ORIGINAL ARTICLE

Evaluation of the Role of Data Increment of Cancer Cases with Computer-Aided Algorithms for Detection of Breast Cancer

Imran Majeed Khan[1*], Hafiz Rafique[2], Abdul Waheed Anwar[3], Basit Attique[2] and Muhammad Jahanzab[2]

[1]*Allama Iqbal Medical College/Jinnah Hospital, Lahore, Pakistan*
[2]*Department of Physics, Punjab University, Lahore, Pakistan*
[3]*Department of Physics, University of Engineering and Technology, Lahore, Pakistan*

Abstract

Background: Cancer is one of the leading causes of death and morbidity all over the world, with 14.1 million new cases and 8.2 million deaths due to cancer.

Objective: Early breast cancer detection is important for the treatment and survival of patients. CAD is a useful tool for earlier cancer detection.

Methods: There are 1863 malignant and benign cases. The nine features are extracted from the DDMS database, and the values are assigned using the BI-RAD mammography lexicon. The research is conducted at Radiation Oncology, AIMC/Jinnah Hospital, Lahore from October 2021 to November 2023. Mammography is an important medical imaging modality used for early diagnosis and detection of breast diseases. The data size plays an important role in applying artificial intelligence to cancer diagnosis. CBR stands for Case-based Reasoning is an established research in the Artificial Intelligence field. The CBR was used at multiple data increments to research its impact on the detection of breast cancer. Principal component Analysis (PCA) is applied to evaluate important features in mammograms to improve the precision and recall for the detection of breast cancer.

Results: The recall of malignant test cases lies in the range of 0.78 to 0.88. The precision and recall for benign test cases vary between 0.82 to 0.89 and 0.85 to 1 respectively.

Conclusion: Finally, the implementation of PCA on data results showed that the precision of malignant test cases increased, and recall decreased. The data increment proves to increase the detection of breast cancer.

Keywords: *Cancer, case-based reasoning, precision, recall, principal component analysis, recall.*

INTRODUCTION

Cancer is one of the leading causes of death and morbidity all over the world, with 19 292 789 new cases and 8.2 million deaths due to cancer. There are 32.6 million people who are living with Cancer during five years of detection. The world has reported 1590000 deaths due to lung cancer, 745000 deaths due to liver, 723000 cancer death due to stomach, 694000 cancer death due to colorectal, 521000 cancer death due to breast and 400000 cancer death due to esophageal [1]. The most five common cancers spotted in men were lung, prostate, colorectal, stomach, and liver cancer, and in women breast, colorectal, lung, cervix, and stomach cancer. It is estimated that yearly cancer cases will increase from 14 million to 22 in the next 2 decades [2].

Early breast cancer detection is important for the treatment and survival of patients. CAD is a useful tool for earlier cancer detection. This section comprises a review of the performance of CAD implementation for breast cancer detection for different medical imaging modalities like mammography, MRI, and ultrasound [3, 4].

**Corresponding author: Imran Majeed Khan, Allama Iqbal Medical College/ Jinnah Hospital, Lahore, Pakistan, Email: bisimran@gmail.com*
Received: June 12, 2024; Revised: October 10, 2024; Accepted: November 19, 2024
DOI: https://doi.org/10.37184/lnjcc.2789-0112.5.21

A mammogram is the most consistent and effective way for breast cancer diagnoses in the early stage. Computer use is important to help radiologists in mammography because of complicated breast architecture, low breast cancer probability, and subtleties present among findings. CAD systems can be implemented for both FFDM and Screen-Film mammography [5].

Joshua J. Fenton *et al.* [6] showed the effectiveness of the CAD system implemented from 1998 to 2006 on screen-film mammograms on 684 956 women who had received above 1.6 million screen-film mammograms at the Breast Cancer Surveillance Consortium. CAD system was applied to 27.8% of screen-film mammograms and results showed decreased specificity (0.5%) and no improved detection rate for invasive breast cancer.

