1st International Congress on Cancer Prevention
ICCP 2024
Zanjan, Iran
18-19 September 2024
University of Zanjan

Model Generation and Prediction of Breast Cancer Malignancy Using Machine Learning Algorithms

Hadi Tabesh[*1], Elham Ansari [1], Ardavan Astanei [1]

[1] Faculty of Life Science Engineering, College of Interdisciplinary Science and Technologies, University of Tehran, Iran
Hadi.tabesh@ut.ac.ir
ansarii.elham@gmail.com
aastaneii@gmail.com

* Corresponding Author
Dr.-Ing. Hadi Tabesh
Associate Professor in Biomedical Engineering,
Faculty of Life Science Engineering, Room 318,
College of Interdisciplinary Science and Technologies, University of Tehran.
North Kargar St., 14399-57131 Tehran, Iran
Office: +98-90-22402250
Mobile: +98-912-1546457

Abstract

Healthcare providers continue to face challenges in identifying breast cancer malignancy, despite using mammography and magnetic resonance imaging, which have limitations. As a result, there is a growing interest in machine learning (ML) for its precision in diagnosis and outcome prediction. This study utilized various ML algorithms to create models for diagnosing breast cancer malignancy, using data from the Wisconsin Diagnostic Breast Cancer database (WDBC). Logistic regression and support vector machines (SVM) models were employed to predict breast cancer malignancy. Logistic regression identified four key parameters: bland chromatin, bare nuclei, marginal adhesion, and clump thickness. It should be mentioned that SVM had higher accuracy and area under the ROC curve (0.99). Both of ML models effectively predicted breast cancer malignancy based on these attributes, making them valuable tools in clinical settings for predicting breast cancer malignancy.

Keywords: Breast cancer, Machine Learning, Malignancy, Logistic Regression, Support vector machine

1- Introduction

Breast cancer is a prevalent and severe disease, with an estimated 2.3 million new cases reported annually. It is predicted to account for 11.7% of all cancer cases by 2023, surpassing lung cancer as the most common cancer worldwide. Shockingly, it is responsible for one in four cancer cases and one in six cancer-related deaths among women [1]. Mammography is a common method for detecting breast cancer, but magnetic resonance imaging (MRI) is even more sensitive than mammography [2, 3]. The World Health Organization (WHO) recommends mammography screening every two years for women aged 50 to 69 [4]. Early detection is crucial for effectively combating this disease, as the 5-year survival rate in developed countries is around 90% when diagnosed early [5].

New methods such as contrast-enhanced mammography and dynamic contrast-enhanced MRI have been developed to improve accuracy, but limitations in screening methods can lead to overdiagnosis and overtreatment. ML has emerged as a highly accurate method for diagnosing and predicting breast cancer [6-8].

ML techniques have been used for over thirty years in cancer diagnosis and prediction, with each research group employing unique approaches and datasets, leading to varied outcomes [9-11]. Algorithms like logistic regression, SVM, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), random forest, and decision trees have proven effective in identifying various cancers, including breast, lung, prostate, brain, and colorectal cancer [12-19].

In this research, we created two ML models for diagnosing breast cancer malignancy, demonstrating and comparing their effectiveness for healthcare professionals and presenting hopeful progress in cancer diagnosis

2- Methodology

2-1- Data selection

The dataset on breast cancer was obtained from the WDBC [20], consisting of 699 samples, with 16 samples excluded due to missing values. The attributes of the WDBC are detailed in Tab. 1.

Table 1 : Wisconsin Diagnostic Breast Cancer dataset attributes

Number	Attribute	Abbreviation	Domain
1	Sample code number		Id number
2	Clump Thickness	CT	1 - 10
3	Uniformity of Cell Size	UCSI	1 – 10
4	Uniformity of Cell Shape	UCSH	1 – 10
5	Marginal Adhesion	MA	1 – 10
6	Single Epithelial Cell Size	SECS	1 – 10
7	Bare Nuclei	BN	1 – 10
8	Bland Chromatin	BC	1 – 10
9	Normal Nucleoli	NN	1 – 10
10	Mitoses	M	1 – 10
11	Class		2 for benign 4 for malignant

2-2- Logistic regression

We utilized Minitab 19 software's logistic regression with a 95% confidence level and ten-fold cross-validation as the sampling strategy. The malignancy indicator (Number 4) was considered the response event, and the class parameter was chosen as the response parameter. All parameters in Table 1, except for the sample code number, were defined as continuous predictors. Predictors with p-values greater than 0.05 were considered statistically insignificant and were removed from the final regression model.

