Original Article

Breast Cancer Risk Prediction with Stochastic Gradient Boosting

Abstract

Breast cancer, which is an important public health problem worldwide, is one of the deadliest cancers in women. This study aims to classify open-access breast cancer data and identify important risk factors with the Stochastic Gradient Boosting Method. The open-access breast cancer dataset was used to construct a classification model in the study. Stochastic Gradient Boosting was used to classify the disease. Balanced accuracy, accuracy, sensitivity, specificity, and positive/negative predictive values were evaluated for model performance. The accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score metrics obtained with the Stochastic Gradient Boosting model were 100 %, 100 %, 100 %, 100 %, 100 %, and 100 %, and 100 % respectively. In addition, the importance of the variables obtained, the most important risk factors for breast cancer were a cave. points_mean, area_worst, and perimeter_worst, concave. points_worst respectively. According to the study results, with the machine-learning model Stochastic Gradient Boosting used, patients with and without breast cancer were classified with high accuracy, and the importance of the variables related to cancer status was determined. Factors with high variable importance can be considered potential risk factors associated with cancer status and can play an essential role in disease diagnosis.

Keywords: *Breast cancer, Machine learning, Ensemble learning, Stochastic gradient boosting*

Mehmet Kivrak[1*]

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

Introduction

Breast cancer, which is a significant public health concern globally, is one of the deadliest cancers in women.[1] Especially in such cancer diseases, early diagnosis is very important and may affect the course of the disease. There are many traditional methods for early detection of breast cancer. First, a physical examination is performed after listening to the patient's medical history. Afterward, ductoscopy (examination of milk ducts by entering very thin fiber-optic systems from the mouth of the canal at the nipple) with imaging methods such as mammography or breast ultrasound may be requested. Additional examinations such as ductography (or galactography, imaging with contrast material from the nipple) and magnetic resonance imaging (MR) may also be requested.[2]

Data mining methods, which are also defined as the process of knowledge discovery from large amounts of data, enable one to make predictions and interpretations about the future by revealing hidden relationship structures.[3] While machine-learning methods that can work based on association rules, classification, and regression perform data-based learning in the training phase, they aim to make predictions about new data in the testing and validation phase.[4]

This study aims to classify patients with and without breast cancer using the Stochastic Gradient Boosting (SGB) method. In addition, it is to determine the risk factors related to breast cancer and to find the variable importance of the cancer-related factor.

Materials and Methods

Dataset

The public dataset "UCI (Machine Learning Repository) Data Set" was obtained from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 to classify the presence or absence of breast cancer via the SGB method in the study.[5] Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. Attribute information is id number, diagnosis (M = malignant, B =

How to cite this article: Kivrak M. Breast Cancer Risk Prediction with Stochastic Gradient Boosting. Clin Cancer Investig J. 2022;11(2):26-31. https://doi.org/10.51847/21qrrkLo4Y

Address for correspondence:
Mehmet Kivrak,
Department of Biostatistics and Medical Informatics,
Faculty of Medicine,
Recep Tayyip Erdogan University,
Rize, Turkey.
E-mail:
mehmet.kivrak@erdogan.edu.tr

Access this article online

Website: www.ccij-online.org

DOI: 10.51847/21qrrkLo4Y

Quick Response Code:

benign), and ten real-valued features are computed for each cell nucleus (radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1)). The mean, standard error, and "worst" or maximum (mean of the three largest values) of these traits were calculated for each image resulting in 30 features. For example, field 3 is Medium Radius, field 13 is SE Radius, and field 23 is Worst Radius. All feature values are recorded in four substantial digits. There are no missing observations in the data set. 63 % of the data set is benign (357) while 37 % is malignant (212).

Stochastic gradient boosting

The boosting algorithm, designed as a meta classifier, is an ensemble learning method that can also make predictions.[6] Stochastic gradient boosting (SGB) is a data processing approach introduced by [7]. SGB is a crucial technique accustomed to creating forecasts and classification tasks and adjusting forecast performance through the appliance of preprocessing procedures. XGB, SGB, and Lasso methods are both widely and successfully used techniques in breast cancer/mammography research. They are also widely used approaches to select critical predictive variables in health and medical informatics applications. XGB or SGB can effectively identify important variables when the variables have nonlinear and/or high dimensional interactions.[8] SGB was implemented in R by the Generalized Boosted Regression Models (GMB) Package.[9] The hyperparameters of the SGB classifier are n.trees, shrinkage, and n.minobsinnode.

