Society of Medical Education & Research

2022, Volume 2, Issue 1, Page No: 17-23

Archive of International Journal of Cancer and Allied Science

Predicting Breast Cancer Risk Using Stochastic Gradient Boosting

Orhan Kayira[1*]

[1] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Recep Tayyip Erdogan University, Rize, Turkey.

***E-mail** ✉ Orhan.Kayira @erdogan.edu.tr

Abstract

Breast cancer is a major global health concern and ranks as one of the leading causes of cancer-related mortality in women. This research focuses on utilizing the stochastic gradient boosting (SGB) technique to classify breast cancer data from open-access sources and identify key risk factors. The dataset was employed to develop a classification model, with SGB used for the disease classification process. The performance of the model was assessed using metrics such as accuracy, balanced accuracy, sensitivity, specificity, and both positive and negative predictive values. The SGB model achieved perfect results with 100% in all key metrics, including accuracy, sensitivity, specificity, and the F1 score. Additionally, the study identified the most significant risk factors, including "cave_points_mean," "area_worst," "perimeter_worst," and "concave_points_worst," which were found to have the highest importance. The findings suggest that the SGB-based model can effectively differentiate between breast cancer patients and healthy individuals while also pinpointing critical risk factors, thus contributing to more accurate diagnosis and risk prediction.

Keywords: Machine learning, Breast cancer, Stochastic Gradient Boosting, Risk factors, Classification

Introduction

Breast cancer remains a major public health issue worldwide and is among the most fatal cancers for women [1]. Early diagnosis is critical, as it can significantly influence treatment outcomes and patient survival. Traditional methods for diagnosing breast cancer typically begin with a physical examination, followed by a review of the patient's medical history. Depending on the findings, further diagnostic tests such as mammography, breast ultrasound, and ductoscopy (which uses a fiber-optic device to examine milk ducts) may be performed. Additional methods like ductography (contrast imaging of the ducts) and magnetic resonance imaging (MRI) may also be employed when necessary [2].

Data mining, also known as knowledge discovery from large datasets, plays a key role in identifying patterns and making predictions based on hidden data relationships [3]. Machine learning algorithms, which focus on classification, regression, and association rule mining, enable systems to learn from historical data and make predictions on new data during the validation and testing phases [4].

This research aims to apply the stochastic gradient boosting (SGB) method to differentiate between patients with and without breast cancer. Additionally, the study seeks to identify significant risk factors for breast cancer and assess the importance of these factors concerning cancer diagnosis.

Materials and Methods

Dataset

The dataset for this study was sourced from the UCI Machine Learning Repository (available at

Access this article online

Website: https://smerpub.com/ **E-ISSN: 3108-4834**

Received: 16 November 2021; **Revised:** 04 February 2022; **Accepted:** 05 February 2022

How to cite this article: Kayira O. Predicting Breast Cancer Risk Using Stochastic Gradient Boosting. Arch Int J Cancer Allied Sci. 2022;2(1):17-23. https://doi.org/10.51847/WIXAgukWkR

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29) and was used to predict the presence or absence of breast cancer through the SGB method [5]. The data includes features derived from fine needle aspiration (FNA) biopsies of breast masses, with each feature representing characteristics of the cell nuclei in the image. Key attributes include an ID number, the diagnosis (M = malignant, B = benign), and ten quantitative features describing the shape and texture of the cell nuclei. These features include the mean radius (distance from the center to the perimeter), texture (variability in gray-scale values), perimeter, area, smoothness (variation in radius), compactness (perimeter^2 / area - 1.0), concavity (depth of indentations on the contour), concave points (number of indentations), symmetry, and fractal dimension (approximates the complexity of the shape). For each image, the mean, standard error, and worst (maximum of the three highest values) of these characteristics were computed, leading to 30 features in total. For example, "field 3" corresponds to the mean radius, "field 13" represents the standard error of the radius, and "field 23" denotes the worst radius. The data entries are recorded with four decimal places, and there are no missing values. The dataset consists of 63% benign cases (357) and 37% malignant cases (212).

