A Machine Learning Approach to Breast Cancer Detection in Mammograms

Humberto A. Salinas-Cortés[1], José M. Martínez-Serrato[1],
Vianney Muñoz-Jiménez[1](✉), and Marco Ramos[1]

[1] Universidad Autónoma del Estado de México, Toluca, Estado de México, México
hsalinasc001@alumno.uaemex.mx; jmartinezs595@alumno.uaemex.mx;
vmunozj@uamex.mx; maramosc@uaemex.mx

Abstract. Breast cancer poses a grave threat, being a primary cause of cancer-related fatalities for women globally. Early detection and precise diagnosis are pivotal in enhancing the likelihood of survival. Computer-aided detection systems can aid radiologists in detecting breast cancer on mammograms. This paper proposes a CAD system using Vector Field Convolution (VFC) and active model deformation for mass segmentation and Random Forest (RF) for breast cancer detection. The CBIS-DDSM database provides the mammogram images. Mammograms are processed with two focuses: the entire mammogram and the region having the mass. Preprocessing encompasses noise reduction, binarization, contour-based masking, erosion dilation and Hough transform for the mammogram, and Rough Set (RS) theory-based noise filtering and interference discernment for the region of interest. Segmentation is achieved using an active model deformation approach that minimizes internal and external energy to capture the desired features. Our research results reveal compelling evidence, suggesting that mammogram segmentation utilizing VFC exhibits superior accuracy compared to manual segmentation. Furthermore, comparing various machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), provides evidence supporting the superior performance of RF over SVM and KNN. Specifically, RF achieved, on its best performance, an accuracy of 83%, proving its effectiveness in mammogram classification tasks using 19 features from segmented mass and the full mammogram.

Keywords: Breast Cancer Mammograms · Pattern Recognition · Random Forest

1 Introduction

According to information provided by the National Institute of Statistics and Geography (INEGI) in México, in 2022, malignant tumors took 87,880 lives in people over the age of 20 years and older. Breast cancer itself is responsible for 9.0% of these deaths (7888).

J. de J. A. Flores Cuautle et al. (Eds.): CNIB 2024, IFMBE P 116, pp. 41–51, 2025.
https://doi.org/10.1007/978-3-031-82123-3_4

99.4% of these cases were diagnosed in women, while 0.6% were diagnosed in men [1]. The World Health Organization (WHO) said that "by decreasing worldwide breast cancer mortality by 2.5% annually, we could prevent 25% of breast cancer fatalities by 2030 and 40% by 2040 in women below 70 years old". The three fundamental elements to accomplish these goals are promoting health for early identification, prompt diagnosis, and all-inclusive management of breast cancer [2].

Early detection is a critical phase in the fight against breast cancer. Mammography is widely used across the globe; it is relatively easy to detect cancerous cells that are small and difficult to palpate, making it one of the most prevalent methods for identifying breast cancer. This method's accuracy largely depends on the radiologist's expertise, with an approximate success rate of 70% [3]. Mammography is a radiographic examination of the mammary gland [4]. From mammography, we can find asymmetry between breasts, architectural distortion, calcifications, mass, mass density, mass margins (circumscribed, micro-lobed, ill-defined, spiculated), and mass shape (round, oval, lobular, irregular). In the procedure usually followed for mass detection, the radiologist is tasked with identifying the lesion, assessing the likelihood of malignancy, and suggesting additional examinations [5]; classifying a mass as malignant or benignant is a complex task, and it is why the interest on developing an efficient and accurate Computer Aided Diagnosis (CAD) systems have expanded over the recent years.

CAD systems are tools that help radiologists improve accuracy and reduce the need for unnecessary invasive procedures. This work focuses on image segmentation and classification by implementing a Random Forest classifier considering 19 features from the mass and from the full mammogram.

2 Related Work

In in this section, we show the most works on breast cancer used according to the facilities used to detect these diseases.