2-3- SVM

We applied the "Classification Learner" app in MATLAB 2019b to implement SVM. For data validation, 10-fold cross-validation was used, and various SVM techniques were studied to determine the most accurate model.

3- Results

3-1- Modeling with logistic regression

Eq. (1) is presenting the initial generated regression model while its coefficient table is presented in Tab. 2.

$$P(Y') = \exp(Y')/(1 + \exp(Y'))$$
$$Y' = -10.10 + 0.535\ \text{clump thickness} - 0.006\ \text{uniformity of cell size} + 0.323\ \text{uniformity of cell shape} + 0.331\ \text{marginal adhesion} + 0.097\ \text{single epithelial cell size} + 0.3830\ \text{bare nuclei} + 0.447\ \text{bland chromatin} + 0.213\ \text{normal nucleoli} + 0.535\ \text{mitosis} \quad \text{Eq.1}$$

Table 2: Coefficients table of regression model for WDBC dataset

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	-10.10	1.17	(-12.41, -7.80)	-8.60	0.000	
clump thickness	0.535	0.142	(0.257, 0.813)	3.77	0.000	1.19
uniformity of cell size	-0.006	0.209	(-0.416, 0.404)	-0.03	0.976	2.82
uniformity of cell shape	0.323	0.231	(-0.129, 0.775)	1.40	0.162	2.76
marginal adhesion	0.331	0.123	(0.089, 0.573)	2.68	0.007	1.19
single epithelial cell size	0.097	0.157	(-0.210, 0.404)	0.62	0.537	1.35
bare nuclei	0.3830	0.0938	(0.1991, 0.5670)	4.08	0.000	1.14
bland chromatin	0.447	0.171	(0.111, 0.783)	2.61	0.009	1.21
normal nucleoli	0.213	0.113	(-0.008, 0.434)	1.89	0.059	1.22
Mitosis	0.535	0.329	(-0.110, 1.179)	1.63	0.104	1.04

As depicted in Tab. 2, five parameters had p-values exceeding 0.05. After removing these parameters, a modified regression model was created, considering the remaining four predictors: bland chromatin, bare nuclei, marginal adhesion, and clump thickness. Eq. (2) and Tab. 3 present the modified regression model and its coefficients, respectively.

$$P(Y') = \exp(Y')/(1 + \exp(Y'))$$
$$Y' = -10.11 + 0.812\ \text{clump thickness} + 0.434\ \text{marginal adhesion} + 0.4814\ \text{bare nuclei} + 0.702\ \text{bland chromatin} \quad \text{Eq.2}$$

Table 3: Coefficients table of the modified regression model

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	-10.11	1.03	(-12.14, -8.09)	-9.79	0.000	
clump thickness	0.812	0.126	(0.565, 1.058)	6.45	0.000	1.09
marginal adhesion	0.434	0.114	(0.211, 0.658)	3.81	0.000	1.07
bare nuclei	0.4814	0.0882	(0.3086, 0.6541)	5.46	0.000	1.06
bland chromatin	0.702	0.152	(0.404, 0.999)	4.62	0.000	1.06

Tab. 3 displays the predictor coefficients of the modified regression model. All predictors have positive coefficients, indicating that an increase in their value positively affects breast cancer malignancy. The high Z-value and positive-value regions of the confidence intervals suggest the vital role of the four predictors in the simulation model. Each predictor in Table 3 has a VIF value of less than 10, ensuring the absence of multicollinearity and a high degree of accuracy in predicting future events.