Arifa Sadaf *et al.* [7] evaluated the performance of CAD with FFDM in the detection of breast cancers applied to 127 mammographic cases that proved breast cancers with biopsy-diagnosed with FFDM. CAD system mounted to FFDM revealed 100% sensitivity in finding microcalcifications and 86% sensitivity for other cancers. The difference in sensitivity is mainly due to lesion size. They concluded that the CAD system with FFDM helped assist the radiologist in early breast cancer detection.

  ISSN: 2789-0120 (Online)

Robert M. Nishikawa *et al.* [8] showed that radiologist sensitivity increased by 10% and a comparable increase in recall rate by the use of CAD system on mammograms above 256 cases. CAD system successfully recognized 71% of cancer cases that were missed by radiologists at screening.

Mohamed Meselhy Eltoukhy *et al.* [9] proposed a statistical t-test method for feature extraction and applied breast cancer detection and classification in mammograms. They used a Support Vector Machine (SVM) to classify (5-fold) by using 70% of the dataset and 30% was used for the calculation of the classification rate. The accuracy rate by the proposed method is 95.84% to classify normal and abnormal tissues and 96.56% to detect benign and malignant tumors using wavelet coefficients. The accuracy rate by the proposed method is 95.98% to classify normal and abnormal tissues and 97.30% to detect benign and malignant tumors using curvelet coefficients.

J. Dheeba *et al.* [10] investigated a new classification method for breast cancer detection by use of a Particle Swarm Optimized Wavelet Neural Network on 216 digital mammograms based on extracting Laws Texture Energy Measures by application of a pattern classifier. The sensitivity and specificity of the new classification method were 94.167% and 92.105% respectively.

Yu-Dong Zhang *et al.* [11] proposed a novel CAD system for detecting abnormalities in breasts on 200 mammogram images. The sensitivity, specificity, and accuracy of their proposed method based on Weighted Type fractional Fourier Transform with Principal Component Analysis in addition to SVM were and respectively.

Zhiqiong Wang *et al.* [12] proposed a CAD detection system established on an extreme learning machine by the implementation of optimum fused features for breast cancer detection. They confirmed the effectiveness of their proposed method on 222 mammograms [13].

The main objectives of this study were to research the impact of data increment on the early detection of breast cancer by the application of computerized algorithms on mammogram databases.

METHODOLOGY

This study was carried out at Jinnah Hospital, Lahore, and was approved by the hospital's ethical committee. The guidelines of the Helsinki Declaration were followed in conducting this research work. Participants. The data included for which consent was available.

The research was conducted during the 2 years 2021-2023. Those mammographic images of women who met the following criteria were included in the research: diagnosed cases of biopsy-proven breast cancer, benign cases, and complete records were available for them.

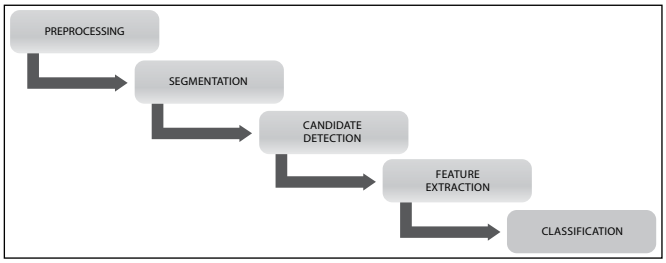

Fig. (1): CAD process.

Total population sampling allowed deep insight into the factors involved in the delay in the treatment and diagnosis of cancer. This is a method through which we included all the patients fulfilling our criteria and excluded those who did not meet the criteria. The steps that are used by CAD for cancer detection are shown in **Fig. (1).** First of all, take the medical image from imaging techniques and after that, the following steps are performed pre-processing, segmentation, candidate detection, feature extraction, and classification by CAD. Radiologists made the final decision about the cancerous area [14].

