IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12810

Automatic Mammographic breast density classification Using machine learning approach

Milind S. Vadagave[1], Suraj K. Patil[2], Goutami S. Kamble[3]

Assistant Professor, CSE (Data Science), DYPCET, Kolhapur, India[1]

Assistant Professor, CSE (Data Science), DYPCET, Kolhapur, India[2]

Lecturer, Computer Science and Engineering, SITP, Yadrav, India[3]

Abstract: Breast cancer is one of the common types of cancer which is affecting health of women population in the world from last few decades. Breast cancer treatments always depend upon early detection, personalized approach and knowledge of disease. From last decade there are many deep learning and machine learning algorithm are implemented by many researchers but accuracy and precision not up to the mark hence mammographic breast density classification is done subjectively by radiologist.

In this research article implementation of machine learning algorithm is proposed for mammographic breast density classification. In this approach input images are preprocessed with help of morphological operations; pectoral muscle is removed by Hough transform and Canny Edge detection techniques are used.

The images are segmented with the help of Gaussian mixture model and features are extracted using GLCM feature extraction method and then SVM classification is performed on the images. With certain modification this algorithm is suitable for clinical practice.

Keywords: Breast Cancer, BI-RADS Classification, Preprocessing, Segmentation, Feature Extraction, Mammographic Breast Density.

I. INTRODUCTION

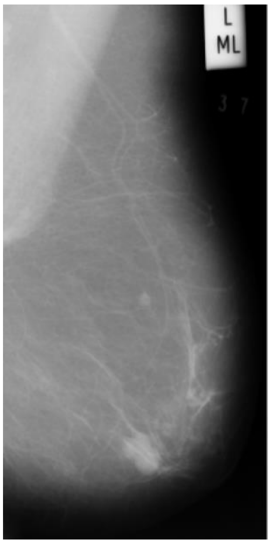
Breast cancer is a kind of most cancers that begins within the breast. It affects health of women population in the world from last few decades.Cancer begins when cells starts to develop out of control.

The cells of breast cancer usually form a tumor that can often be detects on an x-ray[1]. From the data within the year of 2018 near approximately 2.1 million women are newly recognized with bosom cancers at the same time as 0.65 million deaths happened due to this, disorder in year 2018 [2].

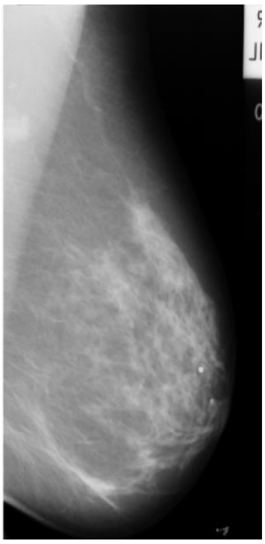
There is possibility of recovery from the cancer when it is detected in early stage. Breast density reflects the quantity of fibrous and glandular tissue in a female's breasts as compared with the amount of fatty tissue inside the breasts, as seen on a mammogram [3].

In the diagnosis of breast cancer mammograms are used to detect the malignancy. In order to remove confusion about mammographic result the "Breast Imaging Reporting and Data System (BI-RADS)" is used to classify four types of breast density classes [4, 5].

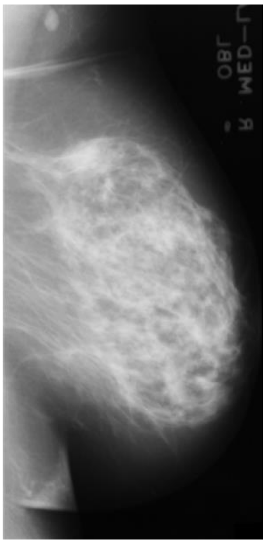
In the diagnosis of breast cancer mammograms are used to detect the malignancy. In order to remove confusion about mammographic result the "Breast Imaging Reporting and Data System (BI-RADS)" is used to classify four types of breast density classes [3, 4].