Data analysis

Henze zirkler test was used for the assumption of multivariate normality. The median (minimum-maximum) was used to summarize quantitative data, and the numbers were used to summarize qualitative variables (percentages). The Mann-Whitney U test was utilized to see if significant difference in the target exits. The relationship between the variables was evaluated with the spearman correlation coefficient. The model's fit was checked with the Likelihood Ratio Test. P-value <0.05 was regarded as significant. IBM SPSS Statistics 26.0 program was employed in the analysis.

Modeling

SGB, one of the machine learning methods, was used in the modeling. Analyzes were carried out using the 100 repeated bootstrap method. Balanced accuracy, accuracy, sensitivity, specificity, positive/negative predictive values, and F1-score were used as performance evaluation criteria. In parameter optimization, while the surface selection was made as the optimization depth, grid search was used as the scanning method. 5-fold cross-validation was used as a resampling method.

Results and Discussion

In the data set used in the study, there are 357 (63 %) benign cancer patients and 212 (37 %) malignant cancer patients. Distributions of the variables are given **(Figure 1)**.

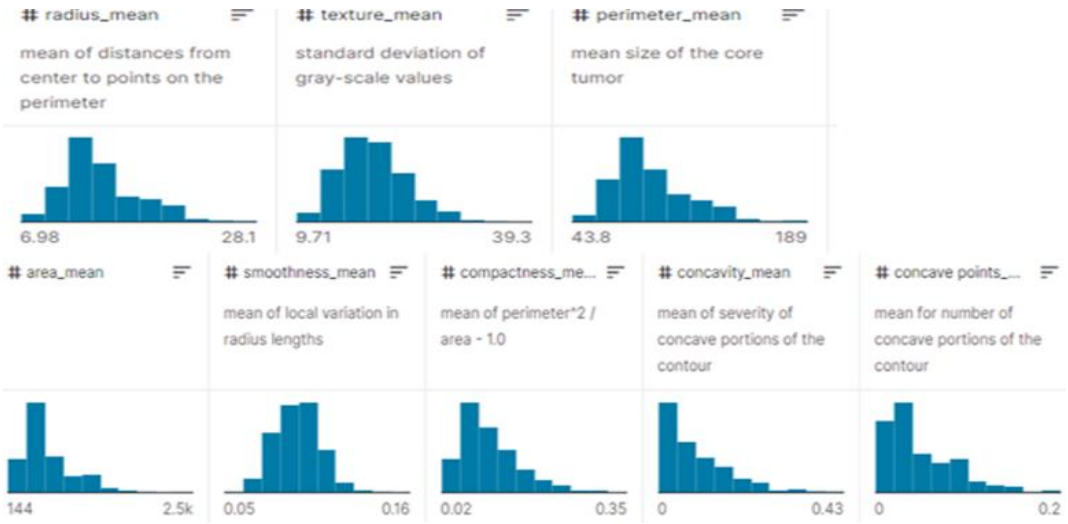

Figure 1. Distributions of the Variables

Descriptive statistics for the variable examined in this study are presented in **(Table 1)**. There is a significant difference between the diagnosis groups regarding other variables (P<0.001) except the radius_mean, fractal_dimension_mean, and smoothness_se variables.

Table 1. Descriptive Statistics for Variables

Variables	Breast Cancer		p
	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 indicated **(Figure 2)** in this study. the matrix that includes a lot of numbers. These numbers range from -1 to 1. A value of 1 mean that the two variables, mean radius and area, are positively interrelated with each other. The zero mean does not correlate with variables such as radial mean and fractal dimension SE. The mean -1 that has two variables, radius and fractal dimension mean are negatively correlated with each other. The correlation between them is not -1, it is -0.3 but the idea is that if the sign of the correlation is negative that means that there is a negative correlation. According to the table, there is no correlation between radius_mean and concavity_mean or there is a very weak negative correlation. However, the relationship is statistically significant. This may be due to the high sample size. Although the result is statistically significant, it may not be clinically significant (r=-0.0917, p=0.029). Similarly, there was no statistically significant correlation between radius_mean and texture_mean variables (r=-0.0113, p=0.789). However, there is a positive, weak and statistically significant correlation between texture_mean and texture_worst variables (r=0.2318, p<0.001). Other variables can be interpreted similarly among themselves.