There are 1863 malignant and benign cases. The eight features are extracted from the DDMS database, and the values are assigned using the BI-RAD mammography lexicon. The following mammogram features are extracted from the data (**Table 1**):

a) Mass Shape (Oval, Round, Irregular)
b) Mass Margin (Circumscribed, Obscured, Microlobulated, Indistinct, Spiculated)
c) Mass Density (High density, Equal density, Low density, Fat-containing)
d) Calcifications Number (Skin, Vascular, Coarse or "Popcorn-Like", Large Rod-Like, Round Rim, Dystrophic, Milk of Calcium, Suture, Suspicious)
e) Calcifications Morphology (Amorphous, Coarse Heterogeneous, Fine Pleomorphic, Fine Linear or Fine-Linear Branching)
f) Calcifications Distribution (Diffuse, Regional, Grouped, Linear, Segmental, Architectural Distortion)
g) Associated Findings
h) Special cases

The features extracted from mammograms are provided by numerical weightage. When the features are converted into numerical values, the values are normalized into the range of (0-1) for the application of algorithms. The CBR system is derived into different systems depending upon the number of cases in the case base named 100--1,300--1,500--1, 700--1, and, 900--1.

Mammography is an important medical imaging modality used for early diagnosis and detection of breast diseases. Mammography is a low-dose X-ray of the breast. X-rays are frequently used for imaging a body part [15]. The image of mammography is called a mammogram. The mammogram of six abnormal breasts is shown in **Fig. (2).**

