Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2024;13(2):245-9

Machine learning-based biomarkers for breast cancer molecular subtypes HER2+ and TNBC

Seyma Yasar

İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Received 11 November 2023; Accepted 08 December 2023
Available online 24 April 2024 with doi: 10.5455/medscience.2023.11.219

Abstract

Breast cancer is the most common type of cancer and the leading cause of death in women in Türkiye and worldwide. Since breast cancer tumor cells have different genetic characteristics, diagnosis and treatment options vary according to molecular subtypes. The primary objective of this study is to classify human epidermal receptor 2 positive (n=71) and triple-negative breast cancer (n=25) molecular subtypes using two different machine learning algorithms based on tissue proteomics data of 96 breast cancer patients. The secondary aim was to identify possible protein biomarkers that could be used to determine the diagnosis and treatment of these molecular subtypes. The upper sampling method was used to overcome the class imbalance in the study data. The least absolute shrinkage and selection operator (Lasso) was used as the variable selection method. In the modeling phase, Extreme Gradient Boosting (XGBoost) and Bootstrap Aggregating Classification and Regression Trees (Bagged CART) machine learning methods were used. The best classification performance is XGBoost, whose accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, MCC and G-mean values are 96.4%, 96.4%, 100%, 92.9%, 93.3%, 100%, 96.6%, 93.1%, 96.6%, respectively. The three protein access codes found to be most important for the classification of the two molecular subtypes in the optimal model XGBoost result are P02042, P00441 and P20231. These three proteins are thought to be clinically useful markers for early diagnosis and individualized treatment.

Keywords: Breast cancer, human epidermal growth factor receptor 2+, triple negative breast cancer, proteomic, machine learning, extreme gradient boosting

Introduction

A class of disorders known as cancer is defined by unchecked cell division that results from genetic alterations in cells. After cardiovascular disorders, it ranks as the second most prevalent cause of death [1]. According to the World Health Organisation (WHO) 2020 reports, cancer has caused the death of approximately 10 million people. Known causes of cancer include genetic, environmental and individual factors [2]. Use of alcohol, cigarettes, high-fat food intake, and insufficient consumption of fruits and vegetables account for one-third of cancer-related fatalities. There are currently over 100 different forms of cancer. Breast, lung, colon, rectal, and prostate cancers are the most prevalent types of cancer. Breast cancer (BC) is a critical medical problem for the health of most women. Breast health reports indicate that the burden of disease and mortality from breast cancer is higher in developing countries. According to WHO statistics, there were 2.3 million cases of BC in 2020, making BC the most common and most frequently diagnosed disease among women [2]. However, BC mortality rates are gradually decreasing in many countries thanks to the ability to diagnose BC at an early stage and imaging methods used in screening. In addition to these, continuously developed medical treatments reduce the mortality rate of BC. The etiology of BC is not known for certain, but genetic, environmental, psychological and hormonal factors are thought to be effective. Risk factors for breast cancer include gender, late menopause, early menarche, alcohol usage, age, fatty diet, delayed age at first pregnancy, not breastfeeding, ionising radiation, race, and hormone therapy [3]. BC can be classified according to a number of factors, but molecular subtypes can be used to narrow the field down to four types: luminal A, luminal B, triple negative breast cancer (TNBC), and human epidermal growth factor receptor 2 (HER2)+ [4]. HER2-+ breast cancer is a subtype characterized by overexpression of the HER2 protein, which can lead to more aggressive tumor growth. Patients with HER2-

CITATION
Yasar S. Machine learning-based biomarkers for breast cancer molecular subtypes HER2+ and TNBC. Med Science. 2024;13(2):245-9.

Corresponding Author: Seyma Yasar, İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
Email: seyma.yasar@inonu.edu.tr

DOI: 10.5455/medscience.2023.11.219

positive tumors may benefit from targeted therapies such as trastuzumab (Herceptin) or pertuzumab, which specifically target the HER2 protein to inhibit its activity. HER2-positive status is determined through testing, and identifying this subtype helps oncologists tailor treatment plans for individuals with breast cancer, improving the chances of successful outcomes. TNBC is a subtype characterized by the absence of estrogen receptors, progesterone receptors, and HER2 protein expression. TNBC is known for its aggressive nature and limited treatment options compared to other breast cancer subtypes, making it a challenging condition to manage. Research in the field of TNBC is ongoing, with a focus on developing targeted therapies to improve outcomes for individuals diagnosed with this particular subtype of breast cancer. The prognoses, treatment options, and metastasis of these subtypes varied greatly. BC spreads via blood and lymph. The organs to which it metastasises the most are bone, lung, liver, skin, adrenal glands, pleura and brain. Every year, BC takes the lives of around 500,000 women globally, mostly as a result of metastasis, despite advancements in early diagnosis and treatment [5]. Current treatments are often insufficient to completely eliminate cancer cells. Thus, it is necessary to identify novel targets for BC prevention and treatment as well as new processes for the development and recurrence of early-stage BC.