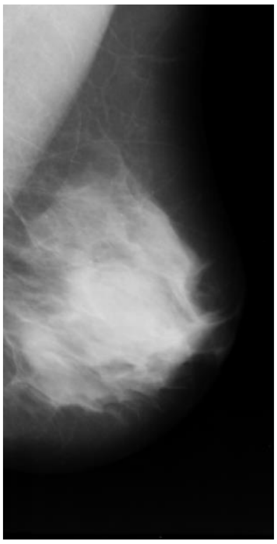
BIRADS Class A
Fatty Breast



BIRADS Class B
Scattered Density



BIRADS Class C
Heterogeneous Density



BIRADS Class D
Extremely Dense

On a mammography description, breast density is attributed to one of the following four categories

BIRADS A Class- Almost entirely fatty breast.
BIRADS B Class - A few regions of dense tissue are scattered thru the breasts.
BIRADS C Class - Evenly dense breasts.
BIRADS D Class- The breasts are extremely dense breast.

Basic objective behind this research article is to develop a machine learning algorithms for automatic breast density measurement towards improving accuracy for breast density measurement and classification.

This research article propose a machine learning algorithm which is divided into four sections: Section 2- preprocessing, section 3- Segmentation, Section 4- Feature extraction and classification, section 5- discussion and finding and section 6 conclude the article.

II. PREPROCESSING

The mammogram background and the pectoral muscle were removed in the pre processing step. Mammograms may be classified into three distinct areas: background, pectoral muscle, and breast region. Precise segmentation of the breast area from different locations is a major step for breast density dimension. The fundamental intention of this pre-processing is to split these unwanted regions. Breast boundary and pectoral muscle segmentation are crucial systems in computer assisted breast cancer detection.

➢ Mammographic images have labels and artifacts in background and these create impact on segmentation of mammogram.
➢ Fibro glandular tissue and pectoral muscle have same opacity for that for getting better accuracy we have to remove pectoral muscle because pectoral muscle creates errors in detection process of malignant tissue.

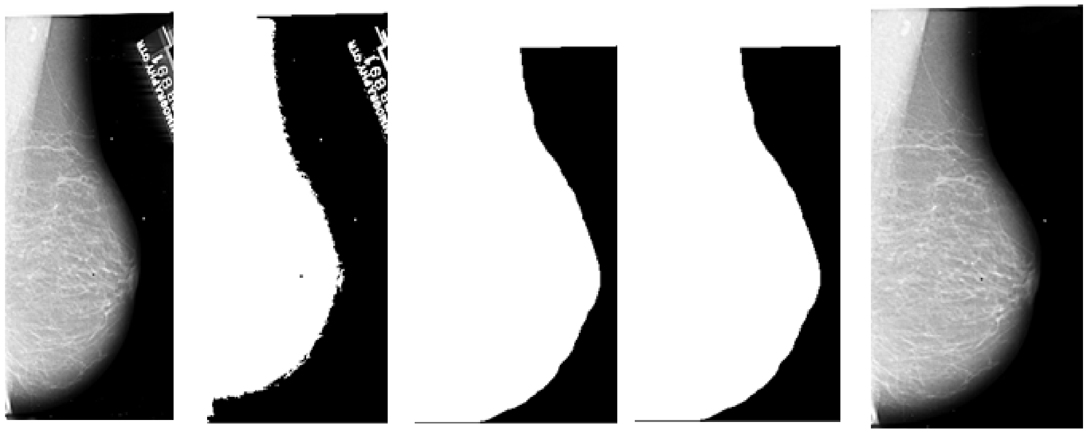
In this work, for removing these noises following operations are performed

Removing Labels and artifacts

Mammogram contains extra parts which are not needed for preprocessing steps; the extra part contains name, date, patient ID or other related information about patient and this information is bright gray as malignant tissue. This will create errors in segmentation process for that we have to remove these noises [9].

Following is the procedure for removing labels and artifacts:

1. Read the mammogram image.
2. Convert the image in grayscale.
3. Perform otsu thresholding algorithm on mammogram image.
4. Perform morphological operations on image.
5. Get largest contour from external contours.
6. Draw all contours as white filled on a black background except the largest as a mask and invert mask.
7. Apply the mask to the input image.
8. Save the results.



