International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Breast Cancer Screening Tool Using Gabor Filter-Based Ensemble Machine Learning Algorithms

P. Narasimhaiah*[1], C. Nagaraju[2]

Submitted: 15/11/2022 **Revised**: 17/01/2023 **Accepted**: 18/02/2023

Abstract: The most common kind of cancer among females that causes death is breast cancer. It's early detection and initial treatment can save the patient's life and also decrease the mortality rate. An efficient approach to finding breast cancer at an initial stage is screening mammography. However the diagnostic procedure is hand-operated, time-taking, and a specialist radiology is required and available only in hospitals, so the patient cannot check at their home with this technology. In the literature, many techniques have existed, but fail to produce a high accuracy rate due to the presence of noise, artefacts, pectoral muscles, and low contrast. Based on these reasons it is difficult for radiologists to find cancer at the initial stage. This paper presents the Gabor filter-based ensemble machine learning technique which gives a high accuracy rate in the presence of noise, artefacts, pectoral muscles, and low contrast. This method is applied on all MIAS Datasets, which consist of 322 mammogram images and produce an accuracy of 98.98%.

Keywords: *Artefact, Cancer, Diagnostic, Gabor Filter, Radiologist, Random Forest.*

1. Introduction

In terms of occurrence and death, the majority natural kind of tumour among women in the world is breast tumour. Current statistics show that a 25% increment in breast tumour occurrence and an increment in death due to breast tumours every year is 20%. Presently the prevailing method used to find breast tumours in the premature phase is screening mammography. Researchers showed that the death rate due to breast tumours can be decreased by detecting breast tumours at an early clocking and timely adjuvant treatment. The various approaches for breast tumour screening with the development of new medical technologies are mammograms, magnetic resonance imaging, ultrasound and computerized tomography. These distinct investigation techniques have distinct supremacy. The most effective and proven approach for screening breast tumours to reduce mortality is mammography [3]. Two different views of low-energy radiograph images of the complete breast in screening mammography are craniocaudal (CC) view and mediolateral oblique (MLO) view taken.

Multiple mammographic images of a patient produced during a mammography procedure are typically checked by radiologists individually. An expert radiologist is required to diagnose a mammogram for a breast tumour. This diagnostic procedure is time-consuming and labour-intensive. To overcome these problems and to assist radiologists to increase the performance of detection computer-aided detection (CAD) systems have been developed [4], [5]. The first stage in developing a CAD system for breast tumours is mammogram pre-processing. The pre-processing stage includes finding a region of interest and removing noise, labels and pectoral muscles. Including these in pre-processing stage produce a better accurate CAD system.

The important stage of pre-processing in mammography-based CAD systems is automatic pectoral muscle boundary (PMB) detection [6], [7]. An early sign of malignancy detection rarely in the breast is the non-mammary tissue region, the pectoral muscle of the breast. The automatic detection of breast tumours is interfered with the presence of micro-calcification, pectoral muscles and masses having similar intensity [8]. For automatic breast density quantification, pectoral muscles should be excluded [9]. The radiologist particularly checks the presence of pectoral muscles to reduce false negatives [10], because the overlying area of pectoral muscles is common to develop cancer. Therefore PMB detection is important before applying any automatic malignancy detection in the mammogram.

Present days researchers proposed various unmanned PMB detection methods, the most prevalent approach is an undeviating PMB line estimation, followed by its depuration [4]. Kowk et al. come up with repetitive starting and gradient tests to estimate the undeviating line of PMB [11]. When a pectoralis is small in size or overlapped by high-density mammary tissue, the straight-

[1]*Research Scholar. Department of CSE,YSR Engineering College of YVU, Proddatur - 516360, INDIA*
ORCID ID : 0009-0001-0587-421X
[2]*Professor,Department of CSE,YSR Engineering College of YVU, Proddatur-516360, INDIA*
Corresponding Author Email: narasimhareddypolu@gmail.com

line estimation technique fails to find PMB. Ferrari et al. come up with an undeviating line approximation of PMB entrenched on the Hough transformation [12]. Bora et al. proposed PMB detection based on a two-stage process [13]. When pectoral muscle density is low and small in size, straight-line approximation based on Hough transformation performance is poor. Morphological operations along with Hough transformation were used to detect PMB in Shi et al. [14]. Rampun et al. come up with a convolutional neural network (CNN) entrenched technique to find PMB [15]. To detect PMB automatically Rehman et al. suggested a two-step procedure, in the first step PMB straight-line approximation was obtained, whereas in the second step using the slope information of straight-line estimation Gabor purifier adjusted in the direction of PMB approximation [7]. Straight-line estimation of PMB based on PTG map-based Hough transformation is not robust. The PMB straight-line estimation detection fails when the pectoralis is small in size. The phase response (PR) and magnitude response (MR) of a multi-directional Gabor filter (MDGF) are used to eliminate the drawbacks of previous works [20], [21]. The efficiency of this method is high to find PMB although it diverges from its regular straight-line approximation.

The most important task of CAD systems that detect abnormality in the breast is to classify the detected tissue as normal, benign or malignant. A traditional CAD system classifies mammary tissue using well-designed handcraft features of mammographic images [16]. Methods based on hand-crafted features achieved great success in classification [17], [18], [19]. When data is complex the methods based on hand-crafted features suffer from the lack of adaptability. This problem can be overcome by using the Gabor filter-based texture feature learning approach.

