THE JOURNAL OF COGNITIVE SYSTEMS
Journal Homepage: https://dergipark.org.tr/en/pub/jcs
Vol.7, No.2, 2022

Genomic Biomarkers of Metastasis in Breast Cancer Patients: A Machine Learning Approach

Feyza İnceoğlu[1], **Fatma Hilal Yagin**[2]

[1] Department of Biostatistics, Faculty of Medicine, Malatya Turgut Özal University, Malatya 44000, Turkey (e-mail: feyza.inceoglu@ozal.edu.tr).
[2] Department of Biostatistics, and Medical Informatics, Faculty of Medicine, Inonu University, Malatya 44000, Turkey (e-mail: hilal.yagin@inonu.edu.tr).

ARTICLE INFO

Received: July.,03.2022
Revised: Sep., 21.2022
Accepted: Oct.,17.2022

Keywords:
Breast Canser,
Gene Analysis,
Machine Learning,
XGBoost

Corresponding author: Feyza Inceoglu
✉ feyza.inceoglu@inonu.edu.tr
☎ +90 506 9316595

ISSN: 2548-0650

DOI: https://doi.org/10.52876/jcs.1211185

ABSTRACT

One of the cancers with the highest incidence in the world is breast cancer (BC). The aim of this study is to identify candidate biomarker genes to predict the risk of distant metastases in patients with BC and to compare the performance of machine learning (ML) based models. In the study; Genomic dataset containing 24,481 gene expression levels of 97 patients with BC was analyzed. Biomarker candidate genes were determined by ML approaches and models were created with XGBoost, naive bayes (NB) and multilayer perceptron (MLP) algorithms. The accuracy values of XGBoost, NB and MLP algorithms were obtained as 0.990, 0.907 and 0.979, respectively. Our results showed that XGBoost has higher performance. The top five genes associated with BC metastasis were AL080059, Ubiquilin 1, CA9, PEX12, and CCN4. In conclusion, when the ML method and genomic technology are used together, the distant metastasis risk of patients with BC can be successfully predicted. The developed XGBoost model can distinguish patients with distant metastases. Identified biomarker candidate genes may contribute to diagnostic, therapeutic and drug development research in patients with metastases.

1. INTRODUCTION

THE cancer is the health problem with the highest incidence of cardiovascular diseases in the world. In the report published by the International Agency for Research on Cancer in 2012, with information on 184 countries, it was announced that 14.1 million new cancer cases and 8.2 million deaths were due to cancer [1, 2]. Breast cancer (BC) is a systemic disease that occurs as a result of the rapid spread and proliferation of cells in the mammary glands and ducts of the breast to tissues in different parts of the body. Cancer types with the highest incidence are lung (13%), breast (11.9%) and colon (9.7%) cancers in the world. The cancer types with the highest mortality are lung (19.4%), liver (9.1%) and stomach (8.8%) cancers. The predicted cancer cases for 2025 are 19.3 million [3, 4].

Risk factors affecting BC are demographic variables, hormonal system changes, lifestyles, and benign breast anomalies, environmental and hereditary factors [4, 5]. Gene analysis has an important place in the determination of genetic factors. Mutations in oncogene/anthioncogene structures affect processes and formations. For BC; HER 2 and HER 1 (c-erbB-2 and 1), Ras, c-Mys, TP53, BRCA1, BRCA2, STK11, PTEN, CDH1, CHRK, ATM, PALB2 genes have been revealed in previous studies [5].

Gene expression, which forms the basis of analyzes in molecular structure studies for BC, is used in the diagnosis and treatment of BC [6, 7]. The use of large-scale genomic analyzes in today's studies reveals complex structures. In most of the studies, the relationship between tumor metastasis and widespread mutational structures were examined clonally [8]. Microarray technology developing depending on the developments in medical technologies; It offers researchers the opportunity to measure more than one gene structure at the same time. The gene structures of the diseased and non-disease groups are clearly analyzed for distinguishing features. Although it is difficult for researchers to analyze with multidimensional data, these difficulties disappear with ML [9].

ML offers researchers a dynamic analysis process and is frequently used. In ML, many different processes such as classification, summarization, clustering of data, methods of

establishing variable models are applied. ML also presents unobservable relationships in large databases to researchers. In the ML database system, this unobservable information is made with database technologies, modeling methods, statistical and mathematical analysis [10, 11].