Original Image Threshold Morphological Operation Masking Result Label removed

Pectoral Muscle Removal

The pectoral muscle located close to the rib cage at upper portion of the rib. Pectoral muscle indicated as straight line, for that we utilized Hough transform and canny edge detection [10].

Hough transform

The Hough transform is a way that is used to extract features in processing. The most important use of the technique is to locate imperfect times of items inside a selected magnificence of shapes by means of a voting process. This voting technique is completed in a parameter area, from which object applicants are received as local maxima in a so-called accumulator space that is explicitly constructed by way of the algorithm for computing the Hough transform.

Canny Edge Detection

The Canny edge detector is an area detection approach used to discover a wide range of edges in mammograms. This is widely used detection algorithm in preprocessing process [12]. Canny Edge Detection is taken into consideration to be a higher aspect detection method than other techniques .This is because of following

1. Minimum Suppression

Edges applicants which aren't dominant of their neighborhood are not considered to be edges.

2. Hysteresis Process

While shifting alongside the applicants, given a candidate that is within the neighborhood of a part the edge is decrease.

Method	*Operating Principle*	*Complexity*	*Accuracy*
Hough transform	Canny Edge	Moderate	Good
Random Transform	Line Integral	Average	Average
Polynomial	Regression	Moderate	Average

Table-Different Preprocessing Techniques

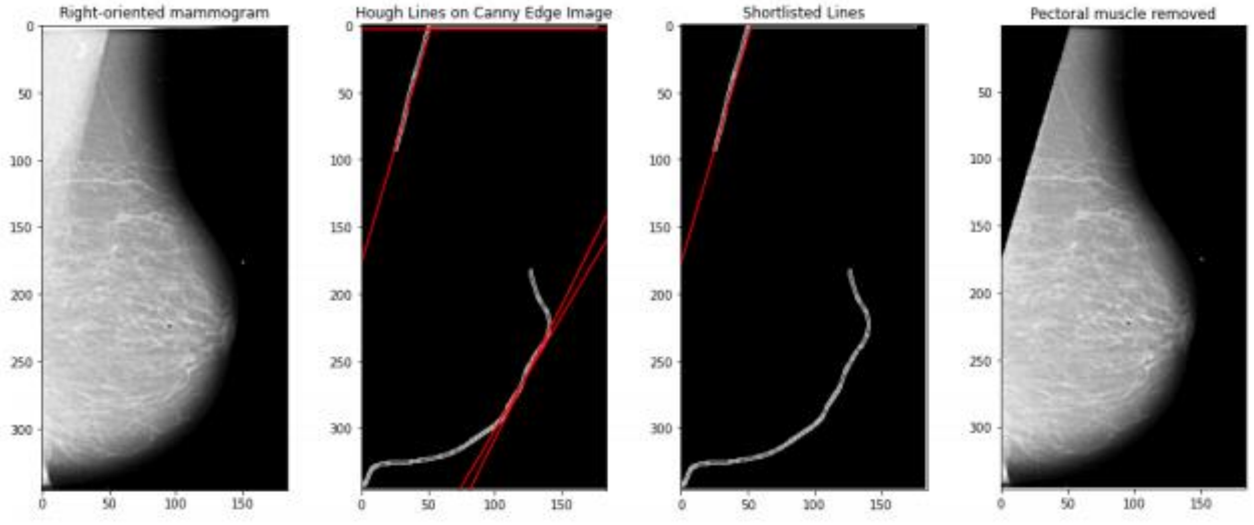


Fig- Remove Pectoral Muscle

III. SEGMENTATION

Segmentation method is involve to extracting a mammogram image into clear-out regions based on the belongings such as brightness, contrast, texture, gray-level and color. The medical segmentation is helpful for recognize anatomical structure, suspicious lesion, and measurement of tissue volume.

Gaussian mixture model

For segmentation of images this is the widely used statistical approach used by the researchers. Petroudi et al. and Ferrari et al. [13 and 14] proposed a Gaussian mixture model to segment fibro glandular tissue which uses gray-level values in a segmented mammogram.