Stochastic gradient boosting

Boosting, an ensemble learning strategy serves as a meta-classifier and is utilized for predictive modeling [6]. Stochastic gradient boosting (SGB), introduced by Schapire [7], is an important tool for improving prediction accuracy and is particularly beneficial in classification tasks when combined with preprocessing steps. Commonly used methods like XGB, SGB, and Lasso are particularly popular in the realm of breast cancer and mammography research and are also highly effective in identifying significant predictive variables within health-related datasets. These techniques are ideal when working with complex interactions among variables that may be nonlinear or have high dimensionality [8]. SGB implementation was achieved using R's Generalized Boosted Regression Models (GBM) package [9]. Key parameters for fine-tuning the SGB classifier include values for n.trees, shrinkage, and n.minobsinnode.

Data analysis

To test for multivariate normality, the Henze-Zirkler test was employed. For summarizing quantitative data, median values (along with minimum and maximum ranges) were used, while categorical variables were reported in terms of counts and percentages. The Mann-Whitney U test was used to assess whether there were significant differences in the target variable. Relationships between the variables were analyzed using the Spearman correlation coefficient. Model fitting was evaluated via the Likelihood Ratio Test, with a significance level set at $P < 0.05$. IBM SPSS Statistics 26.0 was the software used for the analysis.

Modeling

SGB, as one of the machine learning algorithms, was employed in this study's modeling process. The analysis was conducted using the 100 repeated bootstrap method. Performance was evaluated through various metrics, including balanced accuracy, accuracy, sensitivity, specificity, positive/negative predictive values, and F1-score. To optimize parameters, grid search was applied, with surface selection used to determine the optimal depth. The 5-fold cross-validation method was utilized for resampling.

Results and Discussion

The dataset consisted of 357 (63%) benign and 212 (37%) malignant breast cancer patients. **Figure 1** displays the distribution of the variables.

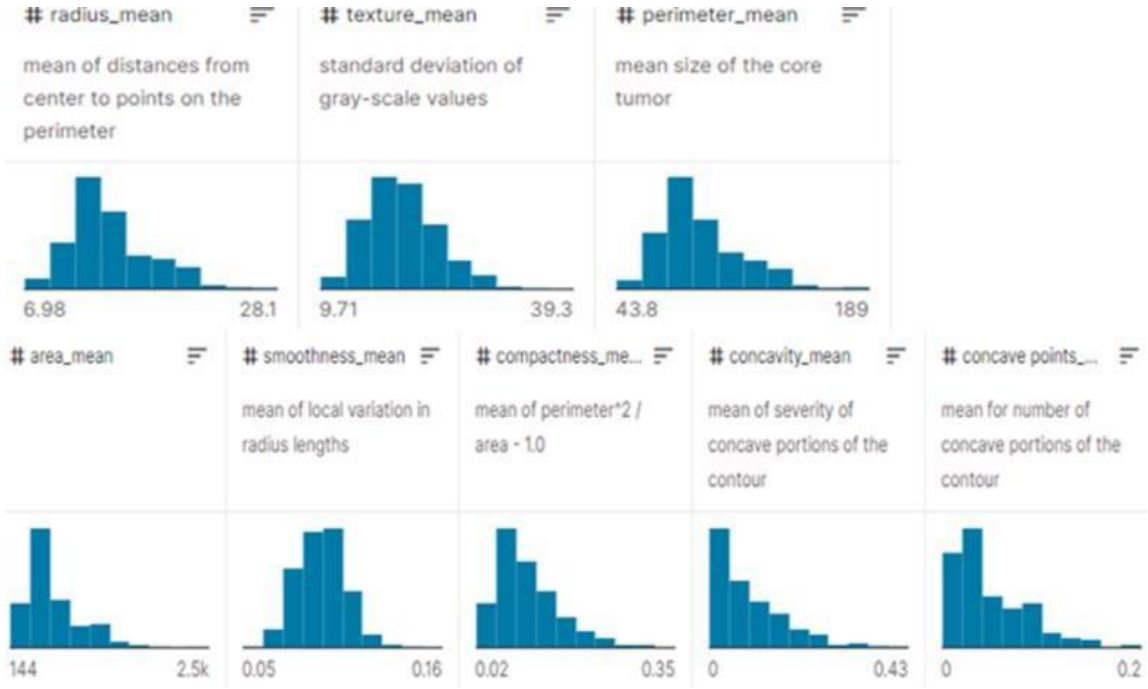

Figure 1. Distributions of the variables

Descriptive statistics for the variables analyzed in this research are shown in Table 1

A notable difference was observed between the diagnosis groups for all variables, except radius_mean, fractal_dimension_mean, and smoothness_se, where no significant variation was found ($P < 0.001$).