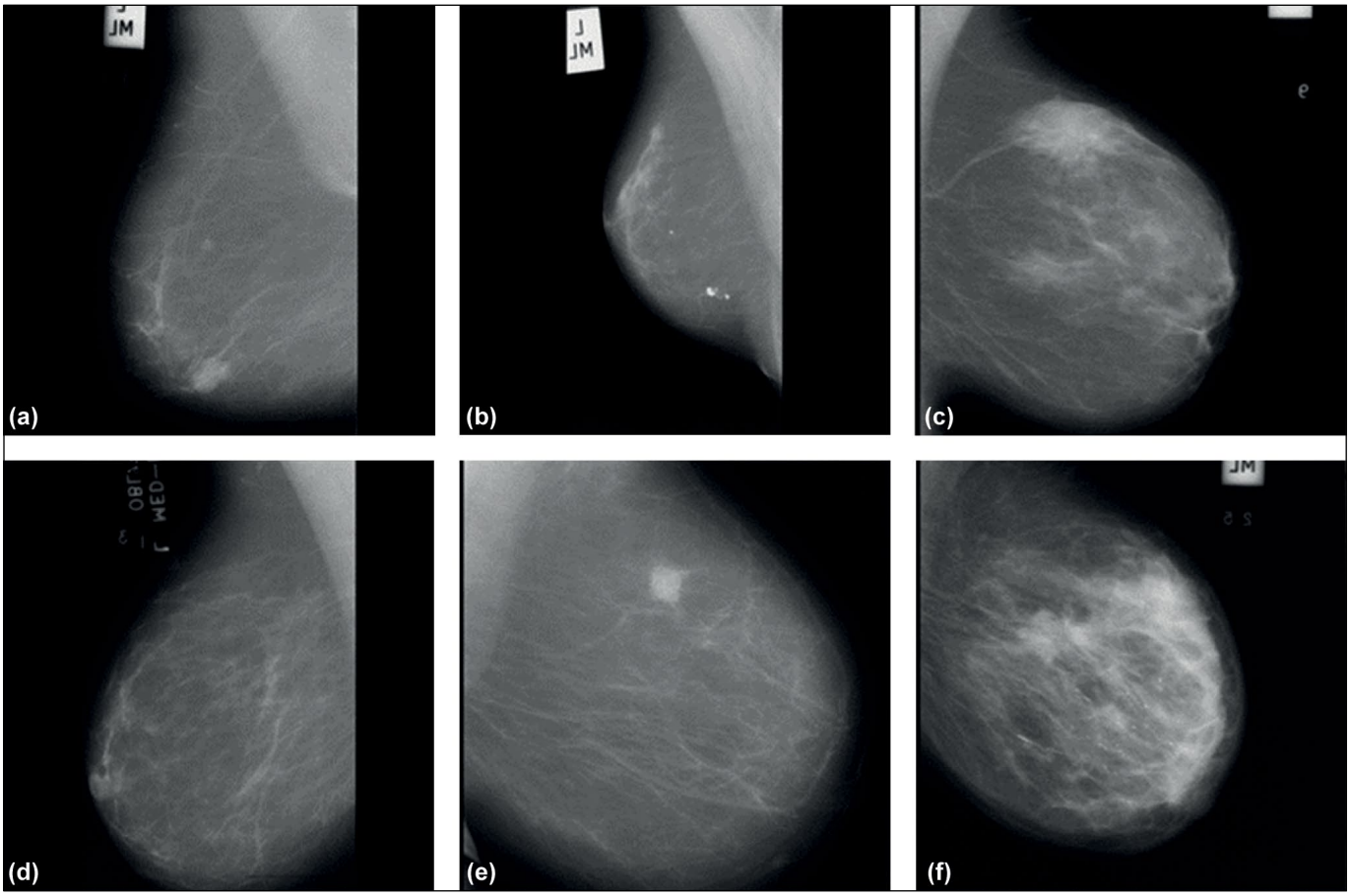

Fig (2): Image of six abnormal types: (a) circumscribed mass, (b) asymmetry, (c) architectural distortion, (d) calcification, (e) ill-defined masses, and (f) spiculated masses [10].

Table 1: Mammography lexicon.

Mass Shape(MS)	Oval Round Irregular
Mass Margin(MM)	Circumscribed Obscured Microlobulated Indistinct Spiculated
Mass Density(MD)	High density Equal density Low density Fat-containing
Calcifications Number (CN)	Skin Vascular Coarse or “popcorn-like” Large rod-like Round Rim Dystrophic Milk of calcium Suture Suspicious
Calcifications Morphology(CM)	Amorphous Coarse heterogeneous Fine pleomorphic Fine linear or fine-linear branching
Calcifications Distribution(CD)	Diffuse Regional Grouped Linear Segmental Architectural distortion
Associate findings(AS)	
Special Cases(SC)	
Age(A)	

Three present advanced mammographies are digital mammography, CAD, and 3D-mamography. In digital mammography, also named FFDM, x-ray film is changed by an electronic system that changes X-rays into mammographic breast pictures that create better images even with a low dose of radiation. The electronic systems used in X-rays are similar to the digital cameras’ electronic systems. Breast Imaging Reporting and Data System (BI-RADS) was established by the American College of Radiology (ACR) to standardize mammographic reporting, to improve communication [16].

CBR

CBR stands for Case-based Reasoning is an established research in the Artificial Intelligence field. It is the study of designing the system on theoretical foundations and its practical application to solve the problem through experience. The essential of every CBR is the case base, which consists of formerly prepared and stored experience, known as cases. A case-based solver resolves new problems with the help of similar past solved problems present in the case base [17]. CBR system selects one or several similar cases. The solutions of selected similar cases are adapted to develop a current problem solution. Finally, a new

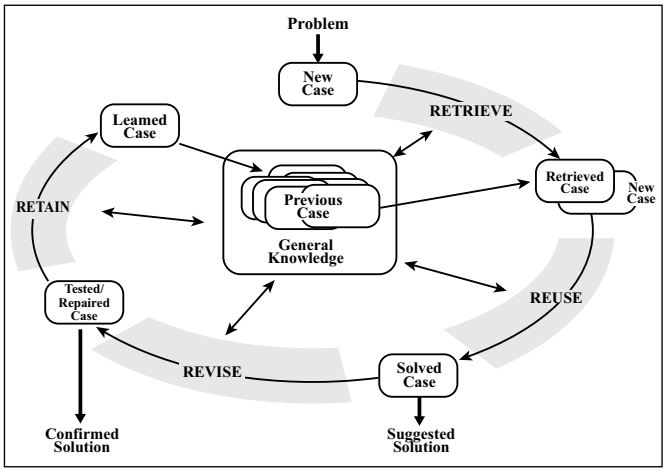

Fig. (3): CBR cycle.

solution to the new problem is stored in the case base to increase its capability.

CBR can be divided into three types depending on their case representation and reasoning technique name as textual CBR, structural CBR, and conversational CBR [18]. In structural CBR the cases are characterized as according to common structured vocabulary (ontology). In textual CBR, the cases are characterized as free text (strings). In conversational CBR, cases are represented by the list of varied questions in cases. Despite different approaches to the CBR systems, the basics of all CBR are a simple and uniform process as shown in **Fig. (3).**

Principal Component Analysis

Principal Component Analysis (PCA) uses the principles of mathematics to convert numerous correlated variables into a smaller number of variables called principal components. PCA is used to analyze the multivariate data. Mathlab is used to perform this analysis. PCA reduces the dimensions of multivariate data using vector space transformation [19]. The data set is interpreted in a few principal components by using mathematical projections. So, it helps the user by lessening data dimension to find trends, outliers, and patterns in the data.