Method	*Computational Complexity*	*Accuracy*
Dyadic wavelet decomposition	High	Moderate
Fuzzy-mean clustering	Moderate	Low
Gaussian mixture model	High	Moderate

Table- Segmentation Methods

IV. FEATURE EXTRACTION AND CLASSIFICATION

Feature extraction includes decreasing the quantity of sources required to describe a large set of information. Feature extraction is related to dimensionality discount. Feature extraction begins from a preliminary set of measured information and builds derived capabilities meant to be informative and non-redundant, facilitating the subsequent mastering and generalization steps.

GLCM feature Extraction

A statistical method of inspecting texture of the photo that examines the spatial courting of pixels is the GLCM (Gray-Level Co-prevalence Matrix). The Gray-Level Co-prevalence Matrix features classify the texture of an mammogram image with the aid of calculating how often pairs of pixel with precise values and in a exact spatial dating occur in an mammogram image, growing a GLCM, and then extracting statistical measures from this matrix.

After create the GLCMs the use of graycomatrix, you could derive some facts from them the usage of graycoprops. These data provide records about the texture of a mammographic image.

Statistics	*Description*
Contrast	Measures the local variations in the gray-level co-occurrence matrix.
Correlation	Measures the joint probability occurrence of the specified pixel pairs.
Energy	Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Table-GLCM Features

Feature	*Expression*
Contrast	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j)(i-j)^2$
Correlation	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i,j) \times (i-j) - \mu x \times \mu y / \sigma x \sigma y$
Energy	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} P(i-j)^2$
Homogeneity	$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \frac{P(i,j)}{1+\mid i-j \mid}$

Table-GLCM Feature Expressions

Classification

In this system, SVM classification algorithms will used. A support vector machine classifier is a supervised system getting to know model that uses classification algorithms for 2-group class issues. After giving SVM model sets of categorized training statistics for every class, they're capable of categorize new textual content. Support vector machines are unique linear classifiers which can be based on the margin maximization principle.

In general SVM classifier defined as

$$\mathbf{F(X) = sign} \ (W^T X + b)$$

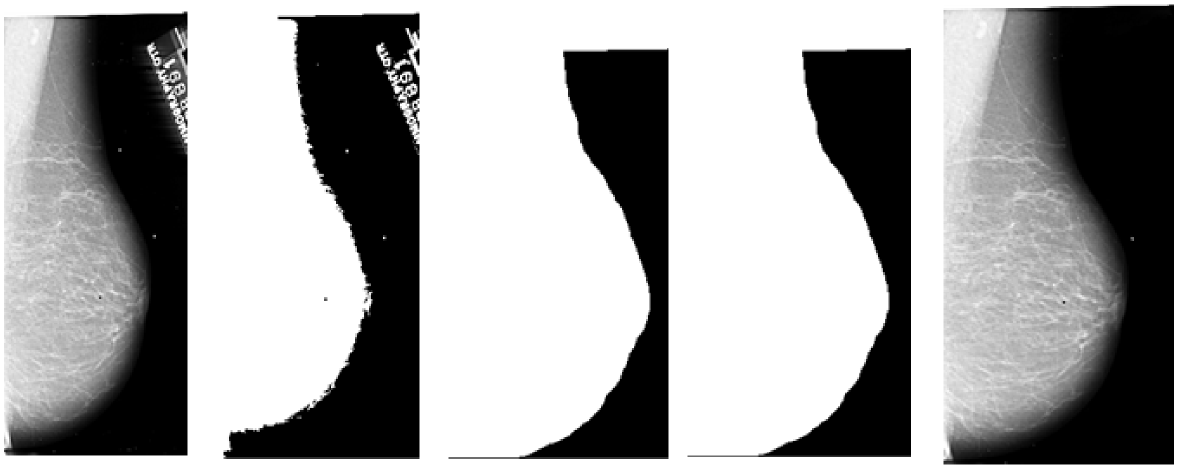
V. IMPLEMENTATION RESULT AND DISCUSSION

In this work we used 300 mammogram images from the source Internet and other clinical databases and created Mixed Dataset. This experiment is performed on 60 fatty, 60 scattered, 60 heterogeneous and 60 extremely dense mammogram images.