Table 1. Descriptive statistics for variables

Variables	Breast Cancer		P-value
	Bening (357)	Malignant (212)	
	Median (Min-Maks)	Median (Min-Maks)	
Radius_mean	21.1 (6.9-47.4)	19.4 (13.0-44.9)	0.108
Texture_mean	19.2 (13.1-47.0)	22.2 (14.2-44.9)	< 0.001
Perimeter_mean	78.1 (43.7-114.6)	114.2 (71.9-188.5)	< 0.001
Area_mean	458.4 (143.5-992.1)	932.0 (361.6-2501)	< 0.001
Smoothness_mean	0.09 (0.05-0.16)	0.1 (0.07-0.14)	< 0.001
Compactness_mean	0.08 (0.02-0.22)	0.13 (0.05-0.35)	< 0.001
Concavity_mean	0.04 (0.0-0.41)	0.15 (0.02-0.43)	< 0.001
Concave points_mean	0.02 (0.0-0.09)	0.09 (0.02-0.2)	< 0.001
Symmetry_mean	0.17 (0.11-0.27)	0.19 (0.13-0.3)	< 0.001
Fractal_dimension_mean	0.06 (0.05-0.1)	0.06 (0.05-0.1)	0.612
Radius_se	0.26 (0.11-0.88)	0.54 (0.19-1.35)	< 0.001
Texture_se	1.14 (0.36-1.76)	1.14 (0.36-1.67)	< 0.001
Perimeter_se	1.94 (0.76-4.56)	3.84 (1.33-4.67)	< 0.001
Area_se	23.24 (6.8-47.1)	60.01 (13.99-44.83)	< 0.001
Smoothness_se	0.01 (0.0-0.02)	0.01 (0.0-0.03)	0.749
Compactness_se	0.02 (0.0-0.11)	0.03 (0.01-0.14)	< 0.001
Concavity_se	0.02 (0.0-0.4)	0.04 (0.01-0.14)	< 0.001
Concave points_se	0.01 (0.0-0.05)	0.01 (0.01-0.04)	< 0.001
Symmetry_se	0.02 (0.01-0.06)	0.02 (0.01-0.08)	0.018
Fractal_dimension_se	0.0 (0.0-0.03)	0.0 (0.0-0.01)	0.007

The correlation matrix is presented in **Figure 2** of this study, which includes a series of numerical values ranging from -1 to 1. A value of 1 indicates a strong positive correlation between variables, such as mean radius and area. A correlation of 0 implies no relationship between variables, like radius_mean and fractal_dimension_se. A value of -1 suggests a perfect negative correlation, although, in this study, the correlation between radius_mean and fractal_dimension_mean is -0.3, indicating a weak negative relationship. This negative correlation still reflects an inverse relationship. The data shows a minimal or no correlation between radius_mean and concavity_mean, but it is statistically significant (r = -0.0917, P = 0.029). This might be due to the large sample size, though the clinical relevance could be questioned. Similarly, no significant correlation was found between radius_mean and texture_mean (r = -0.0113, P = 0.789), but there is a weak positive and statistically significant relationship between texture_mean and texture_worst (r = 0.2318, P < 0.001). Similar interpretations can be made for other variable relationships.

