e-ISSN: 0975-1556, p-ISSN:2820-2643

Available online on www.ijpcr.com

International Journal of Pharmaceutical and Clinical Research 2022; 14(10); 812-818

Original Research Article

Feature Optimization and Breast Cancer Classification using Machine Learning Algorithms

Amit Tak[1], Puran Mal Parihar[2], Shikha Mathur[3], Pranshu Sharma[4]

[1]Assistant Professor, Department of Physiology, RVRS Medical College, Bhilwara, Rajasthan, India

[2]Associate Professor, Department of Pathology, Geetanjali Medical College, Udaipur, India

[3]Associate Professor, Department of Physiology, Mahatma Gandhi Medical College and Hospital, Jaipur, Rajasthan, India

[4]Assistant Professor, Department of Pathology, Ananta Institute of Medical Sciences and Research Centre, Rajsamand, Rajasthan, India

Received: 25-08-2022 / Revised: 25-09-2022 / Accepted: 30-10-2022

Corresponding author: Dr. Pranshu Sharma

Conflict of interest: Nil

Abstract

Background: Breast cancer is the world's most prevalent cancer in females. Statistical models and machine learning (ML) algorithms have been proposed to predict breast cancer. The present study used ML classifiers to classify breast tumors into 'benign' and 'malignant.'

Materials and Methods: The study dataset consists of a random sample of medical records of 569 breast cancer patients. The dataset is publicly available on the Machine Learning Repository website of the University of California Irvine (UCI ML). Thirty features are extracted from a digitized image of cell nuclei obtained from fine needle aspiration cytology of breast tumor. The response variables were 'benign' or 'malignant.' Predictors differed significantly in biostatistical tests and contributed significantly to the logistic regression model were used to train ML classifiers on MATLAB classifier application. The performance metrics of the machine learning classifier were expressed as accuracy, the area under the receiver operator characteristic (AU-ROC) curve, sensitivity, and specificity.

Results: The predictors that contributed significantly to the logistic model include perimeter worst, smoothness worst, texture worst, radius se, symmetry worst, compactness se, and concavity mean. These predictors were used to train various Machine learning classifiers. The logistic regression model showed the best performance. The accuracy, AU-ROC, sensitivity, and specificity were 97.2%, 99%, 98%, and 96%, respectively.

Conclusion: There was a striking improvement in the accuracy of classification of breast cancer achieved with ML algorithms compared to the state-of-the-art model-based approaches.

Keywords: benign, breast cancer, classification, malignant, machine learning, prediction

This is an Open Access article that uses a fund-ing model which does not charge readers or their institutions for access and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0) and the Budapest Open Access Initiative (http://www.budapestopenaccessinitiative.org/read), which permit unrestricted use, distribution, and reproduction in any medium, provided original work is properly credited.

Introduction

In 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. [1] Breast cancer arises in the lining cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. Initially, the cancerous growth is confined to the duct or lobule ("in situ"), where it generally causes no symptoms and has minimal potential for spread (metastasis). Over time, these in situ (stage 0) cancers may progress and invade the surrounding breast tissue (invasive breast cancer) and then spread to the nearby lymph nodes (regional metastasis) or other organs in the body (distant metastasis). If a woman dies from breast cancer, it is because of widespread metastasis. [2] Different techniques are used to capture breast cancer, such as Ultrasound Sonography (ULS), Computerized Thermography (CT), Fine Needle Aspiration Cytology, Biopsy (Histological images), Magnetic-Resonance-Imaging (MRI), and Digital Mammography breast X-ray images (DMG). Various statistical and machine learning models were used to predict disease models using digitized image datasets. [3,4,5,6,7,8,9,10] The present study use machine learning methods for breast cancer classification using digitized image data of cell nuclei obtained from fine needle aspirate of breast lesions.

Materials and Methods

The present study classifies breast cancer mass into 'benign' and 'malignant' using machine learning (ML) classifiers. The study dataset consists of a random sample of medical records of 569 breast cancer patients. The dataset is publicly available on the Machine Learning Repository website of the University of California Irvine (UCI ML). [11] Features are computed from a digitized image of a breast mass's fine needle aspirate (FNA). They describe the characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus, including radius (mean of distances from the 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, and fractal dimension ("coastline approximation" - 1). The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. The response variables were 'benign'(N=357) and 'malignant' (N= 212). The features that differed significantly between the two response variables were found using biostatistical tests. The features that differed significantly between the two groups were chosen to fit the Logistic regression model using the stepwise method. The features that contributed significantly to the logistic regression model were used for ML training and classification. The ML classifier application on MATLAB 2019a was used for classification with 5-fold cross-validation.