Table 4: Model summaries

Models	*AUC	TPR	FPR	FNR	TNR	Accuracy	Deviance R-Sq	Deviance R-Sq (adj)
Regression model	0.9947	-	-	-	-	-	85.78%	85.33%
Linear SVM	0.99	97.6%	2.4%	5%	95%	96.3%	-	-

*AUC: area under the curve; TPR: true positive rate; FPR: false positive rate;
FNR: false negative rate; TNR: true negative rate.

The deviance R-squared and adjusted deviance R-squared of the regression model are nearly equal, indicating its correct construction. Additionally, the model's area under the ROC curve is 0.9947, signifying its high accuracy in determining tumor malignancy. It should be mentioned that accuracy is a metric to evaluate the performance of prediction models. Accuracy is calculated by the following equation:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Eq.3}$$

3-1-1- Goodness-of-Fit Tests

The P-value for the deviance test (Tab. 5) is 1.000, indicating the model's accuracy in predicting future incidents. However, the model did not pass the Hosmer-Lemeshow and Pearson tests, which assess the fitness of the data. Despite this, the model is well-generated, with a small difference between the observed data and the model, except for the Hosmer-Lemeshow and Pearson tests yielding a P-value of 0.

Table 5: The goodness of fit tests

Test	Degree of Freedom	Chi-Square	P-Value
Deviance	678	125.77	1.000
Pearson	678	1130.90	0.000
Hosmer-Lemeshow	8	41.24	0.000

Residual plots help identify skewed or outlier-filled data. The presence of these cases suggests that the generated model was unable to support the theories. As shown in our plot, a long column in one direction indicates the presence of a deviation. Nonetheless, it is preferable to use the normal probability plot for residual sensing because the histogram plot's appearance is correlated with the number of intervals used to group the data. The residual distribution wasn't normal, as shown by the obtained plot's S curve. Additionally, the confidence interval may have been miscalculated.

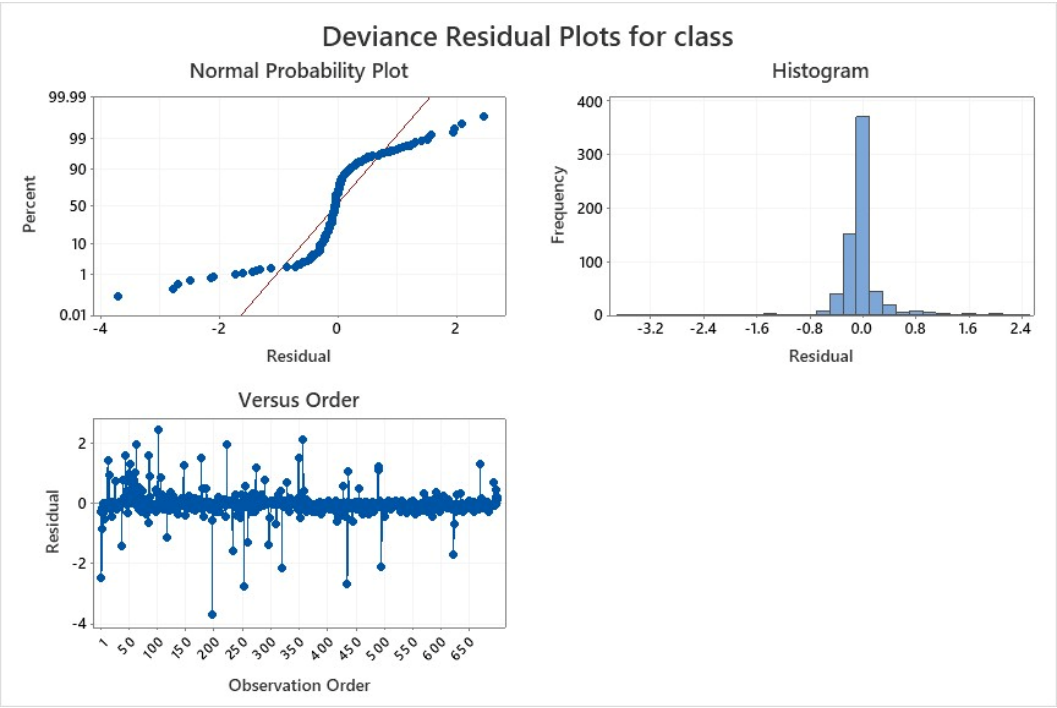

Fig. 1 Deviance residual plots for the class parameter of regression model

3-2- SVM

Linear SVM produced the highest accuracy, 96.3%, among SVM types. As a result, Linear SVM was used to continue the process. Fig. shows a plotted data between bland chromatin and marginal adhesion as an example of scatter plots of linear SVM. According