R. Vijayarajeswari et al. [3] achieved an accuracy of 94% classifying by malignant and benign, proposing a methodology based on the Hough transform for feature extraction, taking into consideration four features from the whole mammogram and Support Vector Machine (SVM) as a classifier, M. Dong et al. [5] focuses on where the suspicious mass is located as a region of interest (ROI), they achieve an accuracy of 97.73% by implementing a vector field convolution (VFC) to segment mass from the background, and applying a Random Forest (RF) classifier which uses 32 characteristics extracted from ROI.

Rough Set (RS) theory has been used in medical image enhancement due to its capacity to manage uncertainty that arises from incorrect, noisy, or incomplete information [5, 6].

In medical images, the contrast has usually been increased by using histogram equalization, Contrast Limited Adaptive Histogram Equalization (CLAHE) is a recommended approach for images with uneven intensity distribution across varied parts of the image [7]. Several approaches to segmentation have been proposed depending on ROI, which is the technique that is going to be used. VFC is a powerful active contour model that leverages a vector field to locate points of energy equilibrium, it incorporates internal and external energy. The advantages of VFC include its low sensitivity to noise; however, it does have a relatively high computational cost [8].

A particular study using Deep learning algorithms presents an innovative algorithm designed for the accurate detection of breast cancer from screening mammograms. This method, initially requiring only lesion annotations, this approach later relies solely on image-level labels, circumventing the need for detailed lesion annotations, which are often scarce [9].

For early detection of breast cancer, recent advances in artificial intelligence (AI) and machine learning (ML), magnetic resonance imaging (MRI) and convolutional neural networks (CNN), have shown promise. Studies have highlighted the development of models such as CNN Improvements for Breast Cancer Classification (CNNI-BCC), significantly improving the accuracy of identifying cancer subtypes. These methods are often hampered by the high computational resources required to process intricate image data associated with breast cancer.

A pioneering study introduced an efficient deep learning model for mammography analysis optimized for varying image densities and designed to conserve computational resources. This approach employed a meticulous feature selection strategy, integrating craniocaudal and mediolateral mammogram views, bolstering diagnostic accuracy, the study demonstrated its superior computational efficiency and precision breast cancer diagnosis, marking a significant step forward in screening technologies [10].

3 Methodology

This section proposes an innovative methodology considering different skills to detect breasts efficiently. This skill of mammograms is processed using two focuses. On one side, the whole mammogram is processed; on the other hand, it focuses on the region where the mass is allocated. Figure 1 shows the methodology proposed. In the next paragraphs, we describe how to implement the different skills.

3.1 Image Acquisition

Mammograms were obtained from the CBIS-DDSM dataset [11], which contains 753 calcification and 891 mass cases, each with a mammogram, ROI image, mask, and CSV file with data divided into train and test cases. This work focuses on mass cases, JPEG images are used for computational efficiency, and mammogram images are standardized to 1024x1024 pixels, where ROI is extracted considering a 200x200 pixels window since all masses at CBISS-DDSM fit on it and considering the computational cost of VFC.

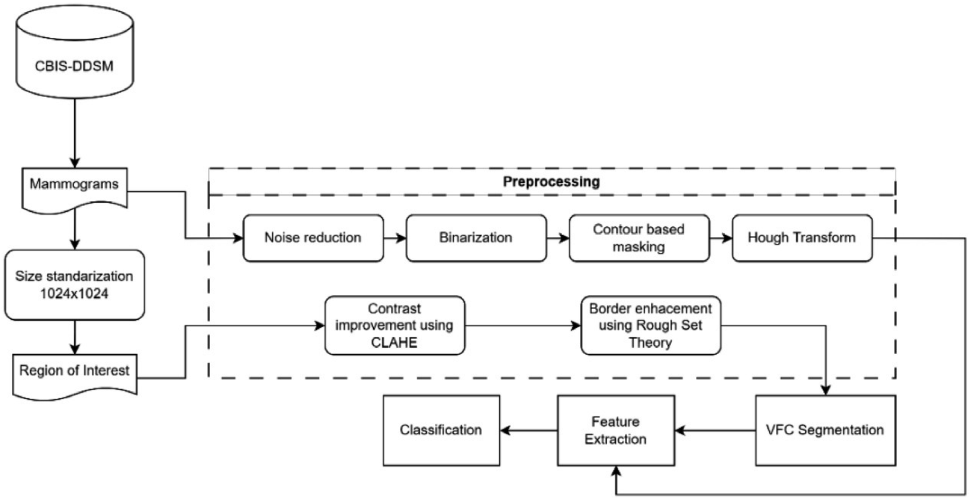

Fig. 1. Methodology proposed for mammography breast cancer.