Machine learning is a field of artificial intelligence that involves the development of algorithms and statistical models, allowing computers to learn from and make predictions or decisions based on data without explicit programming. This field is characterised by the ability of algorithms to learn spontaneously based on large amounts of data. Making data-driven decisions and extracting patterns has great potential using machine learning to solve complex problems and predict future events [6]. The applications of machine learning are numerous. Machine learning combined with proteomics holds great promise for understanding diseases, developing drugs for these diseases and understanding personalized medicine. As technology continues to evolve, machine learning is playing an increasingly important role in unlocking the secrets of the proteome and its impact on human health [7].

In this study, these two molecular subtypes of breast cancer will be classified by two different machine learning methods using the data obtained after label-free proteomic analyses of tumour tissues from 25 TNBC and 71 HER2+ BC patients and possible biomarker candidates will be identified according to the optimal classification model.

Material and Methods

Dataset

The data set used in this study consisted of 1095 protein expression profiles from 71 HER2+ and 25 TNBC tissue samples. Proteins from FFPE (formalin-fixed, paraffin-embedded) tumours were analysed by LC/MS-MS [8]. The İnönü University Health Sciences Non-Interventional Clinical Research Ethics Committee approved this study (approval number: 2023/5306).

Data preprocessing

In machine learning, imbalanced data refers to a situation where the distribution of the data set between different classes or categories is uneven. In this case, the results of machine learning models may be biased because larger classes may dominate over smaller classes. Techniques for dealing with unbalanced data include oversampling or undersampling and synthetic data generation [9]. The dataset used in this study is unbalanced (TNCB=25, HER2+=71). Therefore, upsampling method was used to eliminate bias in the results of the machine learning methods to be created. While applying this method, R programming language library caret was used [10]. Another technique used to improve the performance of machine learning methods is the variable selection method. Another technique used to improve the performance of machine learning methods is the variable selection method. Explanation and interpretation of a model with too many variables can become quite complex. Therefore, by variable selection, the model is transformed into a more easily understandable form. With variable selection, overfitting that may occur in the model can be prevented [11]. In this study, Lasso (least absolute shrinkage and selection operator) was used as a variable selection method [12].

Machine learning methods

Extreme gradient boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a powerful open source ensemble learning algorithm used in machine learning. It can be used especially effectively in classification and regression problems. XGBoost is based on the gradient boosting technique and builds a strong model by combining weak learners (usually decision trees). In terms of speed and performance, XGBoost is highly effective, it can run quickly on large datasets. It includes measures against over-extension, the depth of trees can be limited, data subset sampling can be performed and regularisation techniques can be used. With its inherent cross-validation functionality, it can help determine the best parameter settings [13].

Bootstrap aggregating classification and regression trees (Bagged CART)

Bootstrap Aggregating Classification and Regression Trees (Bagged CART) is an ensemble learning method built on classification and regression trees (CART). This technique randomly samples from the dataset (bootstrap method) and then trains a tree on each sample separately. It then aggregates these trees and makes predictions. This helps to reduce the variance of the model and results in more stable predictions. This method can reduce overfitting due to the trees being trained independently and can also provide better generalisation ability. Bagged CART is one of the ensemble learning techniques and is used in various applications. Bagged CART is successfully used in machine learning applications and provides performance enhancing effect especially on large datasets [14].

DOI: 10.5455/medscience.2023.11.219

Model performance evaluation metrics and statistical analysis

The conformity of the proteins in the proteomics dataset to the normal distribution on a group basis was evaluated by Shapiro Wilk test. Proteins that meet the assumption of normal distribution from quantitative variables are summarised by mean±standard deviation, and proteins that do not meet the assumption of normal distribution are summarised by median (Inter Quantile Range). In the training of the models, 80% of the data set was used as training and 20% as test. Then, 5-fold cross-validation method was used on the train dataset for model validation. Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive value, Matthews correlation coefficient (MCC), G-mean and F1-Score metrics were used in the performance evaluation of XGBoost and Bagged CART machine learning models generated to describe possible biomarkers that can be used in the detection and follow-up of two different BC molecular subtypes (HER2+, TNBC).