To clean the digital mammogram images we performed preprocessing techniques on images. For removing artifacts and labels different morphological operations were performed, figure shows the results of morphological operations. Breast boundary extraction method is used for background removal and Hough transform technique is used for remove Pectoral muscle the result of pectoral muscle is shown in fig.

in this direction could include the pre-processing stage for optimal selection of features inhibiting better textural description of the images. Perhaps, the exploration of advanced feature extraction technique to train the SVM can achieve higher classification accuracy in predicting the optimal segmentation algorithm for mammogram images.

DOI: 10.17148/IJARCCE.2023.12810



Original Image Threshold Morphological Operation Masking Result Label removed

Fig-Morphological Operations

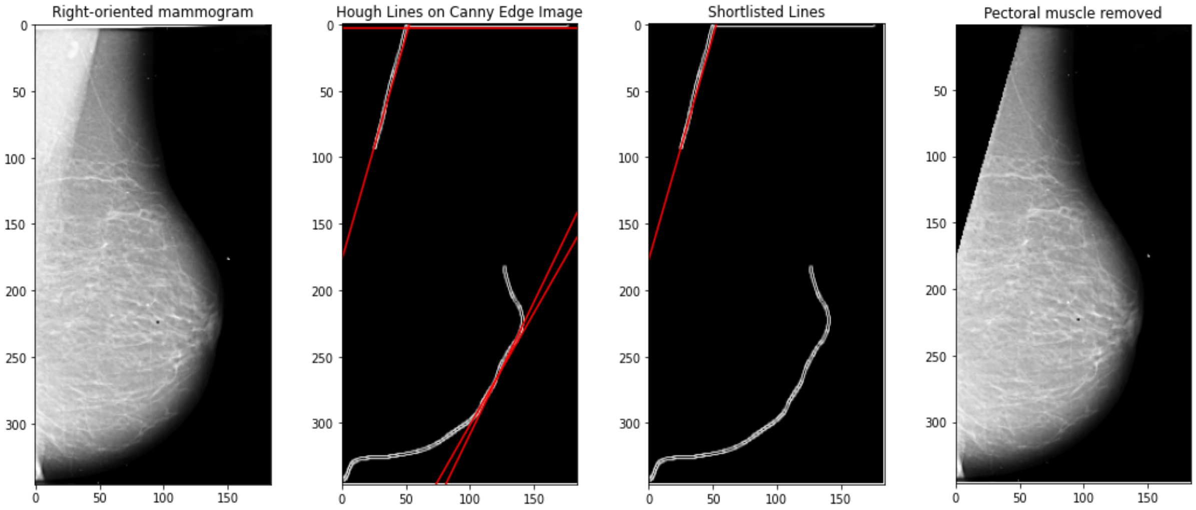


Fig- Removal of Pectoral Muscle

Result After Applying SVM Classifier

Classifiers	*Accuracy*	*Sensitivity*	*Specificity*
SVM	0.74	0.95	0.98

The accuracy, sensitivity and specificity are expressed as follows

1. Accuracy $= \frac{Count\ of\ accurately\ classified\ images}{Total\ number\ of\ images} \times 100\%$
2. Sensitivity $= \frac{True\ Positive}{True\ Positive+False\ Negative} \times 100\ \%$
3. Specificity $= \frac{True\ Negative}{True\ Negative+False\ Positive} \times 100\ \%$

DOI: 10.17148/IJARCCE.2023.12810

VI. CONCLUSION

This thesis demonstrates the flow of measuring breast density from mammogram using SVM. The images are segmented using the widely used threshold and edge detection techniques such as canny edge detection, Otsu threshold and Sobel algorithms. The selection of best segmentation algorithm from these algorithms is necessary since the image segmented accurately by one algorithm may not produce the same result for all the images. The SVM which is considered to be an efficient classifier for pattern recognition is employed to predict the optimal segmentation algorithm for the images. The image features are extracted and represented effectively using the GLCM in order to train the SVM. The highly correlated texture parameters constructed such as energy, contrast, entropy and inverse difference are determined from the GLCM at four different angles. Thus GLCM-based SVM achieved an accuracy of 96% in comparison to the histogram-based SVM with the accuracy of 80% when trained with three segmentation algorithms. The SVM trained using the GLCM illustrated excellent performance due to the additional knowledge extracted from the spatial relations in an image for better classification than using the image histogram alone. Perhaps, the statistical assessment using the confusion matrix states that the GLCM outperformed histogram method. It is evident that the feature descriptors greatly influence the classification accuracy.