3.2 Preprocessing

Given the wide range of gray-level distributions seen across different images, preprocessing medical images presents significant complexity. Each image is enhanced separately, extracting features from the complete mammogram and ROI-cropped images.

Mammogram Mammograms undergo several modifications to enhance their quality and ease subsequent processing. A median filter is applied to reduce noise. As can be seen in Fig. 2, extraneous information such as tags, identifiers, or mammogram specifications are displayed. A binarization is performed to remove this unwanted information. Otsu's adaptive thresholding method was employed. The largest contour length was identified, and a mask was created. The image was then subjected to erosion and dilation operations using a square 3x3 structural element, to delete extraneous information from the edges. The mask was applied, resulting in the acquisition of a refined mammogram image, finally a Hough Transform is applied to use the accumulator to extract features.

Region of Interest CLAHE [7] is performed to improve the ROI contrast and make edge detection an easier task. RS is a significant mathematical tool for navigating uncertainties from imprecise, noisy, or incomplete information [6]. The implementation of RS considers three factors: 1) the ROIs, 2) The gray level of the mass and its surroundings differs and 3) The RS method can manage the ambiguity and uncertainty associated with medical imaging [5].

Let ROI be the universe of disclosure U, C_1 be the gradient attribute, and C_2 be the noise attribute as condition attributes of the RS method. U produces sub-images U_{sub} and U_{base} according to the next indiscernibility relation concept.

$$R_{C_1} = \{(i,j) | : \nabla(i,j) > P\} \quad (1)$$

$$R_{C_2} = U_m U_n \{S_{mn} |: \text{int}|S_{mn} - S_{m\pm1,n\pm1}| \geq Q\} \quad (2)$$

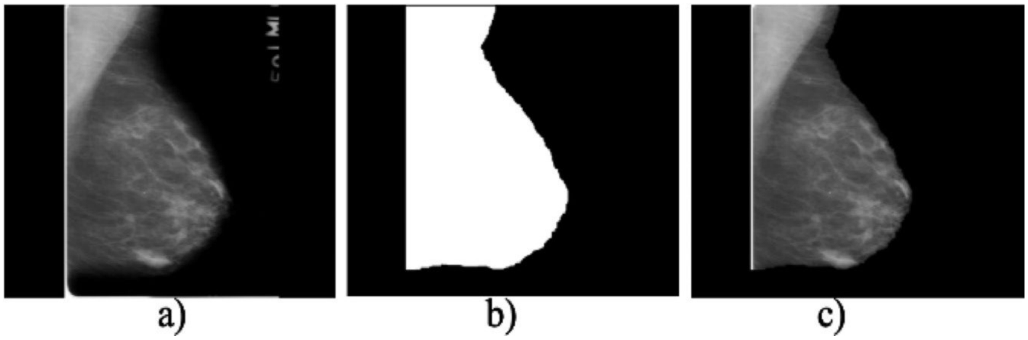

Fig. 2. Preprocessing stage: a) original image, b) binary image, c) preprocessed image.

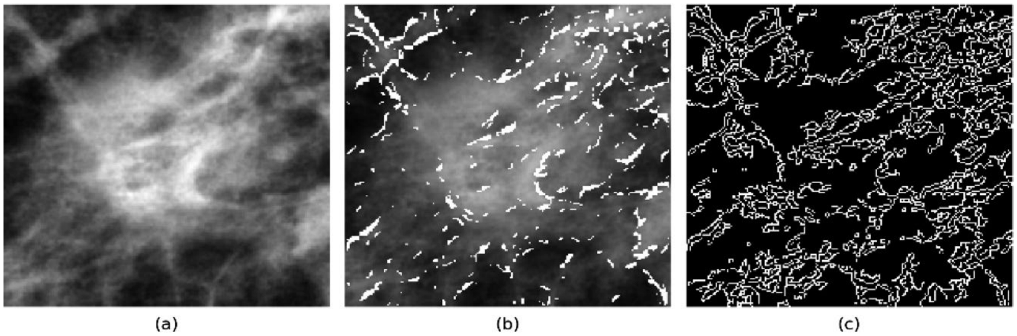

Fig. 3. Preprocessing ROI: a) original mass, b) enhanced mass, c) edge mass.