PCA is a dimensionality decrease technique that is much of the time used to diminish the dimensionality of enormous informational collections, by changing a huge arrangement of factors into a more modest one that contains the majority of the data in the huge set. PCA is straightforward: decrease the quantity of factors of an informational index, while safeguarding however much data as could reasonably be expected. Principal Components are new factors that are built as straight blends or combinations of the underlying factors.

Precision and Recall

A classifier predicts the result as positive or negative in the case of binary decision problems. The classier decision or result is represented in the table named a confusion matrix/contingency table. A confusion matrix is tabulated.

It consists of four classes TP (True positive), FP (False positives), TN (True negatives), and FN (false negatives) **(Fig. 4).** TP is correctly categorized as positive. FP corresponds to negative examples incorrectly

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Fig. (4): Confusion matrix.

categorized as positive. TN refers to negatives correctly categorized as negative. FN corresponds to positive examples incorrectly categorized as negative.

The precision and recall and be calculated by using the following relations in terms of TP, FP, FN, and, TN.

RESULTS

In the proposed system, multiple observations are made by making the test and training cases out of 1863 cases, the two different algorithms are used, and the threshold value 0.62. The test cases both malignant and benign are selected to check the performance of the proposed system calculating the precision (P1, P2) and recall (R1, R2). The average precision (P) and average recall (R) are also calculated and the results are summarized in Table **2** for malignant and Table **3** for benign.

The precision of malignant test cases varies between 0.85 and 1. The recall of malignant test cases lies in the range of 0.78 and 0.88. The precision and recall for benign test cases vary between 0.82-0.89 and 0.85-1 respectively.

The precision and recall of the proposed CAD system calculated for malignant and benign test cases are also shown as bar graphs in **Figs. (5-8)** respectively.

CAD Performance with PCA Implementation

The precision and recall of test cases implemented on about eighteen hundred case bases is shown in Table **4**

Table 2: Precision and recall of malignant test cases.

Malignant	**Percision**			**Recall**		
Cases	P1	P2	P	R1	R2	R
1-100	1	1	1	0.8	.75	0.78
1-300	0.84	0.94	0.89	0.8	0.8	0.8
1-500	0.85	0.89	0.87	0.85	0.85	0.85
1-700	0.81	0.89	0.85	0.85	0.85	0.85
1-900	1	1	1	0.0	0.85	0.88

Table 3: Result for forty test cases on 1863 case base.

Benign	**Percision**			**Recall**		
Cases	P1	P2	P	R1	R2	R
1-100	0.83	0.8	0.82	1	1	1
1-300	0.8	0.83	0.82	0.8	0.8	0.8
1-500	0.85	0.86	0.85	0.85	0.85	0.85
1-700	0.84	0.86	0.85	0.8	0.89	0.85
1-900	0.91	0.87	0.89	1	1	1

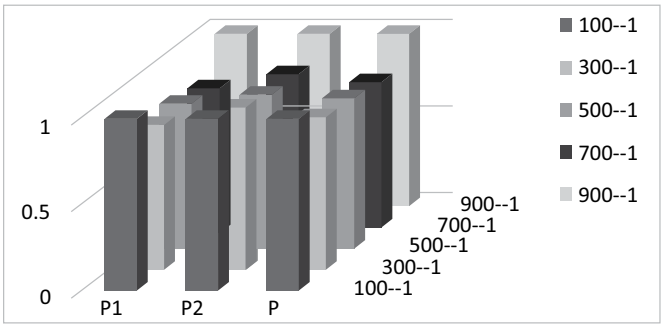

Fig. (5): Precision for malignant cases.