When constructing ML classification models on high-dimensional microarray datasets, biomarker candidate genes associated with the disease of interest must first be identified. It will be possible to improve the performance metrics of the classification model to be created by identifying biomarker candidate genes [12].

Microarray technology, which is used together with ML, is a method that will facilitate the early diagnosis of BC. With Microarray technology, researchers can analyze the expression level of thousands of genes simultaneously and qualitatively. Accurate classification techniques for BC prognosis and treatment process with ML will also help clinicians. A prediction model is created by analyzing complex BC datasets with ML [13].

The aim of this study; to establish a supportive clinical prediction model for early diagnosis by identifying biomarker candidate genes that cause metastasis in patients with BC with ML approaches.

2. MATERIAL AND METHODS

2.1. Data

Gene expression data were obtained from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) database and analyzed. The study examined 24,481 gene expression levels in 97 patients with BC, 46 (47%) of whom had developed distant metastases within 5 years and 51 (53%) were lymph node negative (pN0) who did not develop distant metastases [14, 15].

2.2. Methods

2.2.1. Machine Learning Approach

Feature selection for genomic data are methods used in machine learning to shorten analysis time, identify disease/state of interest-associated biomarker candidate genes, and improve the performance of predictive models. In the study, recursive feature elimination (RFE) method based on logistic regression (LR) classifier was used to select candidate gene biomarkers associated with BC metastasis. RFE is a popular method as it is effective in selecting the features that are most relevant for estimating the target feature RFE is a wrapper feature selection method. The method selects features using a machine learning algorithm [16].
The LR used in the RFE is a method that uses the maximum likelihood estimation approach for regression and classification tasks and estimates the values of the parameters that maximize the probability obtained [17].

XGBoost, naïve bayes (NB), and multilayer perception (MLP) classifiers were used to predict BC metastasis after feature selection. An algorithm based on decision-tree (DT) and gradient-boosting (GB), XGBoost is a faster running algorithm compared to GB algorithms, with different regularization penalties to avoid overfitting [18]. NB is an algorithm based on conditional probability, which is assumed to be equal and independent from each other in the classification of all attributes based on conditional probability [19].

MLP is a type of neural network used to support feed forward neural networks. In MLP, the input layer receives the signal to be processed and the output layer does the estimation and classification [20, 21].

The 10-fold cross validation (CV) method was used to validate the models. The k-fold CV splits the data into k blocks randomly and the algorithm uses the k-1 block as the training set and the remaining single block as the test set. The process continues until all blocks are used as a test set, and the average of all results represents the overall performance [22]. The performance of the models was evaluated with accuracy, Sensitivity, specificity, positive predictive value, negative predictive value, and F1-score, and the performance results of the models were compared.

3. RESULTS

In Table 1, the results of the performance measures for the models created for BC metastasis prediction are given. When Table 1 is examined, the accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values of the XGBoost model were obtained as 0.990, 0.978, 1.000, 1.000, 0.981, and 0.989, respectively. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values for the NB model were obtained as 0.907, 0.978, 0.843, 0.849, 0.977 and 0.909, respectively.

In the MLP model, accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values were obtained as 0.979, 0.957, 0.843, 0.849, 0.977 and 0.909, respectively. The results showed that the XGBoost model had higher performance in predicting BC metastasis compared to the NB and MLP models.

TABLE I
The Performance of the Models

Metric	XGBoost	NB	MLP
Accuracy	0.990	0.907	0.979
Sensitivity	0.978	0.978	0.957
Specificity	1.000	0.843	1.000
Positive predictive value	1.000	0.849	1.000
Negative predictive value	0.981	0.977	0.962
F1-score	0.989	0.909	0.978

NB: Naive Bayes; MLP: Multilayer Perception

In Table 2 and Figure 1, the importance of genes according to their contribution to the prediction of BC metastasis of the XGBoost model was examined. According to the results of the study, the importance of AL080059, Ubiquilin 1, CA9, PEX12, and CCN4 genes were 100, 90.621, 53.731, 46.485, and 45.775, respectively.