Results

After Lasso, one of the variable selection methods applied to the data set used in the study, 12 of the 1095 proteins were selected to be included in the modelling. Descriptive statistics for these proteins are given in Table 1.

Performance metrics of the training and testing models of XGBoost and Bagged CART models created with 12 proteins to classify breast cancer molecular subgroups are given in Table 2.

Taking into account the performance metrics in Table 4, the values of the XGBoost model are higher on the test data set. Therefore, the importance values of possible protein biomarkers that may be associated with breast cancer molecular subtypes determined by the XGBoost model are shown in Table 3 and Figure 1.

Table 1. Descriptive statistics of proteins included in the machine learning models

Accession	Protein name	Group		p-value
		HER2 (+)	TNBC	
		Median (IQR)	Median (IQR)	
Q9UQ35	Serine/arginine repetitive matrix protein 2	17.07 (1.39)	18.44 (1.43)	**<0.001***
Q8N2U0	UPF0451 protein C17orf61	17.40 (2.22)	19.64 (1.13)	**<0.001***
Q96C86	Scavenger mRNA-decapping enzyme DcpS	19.22 (0.78)	20.33 (1.10)	**<0.001***
Q01844	RNA-binding protein EWS	19.82 (1.68)	16.83 (1.11)	**<0.001***
H3BMK9	Tryptase beta-2	21.65 (1.49)	17.20 (2.62)	**<0.001***
P00441	Superoxide dismutase [Cu-Zn]	19.38 (1.03)	17.29 (1.89)	**<0.001***
P02042	Hemoglobin subunit delta	23.42 (1.13)	19.31 (3.90)	**<0.001***
P10809	60 kDa heat shock protein. mitochondrial	21.06 (1.31)	19.44 (3.40)	**<0.001***
P98179	Putative RNA-binding protein 3	20.66 (0.71)	20.09 (2.46)	**<0.001***
Q14103	Heterogeneous nuclear ribonucleoprotein D0	22.51 (0.42)	23.20 (0.91)	**<0.001***
P16403	Histone H1.2	23.34 (1.03)	24.83 (2.48)	**<0.001***
P27635	60S ribosomal protein L10	22.07 (0.58)	23.17 (0.88)	**<0.001***

*: Mann-Whitney U test, IQR: inter quantile range

Table 2. The performance metrics of the training and testing models of XGBoost and Bagged CART models

	Machine learning models			
	XGBoost		Bagged CART	
	Training	Testing	Training	Testing
Accuracy	0.965	0.964	0.991	0.893
Balanced accuracy	0.965	0.964	0.991	0.893
Sensitivity	0.982	1	0.982	0.857
Specificity	0.947	0.929	1	0.929
Positive predictive value	0.949	0.933	1	0.923
Negative predictive value	0.982	1	0.983	0.867
F1-score	0.966	0.966	0.991	0.889
MCC	0.93	0.931	0.983	0.788
G-mean	0.965	0.966	0.991	0.894

MCC: Matthews correlation coefficient

DOI: 10.5455/medscience.2023.11.219

Table 3. The importance values of possible protein biomarkers that may be associated with breast cancer molecular subtypes determined by the XGBoost model

Accession	Gene name	Protein name	Variable importance
P02042	HBD	Hemoglobin subunit delta	100.00
P00441	SOD1	Superoxide dismutase	69.17
P20231	TPSB2	Tryptase beta-2	62.00
Q96C86	DCPS	Scavenger mRNA-decapping enzyme DcpS	24.13
Q14103	HNRNPD	Heterogeneous nuclear ribonucleoprotein D0	18.56
P16403	H1-2	Histone H1.2	15.35
P98179	RBM3	Putative RNA-binding protein 3	15.15
Q8N2U0	TMEM256	UPF0451 protein C17orf61	15.01
Q01844	EWSR1	RNA-binding protein EWS	14.52
P10809	HSPD1	60 kDa heat shock protein, mitochondrial	6.87
P27635	RPL10	60S ribosomal protein L10	3.20