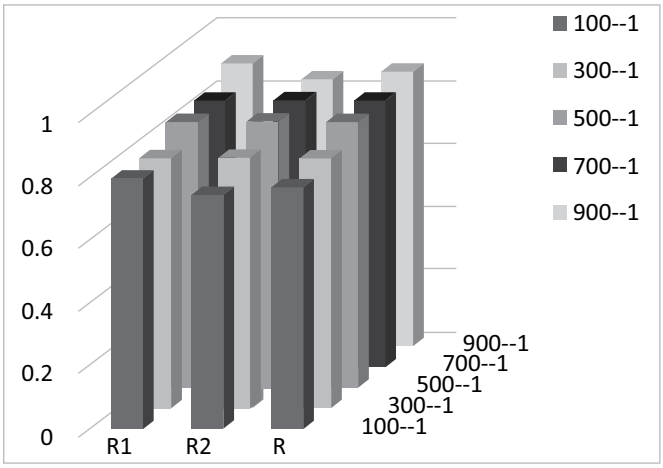

Fig. (6): Recall for malignant cases.

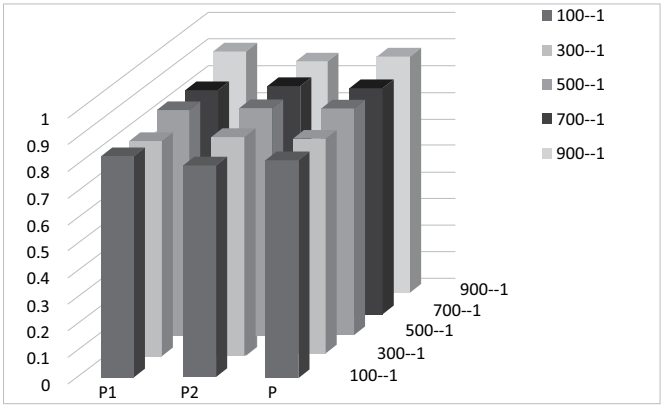

Fig. (7): Precision for benign cases.

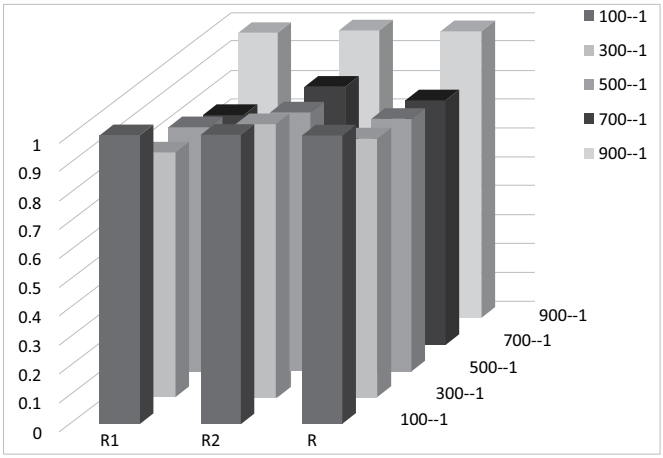

Fig. (8): Recall for benign cases.

Table 4: Result for forty test cases on 1863 case base after PCA implementation.

Malignant						
Precision				Recall		
Cases	P1	P2	P	R1	R2	R
1863	0.76	0.66	0.71	0.95	0.95	0.95
Benign						
Cases	P1	P2	P	R1	R2	R
1863	0.93	0.9	0.92	0.7	0.5	0.6

Table 5: Results of precision and recall of the test cases after the PCA implementation.

Malignant				
Precision			Recall	
Cases	P1	P2	P	R1
1863	0.74	0.81	0.77	0.85
Benign				
Cases	P1	P2	P	R1
1863	0.82	0.84	0.83	0.8

and the results of precision and recall of the test cases after the PCA implementation are summarized in Table **5**.

The result showed that the precision of malignant test cases increased and recall decreased. The precision for benign test cases decreased and recall increased.