Where $\nabla(i,j)$ is the gradient function, P represents the gradient threshold, which is determined as the 95th percentile of the gradient histogram of frequency, S_{mn} denotes intensity mean sub-block 3 × 3, Q represents the noise threshold calculated using the Otsu's threshold method, applied over the image assembled by the mn Q subblocks. Equation (1) represents the possible edges of the mass, Eq. (2) represents these sub-blocks that represent possible noise. Given U, the sub-images are generated as follows (Eqs. 3 and 4).

$$U_{sub} = \left(\overline{R_{C_1}} - R_{C_2}\right) \cdot U \tag{3}$$

$$U_{base} = U - U_{sub} \tag{4}$$

U_{sub} is enhanced by applying the transform Eq. (5) proposed by [5]:

$$U_{sub}\left(i,j\right) = \alpha \cdot U_{sub}\left(i,j\right)^{\beta} \tag{5}$$

Experimentally, setting $\alpha = 1.5$ and $\beta = 5/3$, the boundary contrast between the mass and surroundings is augmented. Final image is generated by the addition of both sub-images. The edge map is generated by applying the Canny operator, See Fig. 3.

3.3 Segmentation

Active models deform within the image domain to capture a specific feature by minimizing an energy function while adhering to specific constraints [8]. Active models have backgrounds in concepts like Parametric Active Contours and External Forces. Active contour models aim to minimize the differences between external and internal forces. The deformation operation is performed by solving the numerical approximation (Eq. 6).

$$\left(I+\tau A\right)v^{t+1}=v^{t}+\tau F^{t} \tag{6}$$

Where I is the identity matrix of dimension $M \times M$, M corresponds to the number of elements (snaxels) along the contour, v^t denotes the current contour, and F symbolizes the external forces acting on each snaxel. A is an $M \times M$ cyclically symmetric pentadiagonal matrix responsible for computing the internal force. Equation (7) defines the external force F:

$$F\left(i,j\right)=VFC\left(i,j\right) \tag{7}$$

Where the VFC is defined as the convolution of a kernel over the edge map of an image. This kernel consists of a magnitude, $m(i,j)$, and a unit vector, $n(i,j)$, pointing to its origin. The magnitude $m(i,j)$, draws inspiration from Newton's law of universal gravitation [8].

The initial approximation of the snake curve is a circle because is the fittest shape to most of the masses, the radius is acquired by obtaining the minor axis of the manual mass segmentation provided by CBIS-DDSM and generating the approximation close to the mass dimensions. The snake curve is then deformed by iteratively applying the deformation Eq. (6), with ten deformations and a remeshing operation performed ten times. The remeshing operation is used to add or remove snake elements (snaxels) as the snake becomes larger or smaller, respectively, see Fig. 4.

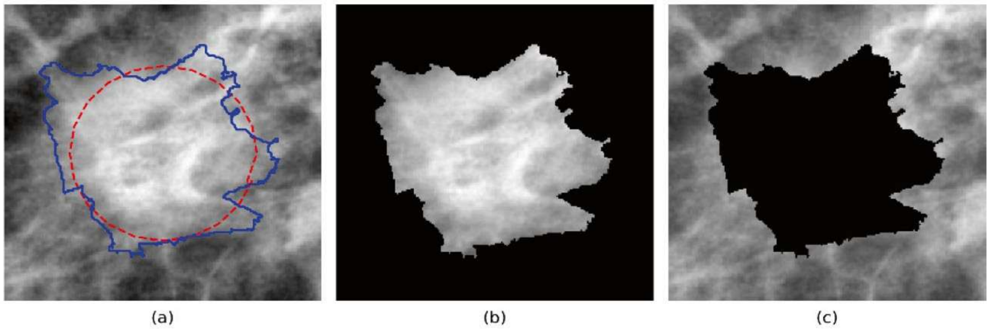

Fig. 4. Segmentation process: a) mass image with final contour (blue) by the initial contour (red), b) mass cropped image, c) background of ROI.