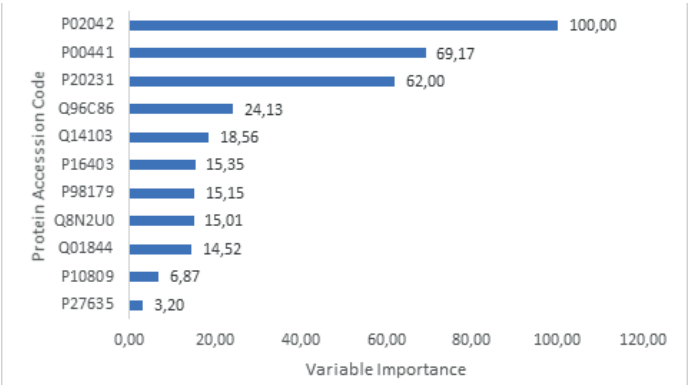

Figure 1. The importance values of possible protein biomarkers that may be associated with breast cancer molecular subtypes determined by the XGBoost model

Discussion

BC is an abnormal and uncontrolled cell proliferation that develops in the ductus and lobules of the breast originating from epithelial cells. BC may develop due to environmental, genetic, hormonal, physiological and sociobiological factors. Although the incidence is increasing, the mortality rate has been decreasing in the last two decades. This decline is a result of the increased use of mammography screening, radiotherapy, surgery and adjuvant chemotherapy. Detection of BC at early stages is very important for successful clinical treatment. Proteomics helps us to understand the changes occurring at the cellular level in breast cancer research [15]. Protein expression in BC cells may be different from healthy cells and this difference may contribute to the development of diagnostic and therapeutic strategies. Analysing and interpreting large datasets obtained as a result of proteomic analyses requires computational approaches such as machine learning techniques. Mass spectrometry data from many biological disciplines is often used to train machine learning algorithms, particularly for different types of cancer. Finding biomarkers to help in the diagnosis, prognosis, and therapy of particular diseases is the goal of this kind of study [16].

Recent omics studies and omics data on the molecular biology of BC have led to the molecular classification of the disease. The inclusion of molecular markers in conventional BC classification systems allows for more effective treatment guidance and the search for potential targets for the development of new therapies [17]. The identification of molecular subtypes of BC is important not only to assess the prognosis of the disease but also to ensure adequate treatment, as subtypes have different drug responses. In this study, two different machine learning methods (XGBoost, Bagged CART) were used to classify two different molecular subtypes of BC, HER2+ and TNBC, using breast tissue proteomic data. Considering the performance metrics of the classification models, XGBoost has the highest value. Accuracy, Balanced Accuracy, Sensitivity, Specificity. Positive Predictive Value. Negative Predictive Value, F1-score, MCC and G-mean values are 96.4%, 96.4%, 100%, 92.9%, 93.3%, 100%, 96.6%, 93.1% and 96.6% respectively. Considering these performance metrics for the optimal model, it is highly successful in classifying two different BC molecular subtypes. As a result of the XGBoost model, P02042, P00441, and P20231 were identified as possible protein biomarker candidates for HER2+ and TNBC classification, which were found to be important and clinically useful in the early diagnosis, prognosis and molecular classification of BC.

In 2021, a study conducted by Lerebous et al. revealed that the HBD gene is upregulated in patients with inflammatory breast cancer (IBC), a very rare form of BC in which the bi hemoglobin protein P02042 accession-encoded hemoglobin subunit delta is expressed, compared to non-IBC patients [18]. On the other hand, biostatistical analyses of the proteomic data used in the study indicated that it was upregulated in the HER2+ molecular subtype of BC and this difference was statistically significant. Therefore, this protein, proposed by the model as a possible biomarker protein candidate, may be clinically useful by investigating this protein in these two molecular types of BC using advanced proteomic analyses.

Based on the research conducted so far, there is limited information indicating a specific association between the SOD1

gene and triple negative breast cancer. However, the role of antioxidant enzymes in the development and spread of cancer in general is a topic under investigation. Studies have shown that high copper levels in tumor tissues of cancer patients are associated with disease progression [19]. The SOD1 gene plays an important role in combating oxidative stress in cells. Oxidative stress is the overproduction of free radicals that damage cells. Copper-dependent superoxide dismutase (SOD1) is an important modulator of oxidative stress in cancer cells [20]. In 2023, Ting et al. showed that upregulation of SOD1 may play a key role in the formation and metastasis of TNBC [21]. Therefore, it is thought that the protein in question can be used in the early diagnosis and prognosis of molecular subtypes in the clinic with proteomic studies to be conducted with a larger sample. Finally, in a study conducted by hao et al. in 2021, they revealed that TPSB2 protein is upregulated in breast cancer patients [22].