DISCUSSION

Many Researchers presented different CAD systems based on different methods to detect various types of cancer on the visual information collected from medical images and proved that their CAD systems had clinical applications [20]. The accuracy of CAD systems implemented on breast mammograms varies between 92% and 97.3%. The accuracy of the CAD system mounted on breast MRI is in the range of 88.42% and 91.67%. The accuracy of the CAD system implemented on breast ultrasound varies between 90% and 100 %. The proposed CAD system sensitivity for lung cancer is between 90% with 0.05 false positives and 98.2% with 9.1 false positives. The accuracy of different developed CAD systems for brain tumors is over 99%. However, most of CAD systems are only used to detect particular cancer types on particular database. The researchers/ oncologists showed that combining different methods could improve the accuracy and effectiveness of CAD systems to detect cancer.

Joshua J. Fenton *et al.* [6] have implemented CAD on a large database of film screen mammograms but specificity was decreased in comparison to our research the specificity and sensitivity are increased.

Arifa Sadaf *et al.* [7] evaluated the performance of CAD with FFDM in the detection of breast cancers applied on 127 mammographic cases and revealed 100% sensitivity in finding microcalcifications and 86% sensitivity for other cancers. Our research was implemented on large data of 1863 cases and achieved better CAD performance.

Robert M. Nishikawa *et al.* [8] showed that radiologist sensitivity increased by 10% and authenticated our research that data size increment helps in diagnosis by CAD algorithms.

Mohamed Meselhy Eltoukhy *et al.* [9] proposed a statistical t-test method for feature extraction and applied breast cancer detection and classification in mammograms. They used a Support Vector Machine and the accuracy rate provided was 95.84% to classify normal and abnormal tissues in comparison we have shown better results by increasing the database.

J. Dheeba *et al.* [10] investigated a new classification method for breast cancer detection by use of a Particle Swarm Optimized Wavelet Neural Network on 216 digital mammograms revealing sensitivity and specificity of the method were 94.167% and 92.105% respectively. Our case-based reasoning approach shows better results and also represents the reasons behind their response.

Yu-Dong Zhang *et al.* [11] proposed an SVM and PCA CAD system for detecting abnormalities in breasts on 200 mammogram images. Our CBR and PCA-based models produced better results with logic using more data.

Zhiqiong Wang *et al.* [12] proposed a machine-learning CAD detection system for 222 mammograms. Our results show with a large database including more benign and malignant cases the accuracy can be improved.

CONCLUSION

The performance of the presented CBR-based CAD system was checked with the help of a confusion matrix by calculating the precision and recall. The precision of malignant test cases varies between 0.85 and 1. The recall of malignant test cases lies in the range of 0.78-0.88. The precision and recall for benign test cases vary between 0.82-0.89 and 0.85-1 respectively.

Finally, the implementation of PCA on data results showed that the precision of malignant test cases increased and recall decreased. However, the precision for benign test cases decreased, and recall increased.

ETHICS APPROVAL

Ethical approval was obtained from the Institutional Review Committee of AIMC/Jinnah Hospital, Lahore (Ref. letter No. 194/23/12/2021/52 ERB Dated: 17 February 22). The research was conducted in this work by following the ethical standards of the institutional and/ or national research committee and the Helsinki Declaration.

CONSENT FOR PUBLICATION

Written informed consent was taken from the participants.

AVAILABILITY OF DATA

The data set may be acquired from the corresponding author upon a reasonable request.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

AUTHORS' CONTRIBUTION

Dr. Imran Majeed contributed to the concept and research, Mr. Basit and Mr. Jahanzeb contributed to the research work, and Dr. H.M. Rafique and Dr. Abdul Waheed Anwar contributed to the research paper.