Table 1. Features extraction.

Feature description	Region (Origin)
Breast density	CBIS-DDSM
Left or right breast	
Mass shape	
Mass margin	
Area	Mass
Shape factor	
Mean	
Standard deviation	
Skewness	
Kurtosis	
Normalized radial length entropy	
Margin gradient mean	
Margin gradient standard deviation	
Fluctuation Mean	ROI
Fluctuation standard deviation	Background mass
Skewness	
Entropy	
Kurtosis	
Gray level mean	Full mammogram

3.4 Feature Extraction

Features are extracted to determine whether the mass is benign or malignant. Our research takes into consideration 19 features; CBIS-DDSM provides breast density, if is left or right breast, mass shape, and mass margin [11]; Dong et al. [5] achieve an accuracy of 97% by using 32 features, from which we take in consideration the mass area, mass shape factor, normalized radial length (NRL) entropy, margin gradient mean and standard deviation, mass gray level mean, standard deviation, skewness, and kurtosis, fluctuation of ROI mean, background fluctuation mean and standard deviation, background kurtosis, and Vijayarajeswari et al. [3] proposes four features, from which we take in consideration complete mammogram gray level mean. Table 1 shows the features computed for this work.

3.5 Classification

Classification becomes the focus of a CAD system, distinguishing between benign and malignant masses. Random Forest (RF) is a classifier proposed in 2001 by Breiman [12]. RF comprises an ensemble of decision trees, each constructed using a random vector sampled independently with an identical distribution across all trees in the forest. The generalization error converges with probability one to a finite limit as the number of trees in the forest increases. The overall error associated with a forest of tree-based classifiers is influenced by the individual tree strength and the degree of correlation among them. RF's advantages are inherited from Decision Trees but complemented since RF contributes to mitigating overfitting by reducing overall variance by incorporating uncorrelated trees. RF handles regression and classification tasks effectively, demonstrating its versatility across various problem types. Additionally, it can estimate missing values.

This work is centered on RF, which has been reported to exhibit excellent performance in breast cancer classification. It is compared to a SVM and a KNN classifier, which have also been reported to achieve acceptable performance.

4 Results

The VFC algorithm has a computational cost of $O(n^2 \; log \; (n))$, where n represents the dimension of pixels on a squared image. For a 200 × 200 pixel image, where $n = 200$, this translates to approximately 92,042 operations per iteration. This work, 100 iterations were performed, resulting in a computational cost of $O(m \cdot n^2 \; log \; (n))$, where m is the number of iterations. With $m = 100$, the total computational cost per image becomes 9,204,200 operations. Considering that a typical processor can handle one million operations per second, the processing time per image is estimated to be 9.2 s. If the original ROI size were kept, the processing time per image would increase significantly, reaching approximately 67 s in the worst case.

Compared to conventional segmentation models, VFC demonstrates enhanced performance, as seen in Fig. 5. Owing to the low contrast often present in enhanced images, VFC surpasses binarization to achieve a more accurate approximation. To ensure the validity of our results, we employ the 5-fold cross-validation technique to evaluate the accuracy. Which randomly divides the data into five folds, were each one is taken as test set while the others 4 as training set, for this, measurement of relevance is *accuracy*, as seen in Table 2, at all five folds, RF outperformed SVM and KNN, being the fold one the best performance for RF. A confusion matrix is a tool that shows us the confusion of a model. It helps detect the problem with our model and allows us to calculate other different metrics.

Given a confusion matrix with true negative (TN), true positive (TP), false positive (FP), and false negative (FN) values. Let the malignant class be our positive class; then the random forest matrix ends as shown in Fig. 6.