to the plot, there is not much intervention in the data, and the classification is almost perfectly ordered. Unfortunately, it was not possible to choose the decision boundary appropriately because of how close the data were.

Tab. 4 displays the results of the linear SVM confusion matrix, which indicates that the model predicted incidence with reasonable accuracy. For the malignancy class, the positive predictive value was 95%. The SVM model can accurately classify the data, as evidenced by the area under the ROC curve for the malignancy class, which was 0.99.

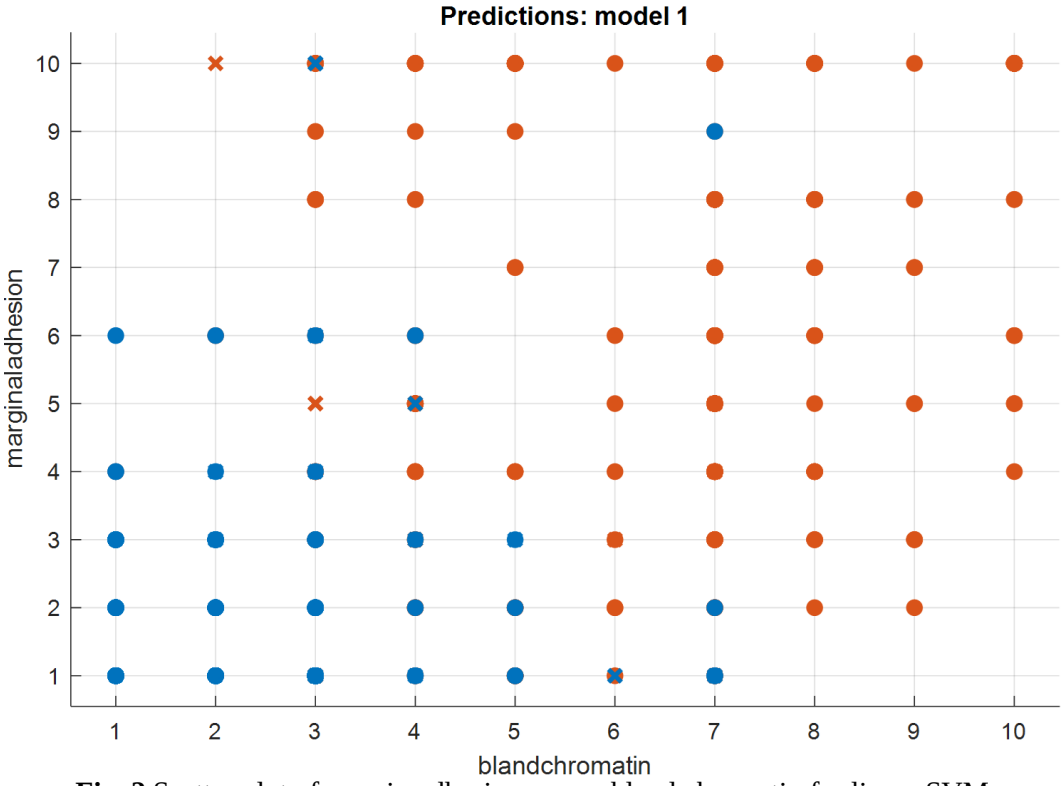

Fig. 2 Scatter plot of margin adhesion versus bland chromatin for linear SVM

4- Discussion

Extensive research has been conducted on the application of ML for breast cancer diagnosis and prognosis. SVM, a widely utilized ML technique, operates by establishing the optimal decision boundary with the greatest margin between two classes [8, 21]. The WBCD and breast cancer Coimbra datasets are the primary datasets employed in breast cancer research. This study focused on logistic regression and SVM models for predicting breast cancer malignancy. These ML models employed 10-fold cross-validation as their sampling strategy.