Conclusion

As a result, these two molecular subtypes were classified with the machine learning method (XGBoost) based on the data obtained as a result of proteomic analysis of tissues obtained from patients with BC molecular subtypes HER2+ and TNBC, and as a result of the model, clinically useful protein biomarkers that can be used to classify the two molecular subtypes were proposed. Therefore, it is recommended that the results of the article be supported by future studies involving more advanced proteomics or different omics (transcriptomics, metabolomics, etc.) technologies.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
The İnönü University Health Sciences Non-Interventional Clinical Research Ethics Committee approved this study (approval number: 2023/5306).

References

1. Arslan Ates E, Turkyilmaz A, Alavanda C, et al. Multigene panel testing in Turkish hereditary cancer syndrome patients. Medeniyet Med J. 2022;37:150-8.
2. Smolarz B, Nowak AZ, and Romanowicz H. Breast cancer-epidemiology, classification, pathogenesis and treatment (review of literature). Cancers. 2022;14:2569.
3. Watkins EJ. Overview of breast cancer. JAAPA. 2019;32:13-7.
4. Dai X, Xiang L, Li T, et al. Cancer hallmarks, biomarkers and breast cancer molecular subtypes. J Cancer. 2016;7:1281-94.
5. Chen W, Hoffmann AD, Liu H, et al. Organotropism: new insights into molecular mechanisms of breast cancer metastasis. NPJ Precis Oncol. 2018;2:4.
6. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. Electron Markets. 2021;31:685-95.
7. Neely BA, Dorfer V, Martens L, et al. Toward an integrated machine learning model of a proteomics experiment. J Proteome Res. 2023;22:681-96.
8. Gámez-Pozo A, Berges-Soria J, Arevalillo JM, et al. Combined label-free quantitative proteomics and microRNA expression analysis of breast cancer unravel molecular differences with clinical implications. Cancer Res. 2015;75:2243-53.
9. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. ICICS. 2020;1:243-8.
10. Kuhn M, Wing J, Weston S, et al. Package 'caret'. R.J. 2020;223:1-219.
11. Cai J, Luo J, Wang S, et al. Feature selection in machine learning: a new perspective. Neurocomputing. 2018;300:70-9.
12. Fonti V, Belitser E. Feature selection using lasso. 2017;1-25.
13. Dhaliwal SS, Nahid AA, Abbas R. Effective intrusion detection system using XGBoost. Information. 2018;9:149.
14. Omurlu IK, Cantas F, Ture M, et al. An empirical study on performances of multilayer perceptron, logistic regression, ANFIS, KNN and bagging CART. Int J Stat Manag Syst. 2020;23:827-41.
15. Yanovich G, Agmon H, Harel M, et al. Clinical proteomics of breast cancer reveals a novel layer of breast cancer classification. Cancer Res. 2018;78:6001-10.
16. Sun R, Ge W, Zhu Y, et al. Proteomic dynamics of breast cancer cell lines identifies potential therapeutic protein targets. Mol Cell Proteomics. 2023;22:100602.
17. Yersal O, Barutca S. Biological subtypes of breast cancer: prognostic and therapeutic implications. World J Clin Oncol. 2014;5:412-24.
18. Lerebours F, Vacher S, Guinebretiere J, et al. Hemoglobin overexpression and splice signature as new features of inflammatory breast cancer?. J Adv Res. 2020;28:77-85.
19. Gupte A, Mumper RJ. Elevated copper and oxidative stress in cancer cells as a target for cancer treatment. Cancer Treat Rev. 2009;35:32-46.
20. Papa L, Hahn M, Marsh EL, et al. SOD2 to SOD1 switch in breast cancer. J Biol Chem. 2014;289:5412-6.
21. Liu T, Li K, Zhang Z, et al. Tetrandrine inhibits cancer stem cell characteristics and epithelial to mesenchymal transition in triple-negative breast cancer via SOD1/ROS signaling pathway. Am J Chinese Med. 2023;51:425-44.
22. Hao Y, Yang W, Zheng W, et al. Tumor elastography and its association with cell-free tumor DNA in the plasma of breast tumor patients: a pilot study. Quant Imaging Med Surg. 2021;11:3518-34.