By here, measurements of relevance for this work are *accuracy*, *true positive rate* (TPR, sensitivity) and *true negative rate* (TNR, specificity). Table 3 shows the accuracy, TPR, and TNR of all classifiers. On the best performances of each model, it shows

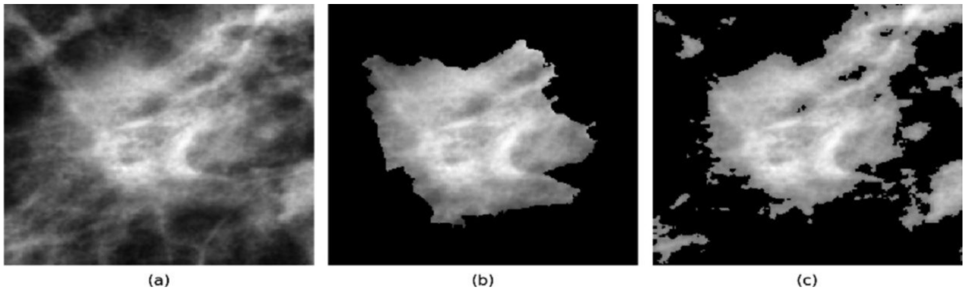

Fig. 5. Approximation: a) Mass image, b) VFC segmented mass, c) binary segmented mass.

Table 2. Cross validation.

Classifier	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVG
RF	**0.83**	**0.82**	**0.79**	**0.79**	**0.80**	**0.80**
SVM	0.78	0.76	0.75	0.73	0.74	0.75
KNN	0.78	0.78	0.74	0.74	0.76	0.76

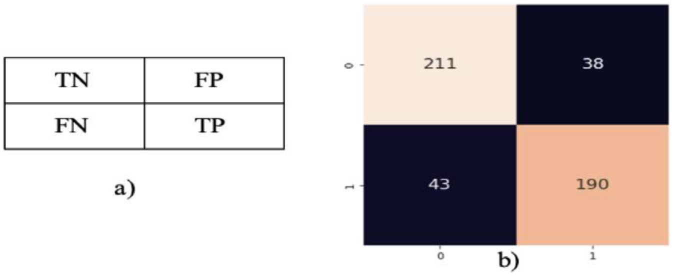

Fig. 6. a) Confusion matrix, b) Random Forest Confusion.

Table 3. Comparison between best results of different classifiers.

Classifier	Accuracy	TPR	TNR
RF	**0.83**	**0.82**	**0.85**
SVM	0.78	0.81	0.76
KNN	0.78	0.76	0.81

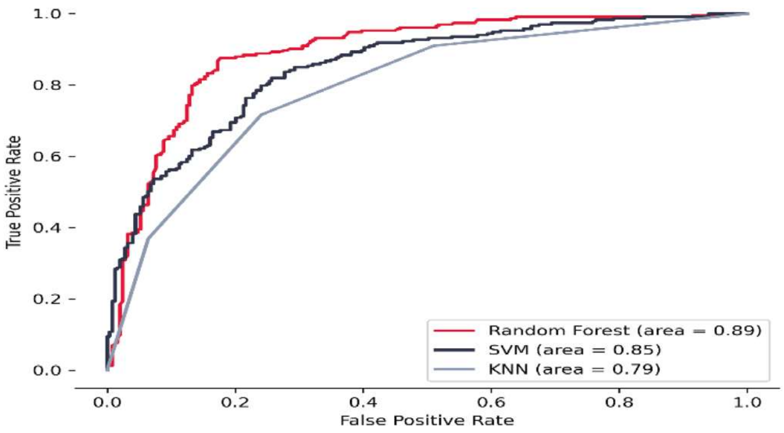

Fig. 7. ROC curve analysis.

that Random Forest outperforms SVM and KNN classifiers by achieving an accuracy of 0.83, and a better prediction nonmalignant case than malignant cases.