The classifiers used in this application include Decision Trees, Discriminant analysis (linear and quadratic), Support Vector Machine (SVM), Logistic regression, Naïve Bayes (Gaussian and Kernel), K-Nearest Neighbors (KNN), and ensemble learning classifiers. The decision trees include complex, medium, and simple tree classifiers. Similarly, the SVMs include linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian classifiers. The ensemble classifiers have boosted trees, bagged trees, and RUS boosted tree classifiers.

Statistical analysis

The quantitative data were expressed in median (IQR) and compared using the non-parametric Mann-Whitney's U test. The logistic regression model with minimum Akaike Information Criterion was selected. The performance metrics of the machine learning classifiers were expressed as accuracy, the area under the receiver operator characteristic (AU-ROC) curve, sensitivity, and specificity. JASP version 0.16.2 was used for statistical analysis. MATLAB Classification Learner application 2019a was used for training and prediction. [12,13] The significance level was considered at 5%.

Results

All the predictors, except fractal dimension, mean, texture se, and smoothness se, varied significantly between the 'benign' and 'malignant' breast cancer groups. (Table 1) A stepwise logistic regression model was fitted using varying predictors.

Table 1: Comparison of features in 'benign' and 'malignant' class of breast cancer patients.

Features	W	p
radius_mean	4729	< .001
texture_mean	16966.5	< .001
perimeter_mean	4019	< .001
area_mean	4668.5	< .001
smoothness_mean	21037	< .001
compactness_mean	10309.5	< .001
concavity_mean	4705.5	< .001
concave points_mean	2691.5	< .001
symmetry_mean	22814	< .001
fractal_dimension_mean	39012.5	0.537
radius_se	9965	< .001
texture_se	36964.5	0.644
perimeter_se	9355	< .001
area_se	5569.5	< .001
smoothness_se	40200.5	0.214
compactness_se	20640.5	< .001
concavity_se	16588.5	< .001
concave points_se	15758	< .001
symmetry_se	42013	0.028
fractal_dimension_se	28737	< .001
radius_worst	2237	< .001
texture_worst	16300	< .001
perimeter_worst	1858	< .001
area_worst	2283.5	< .001
smoothness_worst	18614	< .001
compactness_worst	10421.5	< .001
concavity_worst	5951.5	< .001
concave points_worst	2520	< .001
symmetry_worst	19909.5	< .001
fractal_dimension_worst	23767	< .001

The predictors that contributed significantly to the logistic model include perimeter worst, smoothness worst, texture worst, radius se, symmetry worst, compactness se, and concavity mean. (Table 2)

Table 2: Estimated parameters of the Logistic regression model using predictors with response variable as 'malignant' breast cancer.

Parameter	Estimate	Standard Error	z	Wald Test		
				Wald Statistic	df	p
(Intercept)	-56.4	9.655	-5.842	34.124	1	< .001
perimeter_worst	0.243	0.043	5.618	31.567	1	< .001
smoothness_worst	83.224	23.659	3.518	12.374	1	< .001
texture_worst	0.368	0.076	4.834	23.363	1	< .001
radius_se	12.768	3.174	4.023	16.185	1	< .001
symmetry_worst	16.111	6.31	2.553	6.519	1	0.011
compactness_se	-100.157	28.241	-3.546	12.577	1	< .001
concavity_mean	33.114	10.325	3.207	10.286	1	0.001

df: degrees of freedom; p: p-value

These predictors were used to train various Machine learning classifiers using the Classification learner app on MATLAB. The logistic regression model showed the best performance. The accuracy, AU-ROC, sensitivity, and specificity were 97.2%, 99%, 98%, and 96%, respectively. (Table 3 and Figure 1)

Table 3: Performance metrics of the logistic regression model used for classifying breast cancer lesions into 'benign' and 'malignant.'

Performance metrics	Model: Logistic Regression
Accuracy	97.2%
AUC	0.99
Sensitivity	98%
Specificity	96%

Discussion

According to the World Health Organization (WHO), the number of females that die in 2020 is about 685,000. The number may reach 2.7 million in 2030 globally. [14] The delay in diagnosis is the main reason for the low survival rate. Therefore, early detection and treatment of breast cancer are vital to disease management. There are two types of breast cancer, invasive and non-invasive. The former's prognosis is poor, difficult to treat, and thus malignant. The latter is non-invasive, not spread to other organs, and manageable, therefore benign. The present study used digitized images of cell nuclei obtained from Fine Needle Aspirate. Various features were extracted from these images, and other features were selected. The selected features were used for machine learning classification of breast cancer.