Logistic regression serves as a straightforward and efficient method for predicting cancer malignancy. The model's accuracy was assessed using a confusion matrix, while the

logistic regression model demonstrated an impressive AUC of 0.9947. The equation derived for the model is presented in **Eq. (1)**. Four crucial parameters—clump thickness, marginal adhesion, bare nuclei, and bland chromatin—proved to be significantly important for model generation based on their P-values. These parameters were also utilized in developing SVM model. In a related study, Hernández-Julio et al. [22] identified five key WBCD features using clustered and pivot table ML algorithms: uniformity of cell size, marginal adhesion, single epithelial cell size, bare nuclei, and normal nucleoli. Additionally, Elgedawy et al. [23] created a random forest model using the parameters: cell size, cell shape, clump thickness, and bare nuclei.

The most popular factor for comparing generated models is accuracy [24]. The formula for accuracy is mentioned in the result section. As shown in Tab. 4, it is evident that the linear SVM attained accuracy of 96.3%, respectively. Asri et al. [25] and Islam et al. [26] reported a 97% accuracy for the SVM method using 10-fold cross-validation. Additionally, Tarawneh et al. [27] achieved 97.90% accuracy with the decision tree model, utilizing the Kaggle archive as their dataset. They also evaluated the decision tree model's prediction using ROC, F-measure, recall, TP, and FP rates. The decision tree's performance surpassed that of other methods for the specified dataset. Furthermore, Afolayan et al. [28] achieved 92.26% accuracy with the decision tree model.

Comparing the AUC value provides another means of evaluating the performance of different models. The suggested models demonstrated significant predictability based on the AUC value. Both of the models exhibited an AUC value of 0.99. Notably, Zheng et al. [29] achieved an AUC value of 0.997 for their K-SVM model, while Bazazeh and Shubair [30] and Tarawneh et al. [27] attained the highest accuracy of 99.90% with the random forest method and decision tree method.

Overall, the results indicate that linear SVM exhibited higher accuracy and AUC values, respectively. However, our study has certain limitations. For instance, there is a lack of datasets for clinical validation, and the software we utilized may have constrained the results we obtained. In the future, we aim to bridge the gap and provide a more comprehensive comparison of various ML techniques by working with programming languages such as Python.

5- Conclusions

In this study, two ML algorithms were utilized to develop predictive models for breast cancer malignancy. Logistic regression and SVM were applied on the WBCD to achieve this objective. Logistic regression identified four statistically significant parameters: bland chromatin, bare nuclei, marginal adhesion, and clump thickness, forming the basis for further model development. SVM model demonstrated high accuracy of 96.8%, respectively and area under the ROC curve (0.99). Additionally, the derived logistic regression-based equation proved to be effective for predicting cancer malignancy. Our research suggests that these two ML models exhibit satisfactory accuracy and can serve as practical tools for physicians.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021;71(3):209-49.
2. Bakker MF, de Lange SV, Pijnappel RM, Mann RM, Peeters PH, Monninkhof EM, et al. Supplemental MRI screening for women with extremely dense breast tissue. New England Journal of Medicine. 2019;381(22):2091-102.
3. Veenhuizen SG, de Lange SV, Bakker MF, Pijnappel RM, Mann RM, Monninkhof EM, et al. Supplemental breast MRI for women with extremely dense breasts: results of the second screening round of the DENSE trial. Radiology. 2021;299(2):278-86.
4. Organization WH. WHO position paper on mammography screening: World Health Organization; 2014.
5. Sun Y-S, Zhao Z, Yang Z-N, Xu F, Lu H-J, Zhu Z-Y, et al. Risk factors and preventions of breast cancer. International journal of biological sciences. 2017;13(11):1387.
6. Kornecki A. Current status of contrast enhanced mammography: A comprehensive review. Canadian Association of Radiologists Journal. 2022;73(1):141-56.
7. Puliti D, Duffy SW, Miccinesi G, De Koning H, Lynge E, Zappa M, Paci E. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. Journal of medical screening. 2012;19(1_suppl):42-56.
8. Yue W, Wang Z, Chen H, Payne A, Liu X. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018;2(2):13.
9. Chitre Y, Dhawan AP, Moskowitz M. Artificial neural network based classification of mammographic microcalcifications using image structure features. International Journal of Pattern Recognition and Artificial Intelligence. 1993;7(06):1377-401.
10. Street WN, Mangasarian OL, Wolberg WH. An inductive learning approach to prognostic prediction. Machine Learning Proceedings 1995: Elsevier; 1995. p. 522-30.
11. Wolberg WH, Street WN, Heisey DM, Mangasarian OL. Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. Archives of Surgery. 1995;130(5):511-6.
12. Yassin NI, Omran S, El Houby EM, Allam H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Computer methods and programs in biomedicine. 2018;156:25-45.
13. Mokoatle M, Marivate V, Mapiye D, Bornman R, Hayes VM. A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application. BMC bioinformatics. 2023;24(1):112.
14. Zhang B, Shi H, Wang H. Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. J Multidiscip Healthc. 2023;16:1779-91.