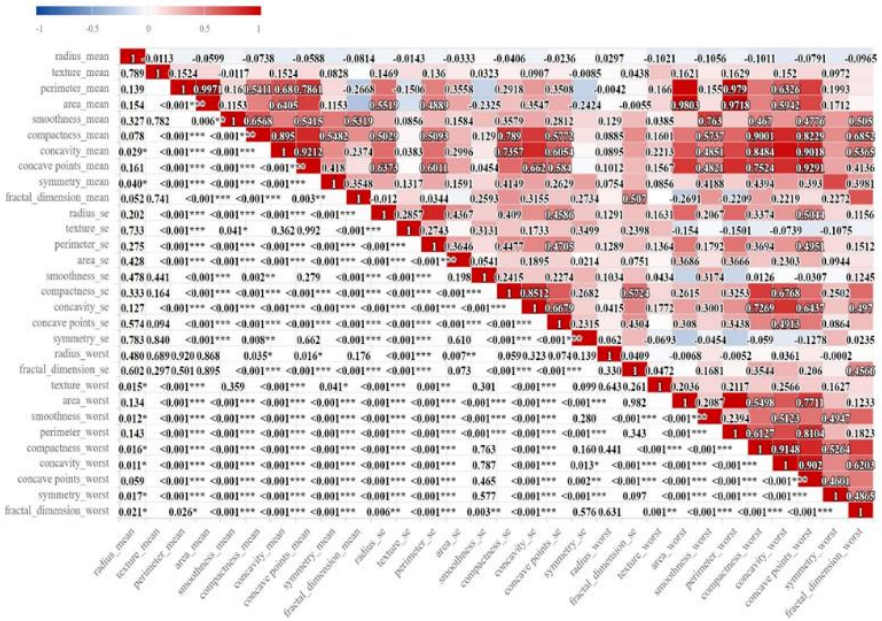

Figure 2. Correlation Matric of Variable

The results of the performance metrics obtained according to the results of the SGB model are given in **(Table 2)**. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score from the SGB model were accounted for 100% respectively. The model's fit was checked with Likelihood Ratio Tests (Chi-Square=751.44, df=1, p-value<0.001).

Table 2. Performance Metrics of 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 Importances 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

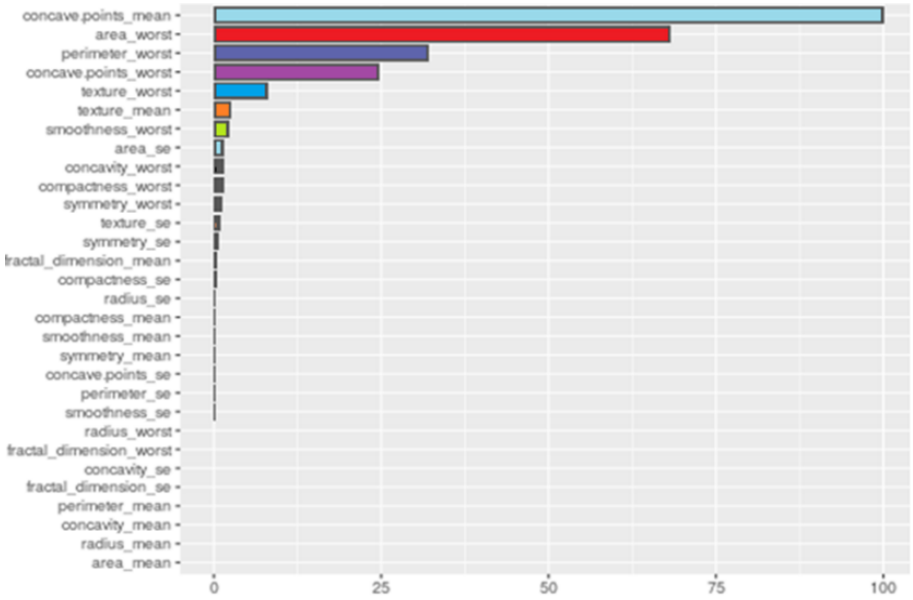

Figure 3. Variable Importances of SGB

In this study, we aimed to develop an algorithm that classifies benign and malignant cells with high accuracy by using the stochastic gradient boosting method with digitized values of cell nuclei obtained from fine needle aspiration (FNA) images on the breast mass.