TABLE II
The Genes and Importance

Feature (Gene)	Importance
AL080059	100
Ubiquilin-1	90.621
CA9	53.731
PEX12	46.485
CCN4	45.775
NMU	40.069
SSX2	38.24
ALDH4A1	36.249
RAB5	35.927
ALDH6A1	34.281
ARL4D	31.954
PHF1	29.893
UBE2T	27.616
AF052087	24.889
KIAA0906	21.185
PRAME	20.766
TGFB3	19.256
CDKN3	8.622
SLC37A1	8.217
SCUBE2	7.583

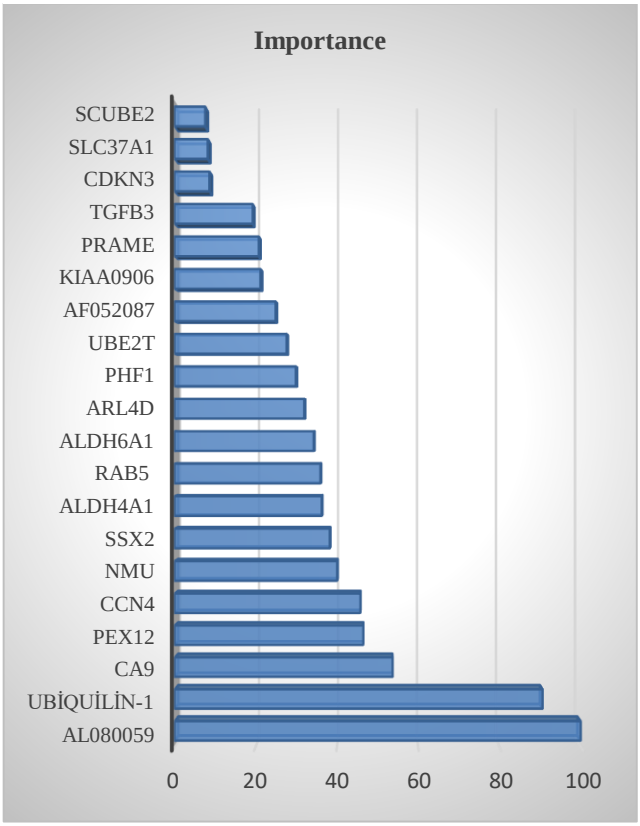

Fig. 1. Importance plot for biomarker genes

4. DISCUSSION

Even though there have been significant advances in the treatment of BC in recent years, the prognosis for the majority of patients who have distant metastasis is still not good. Patients diagnosed with BC who are at the same stage of the disease may have very different responses to treatment and very different outcomes overall. In order to pave the way for earlier detection of metastasis and more effective treatments, it is critical to have an in-depth understanding of the molecular phenotype of distant metastasis. Based on this, the purpose of this study was to predict the presence of distant metastases in BC patients using ML methods that were based on genomic biomarkers [14]. From this point of view, the aim of this study is to predict the presence of distant metastases in BC patients using ML methods based on genomic information and data.

Genomic data including 24,481 gene expression levels of 97 patients with and without metastasis were used in the study. Genomic data containing thousands of gene expression levels belonging to a small number of patients in ML models require some preprocessing at the analysis stage due to their high dimensionality. Therefore, before creating ML models in the study, biomarker candidate genes were selected by LR-based RFE method. As a result of the analyses, 20 genes associated with BC distant metastasis were identified. Models based on XGBoost, NB and MLP algorithms were created with these biomarker candidate genes. Our results showed that XGBoost has higher performance compared to NB and MLP models. The accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score values of the XGBoost model were obtained as 0.990, 0.978, 1.000, 1.000, 0.981, and 0.989, respectively. The first five genes that made a preliminary significant contribution to the prediction performance of the optimal model, XGBoost, and thus to differentiate BC distant metastasis, were AL080059, Ubiquilin 1, CA9, PEX12, and CCN4. The results of the trait significance analysis showed that the significance of the AL080059, Ubiquilin 1, CA9, PEX12, and CCN4 genes were 100, 90,621, 53,731, 46,485, and 45,775, respectively.

Our biomarker gene selection results were similar to the literature. In a study, it was reported that the AL080059 gene had a significant difference in BC patients compared to healthy controls and could be a candidate for a biomarker [23]. In different studies, it was found that UBQLN1 increased abnormally for BC [24, 25]. Similarly, in our study, it was reported that the Ubiquilin 1 (UBQLN1) gene showed a significant difference in BC patients and could be a biomarker candidate. CA9 has an important place in the distribution of tissues in the body and has been found to be an important factor for BC [26]. In our study, CA9 was found to be the third most important risk variable for BC. PEX 12 is an important risk factor in BC as in liver cancer [27-29]. In our study, PEX12 was found to be one of the most important risk variables for BC. CNN proteins enable the activation of signal transduction within the cell. CNN4 gene is effective on cancer because it is effective in cell migration and increases epithelial-mesenchymal transition [30], which was found to be an important risk factor for BC as well in our study. NMU, which affects the invasive capacity of cancer cells, has been identified as a risk factor for BC in different studies [31, 32]. In the factors examined in our study, NMU was found to be a risk factor for BC.