REFERENCES

1. International Agency for Research on Cancer. Agents classified by the IARC monographs, volumes 1-137. Lyon: IARC; Available from: https://monographs.iarc.who.int/agents-classified-by-the-iarc/
2. Majeed I, Ammanuallah R, Anwar A, Rafique H, Imran F. Diagnostic and treatment delays in breast cancer in association with multiple factors in Pakistan. East Mediterr Health J 2021; 27(1): 23-32. DOI: https://doi.org/10.26719/emhj.20.051
3. Ikram A, Pervez S, Khadim M H, Sohaib M, Uddin H, Badar F, *et al.* National Cancer Registry of Pakistan: First comprehensive report of cancer statistics 2015-2019. J Coll Physicians Surg Pak 2024; 34(12): 625-32. DOI: https://doi.org/10.29271/jcpsp.2023.06.625
4. Majeed I, Amanuallah R, Rafique H. Time delay barriers in diagnosis and treatment of cancer. World Cancer Res J 2018; 5(3): e118. DOI: https://doi.org/10.32113/wcrj_20189_1118
5. Majeed I, Imran F, Nadeem A, Ashiq R, Nasir B. Early detection of cancer using mammograms with advanced artificial intelligence (AI) algorithms for breast lesions. J Liaq Natl Hosp 2023; 1(2): 67-73. DOI: https://doi.org/10.37184/jlnh.2959-1805.1.15
6. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, *et al.* B.C S. Consortium. Effectiveness of computer-aided detection in community mammography practice. J Natl Can Inst 2011; 103: 1152-61. DOI: https://doi.org/10.1093/jnci/djr206
7. Sadaf A, Crystal P, Scaranelo A, Helbich T. Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers, Eur J Radiol 2011; 77: 457. DOI: https://doi.org/10.1016/j.ejrad.2009.08.024
8. Nishikawa RM, de Cea MVS, Yang Y. Locally adaptive decision in detection of clustered microcalcifications in mammograms. Phys Med Biol 2018; 63(4): 045014. DOI: https://doi.org/10.1088/1361-6560/aaaa4c
9. Eltoukhy MM, Faye I, Samir BB. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. Comput Biol Med 2012; 42: 123-8. DOI:10.1016/j.compbiomed.2011.10.016
10. Dheeba J, Shamy S. A research on detection and classification of breast cancer using k-means GMM & CNN algorithms. Int J Engin Adv Technol 2019; 8(6S): 501-5. DOI:10.35940/ijeat.F1102.0886S19.
11. Zhang Y-D, Wang S-H, Liu G, Yang J. Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. Adv Mech Engin 2016; 8(2): 1-11. DOI: https://doi.org/10.1177/1687814016634243
12. Górriz JM, Álvarez-Illán I, Álvarez-Marquina, Arco JE, Atzmueller M, Ballarini F, *et al.* Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. Inform Fusion 2023; 100: 101945. DOI: https://doi.org/10.1016/j.inffus.2023.101945

13. Rosenberg SM, Zheng Y, Ruddy K, Poorvu PD, Snow C, Kirkner GJ, *et al.* Helping ourselves, helping others: the Young Women's Breast Cancer Study (YWS) - a multisite prospective cohort study to advance the understanding of breast cancer diagnosed in women aged 40 years and younger. BMJ Open 2024; 14(6): e081157.
DOI: https://doi.org/10.1136/bmjopen-2023-081157

14. Rose C, Turi D, Williams A, Wolstencroft K, Taylor C. Web Services for the DDSM and Digital Mammography Research. In: Astley SM, Brady M, Rose C, Zwiggelaar R, Eds. Digital Mammography. IWDM 2006. Lecture Notes in Computer Science. Vol. 4046. Berlin, Heidelberg: Springer Pp. 376-83.

15. Siu AL.Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Int Med 2016; 164(4): 279-96.
DOI: https://doi.org/10.7326/M15-2886

16. Radswiki T, Elfeky M, Ashraf A, *et al.* Breast calcifications. Available from: Radiopaedia.org (Accessed on 13 Jul 2024).
DOI: https://doi.org/10.53347/rID-13952

17. Paruchuri VA, Granville BC. A case-based reasoning system for aiding physicians in decision making. Intelligent Inform Manag 2020; 12(2): 63-74.
DOI: https://doi.org/10.4236/iim.2020.122005

18. Rawat D, Sharma S, Bhadula S. Case based reasoning technique in digital diagnostic system for lung cancer detection. In: 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1355-61.
DOI: https://doi.org/10.1109/ICCES57224.2023.10192863

19. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E. Principal component analysis. Nat Rev Methods Primers 2022; 2: 100.
DOI: https://doi.org/10.1038/s43586-022-00184-w

20. Zicheng G, Jiping X, Yi W, Min Z, Liang Q, Jiaxuan Y, *et al.* A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis. Open Life Sci 2022; 17(1): 1600-11.
DOI: https://doi.org/10.1515/biol-2022-0517