefore, the model outputs should not be trusted blindly, they should be scrutinized, and the algorithm should be modified appropriately.

References

[1] Sporikova Z, Koudelakova V, Trojanec R, Hajduch M. Genetic markers in triple-negative breast cancer. Clinical breast cancer. 2018 Oct 1;18(5):e841-50., doi: 10.1016/j.clbc.2018.07.023.
[2] Shimelis H, LaDuca H, Hu C, Hart SN, Na J, Thomas A, Akinhanmi M, Moore RM, Brauch H, Cox A, Eccles DM. Triple-negative breast cancer risk genes identified by multigene hereditary cancer panel testing. JNCI: Journal of the National Cancer Institute. 2018 Aug 1;110(8):855-62. doi: 10.1093/jnci/djy106.
[3] Prat A, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. The Breast. 2015 Nov 1;24:S26-35. doi: 10.1016/j.breast.2015.07.008.
[4] Lyons TG. Targeted therapies for triple-negative breast cancer. Current treatment options in oncology. 2019 Nov;20:1-3. doi: 10.1007/s11864-019-0682-x.
[5] Parker JS, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009 Mar 3;27(8):1160. doi: 10.1200/JCO.2008.18.1370.
[6] González-Reymúndez A, de Los Campos G, Gutiérrez L, Lunt SY, Vazquez AI. Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. European Journal of Human Genetics. 2017 May;25(5):538-44. doi: 10.1038/ejhg.2017.12.
[7] Kumar L, Greiner R. Gene expression based survival prediction for cancer patients—A topic modeling approach. PloS one. 2019 Nov 15;14(11):e0224446. doi: 10.1371/journal.pone.0224446.
[8] Brazma A, Vilo J. Gene expression data analysis. FEBS letters. 2000 Aug 25;480(1):17-24. doi: 10.1016/s0014-5793(00)01772-5.
[9] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Jun 21;486(7403):346-52. doi: 10.1038/nature10983.
[10] Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, Saxena A. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nature communications. 2018 Jan 3;9(1):42. doi: 10.1038/s41467-017-02465-5.
[11] Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European heart journal. 2017 Feb 14;38(7):500-7. doi: 10.1093/eurheartj/ehw188.
[12] Raudvere U, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research. 2019 Jul 2;47(W1):W191-8. doi: 10.1093/nar/gkz369.