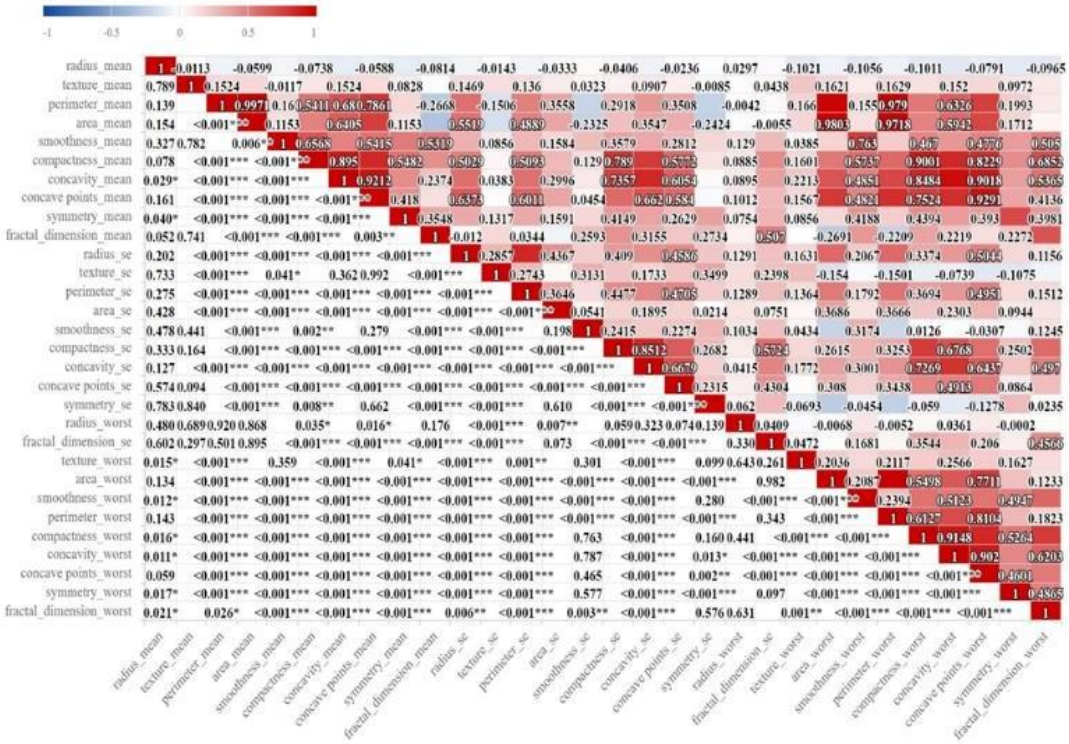

Figure 2. Correlation matrix of variables

The results for the SGB model's performance metrics are shown in **Table 2**. All performance metrics, including accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score, reached 100%. The model's fit was assessed using the likelihood ratio tests (chi-square = 751.44, df = 1, P-value < 0.001).

Table 2. Performance metrics of the SGB model

Metric	Value (%)
Accuracy	100
Balanced Accuracy	100
Sensitivity	100
Specificity	100
Positive predictive value	100
Negative predictive value	100
F1 score	100

Variable importance obtained as a result of SGB modeling is given in **Table 3**. **Figure 3** shows the importance levels of genes that are important for the SGB model.

Table 3. Variable importance of SGB

Variable	Importance	Variable	Importance
Cave.points_mean	100	Radius_se	0.22
Area_worst	68.18	Compactness_mean	0.22
Perimeter_worst	32.19	Smoothness_mean	0.21
Concave.points_worst	24.78	Symmetry_mean	0.18
Texture_worst	8.18	Concave.points_se	0.14
Texture_mean	2.67	Perimeter_se	0.14
Smoothness_worst	2.23	Smoothness_se	0.11
Area_se	1.66	Radius_worst	0.09
Concavity_worst	1.46	Fractal_dimension_worst	0.08
Compactness_worst	1.44	Concavity_se	0.07
Symmetry_worst	1.33	Fractal_dimension_se	0.06
Texture_se	1,00	Perimeter_mean	0.001
Symmetry_se	0.72	Concavity_mean	0.001
Fractal_dimension_mean	0.50	Area_mean	0.00
Compactness_se	0.37	Radius_mean	0.00