 ISSN: 2548-0650 http://dergipark.gov.tr/jcs

As a result, genes identified in the early diagnosis and treatment of BC distant metastasis can be examined and the XGBoost model can successfully differentiate metastases.

5. CONCLUSIONS

In conclusion, with the methodology combined with genomic technology and ml method, the risk of distant metastasis of patients with bc can be successfully predicted. identified biomarker candidate genes may contribute to diagnosis, treatment and drug development research in patients with metastasis. the developed XGBoost model can distinguish patients with distant metastases.

REFERENCES

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394-424.

[2] Cancer, I. A. f. R. o., & Organization, W. H. (2012). Breast Cancer Estimated Incidence, Mortality and Prevalence Worldwide. Globocan 2012: World Health Organization.

[3] Sowunmi, A., Alabi, A., Fatiregun, O., Olatunji, T., Okoro, U. S., & Etti, A. F. D. (2018). Trend of cancer incidence in an oncology center in Nigeria: West African Journal of Radiology, 25(1), 52.

[4] Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer: Targets and Therapy, 11, 151.

[5] Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., . . . Zhu, H.-P. (2017). Risk factors and preventions of breast cancer: International Journal of Biological Sciences, 13(11), 1387.

[6] Holleczek, B., Stegmaier, C., Radosa, J. C., Solomayer, E.-F., & Brenner, H. (2019). Risk of loco-regional recurrence and distant metastases of patients with invasive breast cancer up to ten years after diagnosis–results from a registry-based study from Germany: Bmc Cancer, 19(1), 1-14.

[7] Anwar, S. L., Avanti, W. S., Nugroho, A. C., Choridah, L., Dwianingsih, E. K., Harahap, W. A., . . . Wulaningsih, W. (2020). Risk factors of distant metastasis after surgery among different breast cancer subtypes: a hospital-based study in Indonesia: World Journal of Surgical Oncology, 18(1), 1-16.

[8] Savas, P., Teo, Z. L., Lefevre, C., Flensburg, C., Caramia, F., Alsop, K., . . . Silva, M. J. (2016). The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program "CASCADE": PLoS medicine, 13(12), e1002204.

[9] Xu, C., Meng, L. B., Duan, Y. C., Cheng, Y. J., Zhang, C. M., Zhou, X., & Huang, C. B. (2019). Screening and identification of biomarkers for systemic sclerosis via microarray technology: International journal of molecular medicine, 44(5), 1753-1770.

[10] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare: Paper presented at the Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics.

[11] Yağin, F. H., Yağin, B., Arslan, A. K., & Çolak, C. (2021). Comparison of Performances of Associative Classification Methods for Cervical Cancer Prediction: Observational Study: Turkiye Klinikleri Journal of Biostatistics, 13(3).

[12] Khaire, U. M., & Dhanalakshmi, R. (2020). High-dimensional microarray dataset classification using an improved adam optimizer (iAdam): Journal of Ambient Intelligence and Humanized Computing, 11(11), 5187-5204.

[13] Vaka, A. R., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning: ICT Express, 6(4), 320-324.

[14] Akbulut, S., Yağın, F. H., & Çolak, C. (2022). Prediction of COVID-19 Based on Genomic Biomarkers of Metagenomic Next-Generation Sequencing (mNGS) Data using Artificial Intelligence Technology: Erciyes Medical Journal.

[15] Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer: nature, 415(6871), 530-536.

[16] Lee, M., Lee, J.-H., & Kim, D.-H. (2022). Gender recognition using optimal gait feature based on recursive feature elimination in normal walking: Expert Systems with Applications, 189, 116040.

[17] Yilmaz, R., & Yağin, F. H. (2022). Early detection of coronary heart disease based on machine learning methods: Medical Records, 4(1), 1-6.

[18] Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Computers: Environment and Urban Systems, 96, 101845.