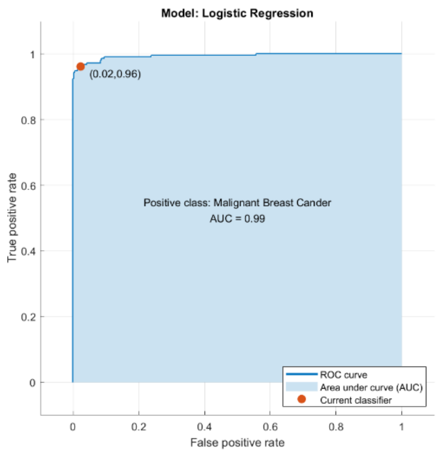

Figure 1: The area under the receiver operator characteristic curve for the logistic regression model.

The present study used various machine learning classifiers, including decision trees, discriminant analysis support vector machine, logistic regression, Naïve Bayes, K-nearest neighbors, and ensemble learning classifiers to classify breast cancer. The logistic regression showed the highest accuracy. Wu et al. proposed using a machine learning (ML) approach to classify triple-negative breast cancer

and non-triple negative breast cancer patients using gene expression data. RNA-Sequence data from 110 triple negative and 992 non-triple negative breast cancer tumor samples were used to train four classification models, including Support Vector Machines, K-nearest neighbor, Naïve Bayes, and Decision tree. The Support Vector Machine algorithm could classify breast cancer more accurately among the four ML algorithms evaluated. [15] Amrane et al. classify breast cancer into benign and malignant. Researchers used the Naive Bayes (NB) classifier and K-Nearest Neighbor (KNN) for breast cancer classification. KNN gives the highest accuracy (97.51%) with the lowest error rate than the NB classifier (96.19 %). [16] Omandiagbe et al. classified breast cancer with Support Vector Machine (using radial basis kernel), Artificial Neural Networks, and Naïve Bayes using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset. The authors used a hybrid approach for breast cancer diagnosis by reducing the high dimensionality of features using linear discriminant analysis (LDA) and then applying the new reduced feature dataset to the Support Vector Machine. The proposed approach obtained an accuracy of 98.82%, a sensitivity of 98.41%, a specificity of 99.07%, and an area under the receiver operating characteristic curve of 0.9994. [17] Aljuiad et al. presented a novel computer-aided diagnosis method for breast cancer classification (both binary and multi-class), using a combination of deep neural networks (ResNet 18, ShuffleNet, and Inception-V3Net) and transfer learning on the BreakHis publicly available dataset. The best average accuracy for binary classification of benign or malignant cancer cases of 99.7%, 97.66%, and 96.94% for ResNet, InceptionV3Net, and ShuffleNet, respectively. Average accuracies for multi-class classification were 97.81%, 96.07%, and 95.79% for ResNet, Inception-V3Net, and ShuffleNet, respectively. [18] Naji et al. applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic dataset. It is observed that the Support vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%). [19] Jabbar et al. developed a decision support system using the ensemble model built with Bayesian network and Radial Basis Function using "Wisconsin Breast Cancer Data set (WBCD)." The model showed an accuracy of 97% in classifying breast cancer data. [20] Nawaz et al. showed a deep learning approach based on a Convolutional Neural Network (CNN) model for multi-class breast cancer classification. Researchers classified breast tumors as benign or malignant but predicted subclass of cancer like fibroadenoma, lobular carcinoma, etc. Experimental results on histopathological images using the BreakHis dataset show that the DenseNet CNN model achieved high processing performances with 95.4% accuracy in the multi-class breast cancer classification task compared with state-of-the-art models. [21,22]

Conclusion

There was a striking improvement in the accuracy of classification of breast cancer achieved with ML algorithms compared to the state-of-the-art model-based approaches. High-accuracy prediction techniques are essential in personalized medicine because they facilitate stratification of prevention strategies and individualized clinical management.

Research Quality and Ethics Statement: Research Quality and Ethics Statement- The present study with report quality, formatting, and reproducibility guidelines set forth by the EQUATOR Network. The data used was acquired from the UCL repository, so exempted from an Institutional Review Board / Ethics Committee review.