Although machine learning gives good results in the classification of histopathological images of breast cancer, the increase in data size and architectural edifice affects the category performance results of these techniques. Ensemble-learning systems, a sub-branch of machine learning, on the other hand, are layer-based learning machines than traditional machine learning layers. Therefore, the systems perform better in classifying cancer images compared to ordinary machine learning techniques generated by ensemble learning algorithms. an important feature of community learning. This significant improvement will further facilitate the widespread adoption of ensemble learning in medicine.

In a recent study, support vector machines, k-nearest neighbors, and probabilistic neural network classifiers were used to differentiate benign and malignant tumors of the breast with signal-to-noise ratio feature sequencing, sequential forward selection-based feature selection, and principal component analysis with feature extraction. are combined. The complete accuracy of breast cancer diagnosis is achieved with a support vector machine classifier model against two commonly used breast cancer comparison datasets equaling 98.80 % and 96.33 %, respectively.[10]

In another study, two different classifiers for breast cancer classification were presented: Naive Bayes (NB) classifier and nearest neighbor (KNN). A comparison of two novel implementations was proposed and their accuracy was evaluated using cross-validation. The outcomes illustrate that KNN gives the highest accuracy (97.51 %) with the lowest error rate after the NB classifier (96.19 %).[11]

In an article, a performance comparison between different machine learning algorithms, Support Vector Machine (SVM) in Wisconsin Breast Cancer, Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbor (k-NN) (original) data sets were made. The main purpose is to evaluate the accuracy of each algorithm in classifying data according to its efficiency and effectiveness in terms of accuracy, precision, sensitivity, and specificity. Experimental outcomes display that SVM provides the maximum accuracy (97.13 %) with the lowest error rate. All experiments are conducted in a simulated environment and executed with the WEKA data mining tool.[12] Similarly, Aamir *et al*. developed a cancer risk classification model using supervised machine learning methods such as SVM, Random Forest, Gradient Boost, Artificial Neural Network and Multilayer Perceptron Model in the study called Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. With an accuracy of 99.12%, the Multilayer Sensor Model outperformed all other models reviewed.[13]

In another paper, models were created using machine learning techniques to detect and visualize important prognostic indicators of breast cancer survival rate. In terms of both model accuracy and calibration measurement, all algorithms produced close results from the lowest decision tree (accuracy = 79.8) and highest random forest (accuracy = 82.7).[14] Similarly, in their study on an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer, mirsadeghi *et al*. examined somatic mutation data from 450 metastatic breast tumor samples from the Bio Cancer Genomics Portal. First, the drivers and passengers predicted by SVM, ANN, RF and EARN are introduced. Comparison between outcomes shows that ROC-AUC reaches 99.24% when EARN is used for MBCA and 99.79% for breast cancer.[15]

An important difference between the proposed studies from the aforementioned studies is the open source analysis in the R program. The proposed algorithm shows that the benign and malignant breast cancer classification performances of the SGB model are high. Therefore, it is suggested that SGB architecture can be used in the classification of benign and malignant breast cancer.

Although the model has high classification performance metrics, the system can be retrained by feeding more data or applying other ensemble learning methods to increase model validity and reliability. Thus, it is aimed to obtain high-precision diagnostic results by integrating the developed model into decision-making processes.

Conclusion

Negative emotional responses to stressful life events are a normal short-lived emotional-behavioral response to threat; Anger and fear are not signs of a psychopathological disorder. Breast cancer can have a significant impact on patients' lives, including their health-related quality of life. In addition, psychological responses to diagnoses and treatments can vary significantly over time and with clinical aspects of the healthcare pathway. Throughout the diagnosis and survival process, breast cancer patients report significant declines in their psychological health and quality of life, including greater pain and fatigue, insomnia, and greater interference of stressors (physical and emotional) with social activities.[16]

In this study, a boosting algorithm with a k-fold cross-validation resampling method using all data was used. High accuracy classification of the variables obtained by digitizing the histopathological images with the Stochastic Gradient Boosting method, which is an ensemble learning algorithm, is provided.

In summary, according to the results obtained from the study, it was determined that the model built and designed on boosting gave promising predictions in the classification of breast cancer (benign/malignant) based and could be used for this.

Acknowledgments

In this study, I would like to express my deepest gratitude to the Machine Learning Repository (UCI) open access databases that provided the datasets.

Conflict of interest

None.

Financial support

None.

Ethics statement

Ethics committee approval is not required as a retrospective study is planned on data sets obtained from open access databases.

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. In2015 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. In2010 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. In2018 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