The ROC curves are used to evaluate the accuracy of each classifier (see Fig. 7). As can be seen in Fig. 7 and Table 3, our methodology effectively identifies malignant masses in the Curated Image Subset of the Digital Database for Screening Mammography (CBIS-DDSM), surpassing the performance of alternative classification methodologies such as SVM and KNN. This advance allows for a second evaluation, reducing incorrect diagnoses and financial burdens in breast pathology diagnosis.

The successful application of our proposed methodology can be attributed to the combined implementation of a precise segmentation method and an effective classification algorithm. The rough set method, which was used for the segmentation process, was found to be particularly adept at improving boundary contrast, while the CLAHE technique exhibited proficiency in improving gray-level contrast. When combined, these techniques create a robust and reliable foundation for feature extraction, which provides an alternative to traditional image enhancement methods. Subsequently, a random forest was used for classification. The improved classification accuracy achieved by our method can be ascribed to the inherent randomization in the processing, which effectively reduced the correlation between individual learners within the ensemble and provided variance reduction through averaging across a diverse set of learners.

5 Conclusions

The proposed methodology effectively combines a VFC-based segmentation approach with a random forest classifier to identify malignant masses in breast mammograms accurately. VFC and CLAHE allow for precise feature extraction through enhanced boundary and gray-level contrast, respectively. The integration of random forest facilitates robust classification, with superior performance demonstrated by an accuracy of 83%. Using the strengths of segmentation and classification techniques, our methodology presents a valuable tool for improving breast cancer diagnosis.

The segmentation phase is crucial in achieving better accuracy in medical imaging tasks. The segmentation process involves dividing the medical image into meaningful regions or structures, which can then be used for feature extraction. Accurate segmentation is essential for many medical applications, such as disease diagnosis, treatment planning, and surgical guidance.

The segmentation phase presents us with a gateway to future exploration and advancement. The methodology presented in this study, with its potential applications in other types of cancer, such as prostate cancer, inspires excitement and opens new avenues for research and development in the field of medical imaging.

Although the results presented are promising in detecting breast cancer, one limitation observed in this work is that the segmentation is directly linked to the BD; therefore, the initial segmentation requires adaptation to the BD that you want to treat. That is why future work consists of automating the segmentation for any Database you wish to process, regardless of the format of the images.

Validation of the medical sector is desirable when distributing this system in general care offices and specialty hospitals. Therefore, it is necessary to extend the learning system to enable multiple databases to predict breast cancer.

References

1. Estadísticas a propósito del día internacional de la lucha contra el cáncer de mama (19 de octubre). INEGI, 10 2023
2. WHO. Breast cancer. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
3. Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., Basha, A.A.: Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measure: J. Int. Measure. Confed. **146**, 800–805, 11 (2019)

4. Lopez, V.R.: Análisis de imágenes de mamografía para la detección de cáncer de mama 2012
5. M. Dong, X. Lu, Y. Ma, Y. Guo, Y. Ma, and K. Wang, "An efficient approach for automated mass segmentation and classification in mammograms," J. Digit. Imaging, vol. 28, pp. 613–625, 10 2015. https://doi.org/10.1007/s10278-015-9778-4
6. Yong, Y., Wang, B., Zhang, W., Peng, Z.: Low-contrast small target image enhancement based on rough set theory. In: Zhou, L., Li, C.-S., Yeung, M.M. (eds.), vol. 11, p. 68332I (2007)
7. Salem, N., Malik, H., Shams, A.: Medical image enhancement based on histogram algorithms. Procedia Computer Science. **163**, 300–311, 1 (2019)
8. Li, B., Acton, S.T.: Active contour external force using vector field convolution for image segmentation. In: IEEE Transactions on Image Processing, vol. 16, pp. 2096–2106, 8 (2007)
9. Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh, W.: Deep learning to improve breast cancer detection on screening mammography. Sci. Rep. **9**, 12 (2019)
10. Khalid, A., Mehmood, A., Alabrah, A., Alkhamees, B.F., Amin, F., AlSalman, H., Choi, G.S.: Breast cancer detection and prevention using machine learning. Diagnostics. **13** (2023)
11. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data. **4**, Article number: 170177 (2017)
12. L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5–32, 10, 2001. https://doi.org/10.1023/A:1010933404324