REFERENCES

[1]. Center for disease prevention control https://www.cdc.gov/cancer/breast/basic_info/dense-breasts.htm

[2]. National Breast Cancer Foundation .http://www.nationalbreastcancer.org

[3]. Mayo clinic website https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[4]. Nithya Inger T. Gram, Ellen Funkhouser , Laszlo Tabar. “The Tabar classification of mammographic parenchymal patterns” https://pubmed.ncbi.nlm.nih.gov/9097055

[5]. John n. wolf. ”Risk for breast cancer development determined by mammographic parenchymal pattern”. https://pubmed.ncbi.nlm.nih.gov/1260729

[6]. Mona Jeffreys, Jennifer Harvey, Ralph Highnam, “Comparing a New Volumetric Breast Density Method (VolparaTM) to Cumulus”.https://link.springer.com/chapter/10.1007/978-3-642-13666-5_55

[7]. Michael A. Wirth, Jennifer Lyon, Dennis Nikitenko, Alexei Stapinski. “Removing radiopaque artifacts from mammograms using area morphology”. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5370/0000/Removing-radiopaque-artifacts-from-mammograms-using-area-morphology/10.1117/12.535372.short?SSO=1

[8]. Michael Wirth, Dennis Nikitenko, “Suppression of stripe artifacts in mammograms using weighted median filtering”. https://dl.acm.org/doi/10.1007/11559573_117.

[9]. Brad M. Keller, Diane L. Nathan, Yan Wang, Yuanjie Zheng, James C. Gee, Emily F. Conant, and Despina Kontos. “Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation” https://pubmed.ncbi.nlm.nih.gov/22894417/

[10]. Weibin Rong, Zhanjing Li, Wei Zhang and Lining Sun “An Improved Canny Edge Detection Algorithm”. https://www.ire.pw.edu.pl/~arturp/Dydaktyka/PPO/pomoce/06885761.pdf

[11]. Styliani Petroudi and Michael Brady, “Breast Density Characterization using Texton Distributions” https://pubmed.ncbi.nlm.nih.gov/22255462.

[12]. R J Ferrari , R M Rangayyan, R A Borges, A F Frère “Segmentation of the fibro-glandular disc in mammograms using Gaussian mixture modeling”. https://pubmed.ncbi.nlm.nih.gov/15191084/

[13]. Gaurav Kumar , Pradeep Kumar Bhatia “A Detailed Review of Feature Extraction in Image Processing Systems”.https://www.researchgate.net/publication/260952140_A_Detailed_Review_ofFeature_Extraction_in_Image_Processing_Systems/link/02e7e532be63da323 8000000/download.

[14]. P. Mohanaiah , P. Sathyanarayana , L. GuruKumarImage “Texture Feature Extraction Using GLCM Approach”. http://www.ijsrp.org/research-paper- 0513/ijsrp-p1750.pdf

[15]. T.S. Subashini , V. Ramalingam, S. Palanivel “Automated assessment of breast tissue density in digital mammograms” www.elsevier.com/ locate/cviu

[16]. Li Liu, Jian Wang, Kai He “Breast Density Classification Using Histogram Moments of Multiple Resolution Mammograms”.https://ieeexplore.ieee.org/document/5639662

[17]. Keir Bod, Sameer Singh, Jonathan Fieldsend, Chris PinderIdentification of masses in digital mammograms with MLP and RBF Nets. https://ieeexplore.ieee.org/document/857859

[18]. R. Nithya, and B. Santhi. “Computer Aided Diagnosis System for Mammogram Analysis: A Survey”. https://www.ingentaconnect.com/content/asp/jmihi/2015/00000005/00000004/art00001