[19] Paksoy, N., & Yagin, F. H. (2022). Artificial Intelligence-based Colon Cancer Prediction by Identifying Genomic Biomarkers: Medical Records, 4(2), 196-202.

[20] Yilmaz, R., & Yagin, F. H. (2021). A comparative study for the prediction of heart attack risk and associated factors using MLP and RBF neural networks: The Journal of Cognitive Systems, 6(2), 51-54.

[21] Akbulut, S., Yagin, F. H., & Colak, C. (2022). Prediction of Breast Cancer Distant Metastasis by Artificial Intelligence Methods from an Epidemiological Perspective: Istanbul Medical Journal, 23(3).

[22] Perçin, İ., Yağin, F. H., Arslan, A. K., & Çolak, C. (2019). An interactive web tool for classification problems based on machine learning algorithms using java programming language: data classification software: Paper presented at the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).

[23] Song, Q., Jing, H., Wu, H., Zou, B., Zhou, G., & Kambara, H. (2016). Comparative Gene Expression Analysis of Breast Cancer-Related Genes by Multiplex Pyrosequencing Coupled with Sequence Barcodes Advances and Clinical Practice in Pyrosequencing: Springer, 315-325.

[24] Feng, X., Cao, A., Qin, T., Zhang, Q., Fan, S., Wang, B., . . . Li, L. (2021). Abnormally elevated ubiquilin-1 expression in breast cancer regulates metastasis and stemness via AKT signaling: Oncology Reports, 46(5), 1-14.

[25] Jantrapirom, S., Lo Piccolo, L., Pruksakorn, D., Potikanond, S., & Nimlamool, W. (2020). Ubiquilin networking in cancers: Cancers, 12(6), 1586.

[26] Hu, Z., Li, X., Yuan, R., Ring, B. Z., & Su, L. (2010). Three common TP53 polymorphisms in susceptibility to breast cancer, evidence from meta-analysis: Breast cancer research and treatment, 120(3), 705-714.

[27] Moelans, C. B., De Weger, R. A., & Van Diest, P. J. (2010). Absence of chromosome 17 polysomy in breast cancer: analysis by CEP17 chromogenic in situ hybridization and multiplex ligation-dependent probe amplification: Springer, 120, 1-7.

[28] Smeets, A., Daemen, A., Vanden Bempt, I., Gevaert, O., Claes, B., Wildiers, H., . . . De Moor, B. (2011). Prediction of lymph node involvement in breast cancer from primary tumor tissue using gene expression profiling and miRNAs: Breast cancer research and treatment, 129(3), 767-776.

[29] Daskalaki, I., Gkikas, I., & Tavernarakis, N. (2018). Hypoxia and selective autophagy in cancer development and therapy: Frontiers in Cell and Developmental Biology, 6, 104.

[30] Nivison, M. P., & Meier, K. E. (2018). The role of CCN4/WISP-1 in the cancerous phenotype: Cancer Management and Research, 10, 2893.

[31] Wu, Y., McRoberts, K., Berr, S., Frierson, H., Conaway, M., & Theodorescu, D. (2007). Neuromedin U is regulated by the metastasis suppressor RhoGDI2 and is a novel promoter of tumor formation, lung metastasis and cancer cachexia: Oncogene, 26(5), 765-773.

[32] Garczyk, S., Klotz, N., Szczepanski, S., Denecke, B., Antonopoulos, W., Von Stillfried, S., . . . Dahl, E. (2017). Oncogenic features of neuromedin U in breast cancer are associated with NMUR2 expression involving crosstalk with members of the WNT signaling pathway: Oncotarget, 8(22), 36246.

BIOGRAPHIES

Feyza İnceoğlu obtained her BSc. degree in statistics from Gazi University (GU) in 2010. Feyza İnceoğlu obtained her BSc degree in statistics from Gazi University (GU) in 2010. She received the MSc ve PhD diploma in Biostatistics and Medical Informatics from the İnönü University in 2013 and 2018. She started to work as a statistician in the Republic of Turkey State Railways in 2012-2021. In 2021, she started to work as an Assistant Professor in Malatya Turgut Özal University, Faculty of Medicine, Department of Biostatistics and is still working. She works in the fields of validity, reliability, artificial intelligence.

Fatma Hilal Yagin obtained her BSc. degree in Statistics from Gazi University in 2017. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2020. She currently continues Ph.D. education in biostatistics and medical informatics from the Inonu University. In 2019, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, and image processing.