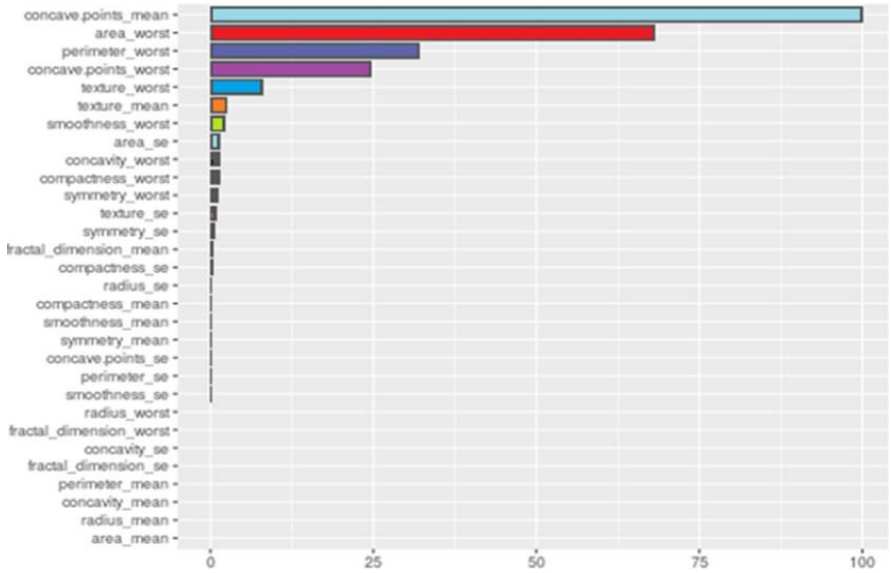

Figure 3. Variable importance of SGB

This study focused on developing a model for accurately distinguishing between benign and malignant breast cancer cells using stochastic gradient boosting (SGB) and digitized data from fine needle aspiration (FNA) images of breast mass samples.

Although machine learning techniques have shown strong performance in classifying breast cancer histopathological images, the effectiveness of these models can be impacted by the size of the data and the complexity of the model architecture. Ensemble learning techniques, a specific branch of machine learning, offer advantages over traditional methods due to their layered approach, which tends to improve classification performance for cancer detection. This enhanced

performance is a significant reason why ensemble learning is gaining popularity in medical fields.

In a recent study, various classification methods, such as support vector machines, k-nearest neighbors, and probabilistic neural networks, were employed to differentiate between benign and malignant breast tumors. These methods incorporated feature selection techniques like signal-to-noise ratio sequencing, forward selection, and principal component analysis. The support vector machine classifier demonstrated impressive diagnostic accuracy, achieving 98.80% and 96.33% accuracy in distinguishing tumors from two widely used breast cancer datasets [10].

A different study compared the performance of two classifiers, Naive Bayes (NB) and k-nearest neighbors (KNN), for breast cancer classification. When evaluated using cross-validation, the KNN classifier yielded the highest accuracy of 97.51%, followed by NB at 96.19% [11].

In another paper, a comparison was made between different machine learning algorithms— support vector machine (SVM), decision trees (C4.5), Naive Bayes (NB), and k-nearest neighbor (k-NN)—using the Wisconsin breast cancer dataset. The results indicated that SVM outperformed the others with an accuracy of 97.13% and the lowest error rate. All experiments were performed using the WEKA data mining tool in a simulated environment [12].

Similarly, Aamir *et al.* developed a model to classify breast cancer risk using various supervised machine learning algorithms, including SVM, random forest, gradient boosting, artificial neural networks, and multilayer perceptron models. Their results showed that the multilayer perceptron model achieved the highest accuracy at 99.12%, outperforming all other models tested [13].

Another study focused on predicting breast cancer survival by using machine learning models to identify key prognostic indicators. Among the models evaluated, random forests provided the highest accuracy at 82.7%, surpassing other models like decision trees, which showed lower performance [14]. Similarly, Mirsadeghi *et al.* [15] employed ensemble learning algorithms to identify driver genes in metastatic breast cancer. The use of algorithms like SVM, ANN, Random Forest, and EARN resulted in outstanding performance, with ROC-AUC scores of 99.24% for metastatic breast cancer and 99.79% for breast cancer overall [15].

An important distinction of this study from previous works is the open-source analysis conducted using the R programming environment. The proposed algorithm demonstrates that the stochastic gradient boosting (SGB) model achieves high classification performance for distinguishing between benign and malignant breast cancer. Consequently, it is recommended that the SGB architecture be considered for applications in breast cancer classification.