References

1. Breast cancer [Internet]. [cited 2022 Sep 28]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. What Is Breast Cancer? | CDC [Internet]. [cited 2022 Sep 28]. Available from: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm
3. Ho JE, Enserro D, Brouwers FP, Kizer JR, Shah SJ, Psaty BM, et al. Predicting Heart Failure with Preserved and Reduced Ejection Fraction: The International Collaboration on Heart Failure Inftypes. Circ Hear Fail [Internet]. 2016 Jun 1 [cited 2022 Aug 8];9(6).
4. Bhandari S, Tak A, Singhal S, Shukla J, Shaktawat AS, Gupta J, et al. Patient Flow Dynamics in Hospital Systems During Times of COVID-19: Cox Proportional Hazard Regression Analysis. Front Public Heal. 2020 Dec 8;8:820.
5. Bhandari S, Shaktawat AS, Tak A, Patel B, Shukla J, Singhal S, et al. Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. Ibnosina J Med Biomed Sci [Internet]. 2020 [cited 2022 Mar 26];12(2):123.
6. Bhandari S, Singh Shaktawat A, Tak A, Patel B, Gupta J, Gupta K, et al. Independent Role of CT Chest Scan in COVID-19 Prognosis: Evidence From the Machine Learning Classification (1) (2) (3) (4) (5) (6) (7). Scr Med. 202 1;52(4):273–81.
7. Tak A, Dia S, Dia M, Wehner TC. Indian COVID-19 Dynamics: Prediction Using Autoregressive Integrated Moving Average Modelling ARTICLE INFO (1) (2). Scr Med [Internet]. 2021 [cited 2022 Apr 13];52(1):6–14.
8. Tak A, Punjabi P, Yadav A, Ankhla M, Mathur S, Dave HS, et al. Prediction of Type 2 Diabetes Mellitus Using Soft Computing. Mod Med [Internet]. 2022 Jun 22 [cited 2022 Aug 12];29(2):135–43. Available from: https://medicinamoderna.ro/prediction-of-type-2-diabetes-mellitus-using-soft-computing/
9. Darshan Shah K, Pancharia A, Bamaniya H, Sharma A, Somani S, Tak A. Evaluation of Risk Factors for Ten-Year Coronary Heart Disease using Logistic Regression Modeling International Journal of Pharmaceutical and Clinical Research. Int J Pharm Clin Res [Internet]. 2022 [cited 2022 Aug 12];14(6):764–71.
10. Tak A, Parihar PM, Singh Fatehpuriya D, Singh Y. Optimised Feature Selection and Cervical Cancer Prediction Using Machine Learning Classification. Scr Med. 2022; 53(3): 205–16.
11. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set [Internet]. [cited 2022 Sep 28]. Available from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
12. JASP Team. JASP (Version 0.16.2.0) [Computer software] [Internet] . 2022. Available from: https: //jasp-stats.org/
13. MATLAB - MathWorks - MATLAB & Simulink [Internet]. [cited 2022 Apr 3]. Available from: https://in.mathworks.com/products/matlab.html
14. Breast cancer [Internet]. [cited 2022 Oct 3]. Available from: https:// www.who.int/ news-room/fact-sheets/ detail/breast-cancer
15. Wu J, Hicks C. Breast Cancer Type Classification Using Machine Learning. J Pers Med [Internet]. 2021 Feb 1 [cited 2022 Sep 24];11(2):1–12.
16. Amrane M, Oukid S, Gagaoua I, Ensari T. Breast cancer classification using machine learning. 2018 Electr Electron Comput Sci Biomed Eng

Meet EBBT 2018. 2018 Jun 20;1–4.

17. Omondiagbe DA, Veeramani S, Sidhu AS. Machine Learning Classification Techniques for Breast Cancer Diagnosis. IOP Conf Ser Mater Sci Eng. 2019;495(1).
18. Aljuaid H, Alturki N, Alsubaie N, Cavallaro L, Liotta A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. Comput Methods Programs Biomed. 2022 Aug 1;223:106951.
19. Naji MA, Filali S El, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Comput Sci. 2021 Jan 1;191:487–92.
20. Jabbar MA. Breast Cancer Data Classification Using Ensemble Machine Learning. Eng Appl Sci Res [Internet]. 2021 Jan 27 [cited 2022 Sep 24];48(1):65–72. Available from: https ://ph01.tci-thaijo.org/index.php/easr/article/view/234959
21. Nawaz M, Sewissy AA, Soliman THA. Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network. Int J Adv Comput Sci Appl. 2018 [cited 2022 Sep 24];9(6):316–22.
22. IJ, O., J, O. J., & U, O. B. Evaluation of the Effectiveness of Intra-operative Low Dose Ketamine Infusion on Post-operative Pain Management Following Major Abdominal Gynaecological Surgeries. Journal of Medical Research and Health Sciences, 2022; 5(10): 2269–2277.