15. Hammad A, Elshaer M, Tang X. Identification of potential biomarkers with colorectal cancer based on bioinformatics analysis and machine learning. Mathematical Biosciences and Engineering. 2021;18(6):8997-9015.
16. Ismael SAA, Mohammed A, Hefny H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artificial intelligence in medicine. 2020;102:101779.
17. Kennion O, Maitland S, Brady R. Machine learning as a new horizon for colorectal cancer risk prediction? A systematic review. Health Sciences Review. 2022;4:100041.
18. Ting W-C, Lu Y-CA, Ho W-C, Cheewakriangkrai C, Chang H-R, Lin C-L. Machine learning in prediction of second primary cancer and recurrence in colorectal cancer. International Journal of Medical Sciences. 2020;17(3):280.
19. Urbanos G, Martín A, Vázquez G, Villanueva M, Villa M, Jimenez-Roldan L, et al. Supervised Machine Learning Methods and Hyperspectral Imaging Techniques Jointly Applied for Brain Cancer Classification. Sensors. 2021;21(11):3827.
20. Wolberg. DWH. Breast Cancer Wisconsin (Original) Data Set.
21. Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science. 2021;191:487-92.
22. Hernández-Julio YF, Prieto-Guevara MJ, Nieto-Bernal W, Meriño-Fuentes I, Guerrero-Avendaño A. Framework for the development of data-driven Mamdani-type fuzzy clinical decision support systems. Diagnostics. 2019;9(2):52.
23. Elgedawy MN. Prediction of breast cancer using random forest, support vector machines and naïve Bayes. International Journal of Engineering and Computer Science. 2017;6(1):19884-9.
24. Salod Z, Singh Y. A five-year (2015 to 2019) analysis of studies focused on breast cancer prediction using machine learning: A systematic review and bibliometric analysis. Journal of Public Health Research. 2020;9(1):jphr. 2020.1772.
25. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016;83:1064-9.
26. Islam MM, Poly TN. Machine learning models of breast cancer risk prediction. BioRxiv. 2019:723304.
27. Tarawneh O, Otair M, Husni M, Abuaddous HY, Tarawneh M, Almomani MA. Breast cancer classification using decision tree algorithms. International Journal of Advanced Computer Science and Applications. 2022;13(4).
28. Afolayan JO, Adebiyi MO, Arowolo MO, Chakraborty C, Adebiyi AA. Breast cancer detection using particle swarm optimization and decision tree machine learning technique. Intelligent Healthcare: Infrastructure, Algorithms and Management: Springer; 2022. p. 61-83.
29. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications. 2014;41(4):1476-82.

30. Bazazeh D, Shubair R, editors. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 2016 5th international conference on electronic devices, systems and applications (ICEDSA); 2016: IEEE.