While the model shows excellent classification performance, there is still room for improvement. The system can be enhanced by incorporating more data or experimenting with other ensemble learning techniques to further boost its validity and reliability. This approach aims to provide high-precision diagnostic results and could potentially be integrated into clinical decision-making processes for better patient outcomes.

Conclusion

Negative emotional reactions, such as anger and fear, to stressful life events, are normal, temporary responses to perceived threats and do not necessarily indicate a psychopathological condition. In the case of breast cancer, the emotional and psychological toll can be profound, affecting patients' health-related quality of life. Psychological responses to the diagnosis and treatment of breast cancer can evolve, with significant fluctuations depending on the stage of treatment and individual clinical circumstances. Patients often experience considerable psychological strain, including pain, fatigue, insomnia, and emotional stress that affects their social interactions and activities [16].

In this study, a boosting algorithm paired with a k-fold cross-validation method was utilized for classifying data derived from digitized histopathological images of breast cancer. The Stochastic Gradient Boosting method, an ensemble learning approach, enabled highly accurate classification results.

In conclusion, the study highlights that the boosting-based model provided promising predictions for classifying breast cancer (benign vs. malignant) and could serve as a reliable tool for clinical breast cancer diagnosis in the future.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424. doi:10.3322/caac.21492
2. Gazioğlu D, Büyükaşik O, Hasdemir AO, Kargici H. BIRADS 3 ve 4 meme lezyonlarına yaklaşım: Hangi olgulara biyopsi yapılmalı?. J Turgut Ozal Med Cent. 2009;16(1):19-24.
3. Akpınar H. Veri tabanlarında bilgi keşfi ve veri madenciliği. İstanbul Üniv İşlet Fak Derg. 2000;29(1):1-22.
4. Polikar R. Ensemble learning. InEnsemble machine learning 2012 (pp. 1- 34). Springer, Boston, MA.
5. Wolberg WH, Street WN, Heisey DM, Mangasarian OL. Computer-derived nuclear features distinguish malignant from benign breast cytology. Hum Pathol. 1995;26(7):792-6. doi:10.1016/0046-8177(95)90229-5
6. Arslan A, Şen B. Detection of non-coding RNA's with optimized support vector machines. In 2015 23nd Signal Processing and Communications Applications Conference (SIU) 2015 May 16 (pp. 1668-1671). IEEE. doi:10.1109/SIU.2015.7130172
7. Schapire RE. The boosting approach to machine learning: An overview. Nonlinear estimation and classification. 2003:149-71.
8. Sun CK, Tang YX, Liu TC, Lu CJ. An Integrated Machine Learning Scheme for Predicting Mammographic Anomalies in High-Risk Individuals Using Questionnaire-Based Predictors. Int J Environ Res Public Health. 2022;19(15):9756. doi:10.3390/ijerph19159756
9. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367-78.
10. Osareh A, Shadgar B. Machine learning techniques to diagnose breast cancer. In 2010 5th International Symposium on Health Informatics and Bioinformatics 2010 Apr 20 (pp. 114-120). IEEE. doi:10.1109/HIBIT.2010.5478895
11. Amrane M, Oukid S, Gagaoua I, Ensari T. Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) 2018 Apr 18 (pp. 1-4). IEEE. doi:10.1109/EBBT.2018.8391453
12. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci. 2016;83:1064-9. doi:10.1016/j.procs.2016.04.224
13. Aamir S, Rahim A, Aamir Z, Abbasi SF, Khan MS, Alhaisoni M, et al. Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. Comput Math Methods Med. 2022;2022. doi:10.1155/2022/5869529
14. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak. 2019;19(1):1-7. doi:10.1186/s12911-019- 0801-4
15. Mirsadeghi L, Haji Hosseini R, Banaei-Moghaddam AM, Kavousi K. EARN: an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer. BMC Med Genomics. 2021;14(1):1-9.
16. Ranieri J, Di Giacomo D, Guerra F, Cilli E, Martelli A, Ciciarelli V, et al. Early Diagnosis of Melanoma and Breast Cancer in Women: Influence of Body Image Perception. Int J Environ Res Public Health. 2022;19(15):9264. doi:10.3390/ijerph19159264