The features obtained from the Gabor filters are fed to ML classifiers to classify the breast masses. To categorise the mammary glands as normal, not cancerous, or cancerous, the ML algorithms like decision trees, logistic regression and SVM are used. In this proposed work GFEML technique was used and it achieved high accuracy on the MIAS dataset. To obtain better performance two-phase Gabor filter is used to detect correctly fuzzy pectoral muscle boundaries and remove them properly. The features are extracted by using the Gabor filter and are fed to RF, LGBM and XGBoost ensemble machine-learning algorithms to classify normal, benign, or malignant. Here we choose an RF ensemble machine learning algorithm that gives the best performance.

2. Proposed Method

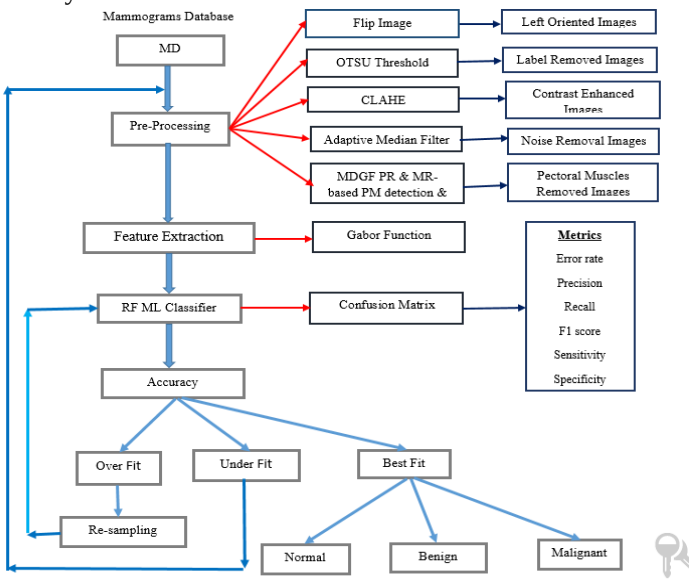


Fig1. Proposed system architecture

2.1. Pre-Processing

The mammographic database before being fed to machine learning algorithms, should be pre-processed to achieve the best prediction performance. The pre-processing of mammographic data involves i) change orientation ii) artefact removal iii) image enhancement iv) removing noise v) finding the region of interest vi) pectoral muscle boundary detection and elimination.

2.1.1. Change Orientation

To simplify our method instead of taking left and right-oriented mammograms only take left-oriented mammograms and we are looking for that direction. The mammogram orientation is determined using the method introduced by Shah. This method initially divides the mammographic image into two equal parts (left, right), if

the left part intensity sum is larger than the right part intensity sum, then the breast orientation is left, otherwise, the breast orientation is right. If the breast orientation is right then left oriented breast image is achieved by flipping the mammographic image.

Left-oriented breast image = { Original mammographic image If the left part intensity sum > the right part Intensity sum ; Flipped mammographic image If the left part intensity sum > the right part Intensity sum }

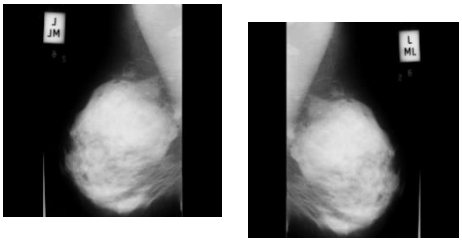


Fig 2. Result of the orientation for the image

2.1.2. Artefact removal

Scanned mammograms have different kinds of artefacts, which are generated at the time of the scanning process. Before applying any automated systems to the mammogram these artefacts should be removed. Distinct threshold values of mammographic images are obtained using the Otsu threshold method. The hard threshold of mammogram (M) is determined as 0.5 * lowest threshold value.

$$M(u, v) = \begin{cases} M(u, v) & \text{if } M(u, v) > 0.5*Tsh1 \\ 0 & \text{if } M(u, v) < 0.5*Tsh1 \end{cases}$$

In scanned mammogram hide out a very low-intensity background region with this small threshold value, thus separating artefacts from the mammary region. Thus mammogram region is selected as the largest connected area by eliminating different kinds of artefacts.

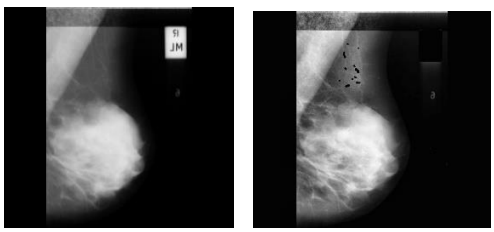


Fig 3. Result of artefact removal for the image

2.1.3. Image Enhancement

The upper right portion of mammographic images contains zero-intensity pixels corresponding to the background region. This background region of the mammogram forms a solid ferocity edge close to the mammary tissue area. The proposed automatic cancer detection system efficiency is affected by these strong intensity edges. A mammographic image with zero-intensity pixels is a low-contrast image. The contrast-limited adaptive histogram equalization (CLAHE) technique is used to increase the contrast of the mammogram.

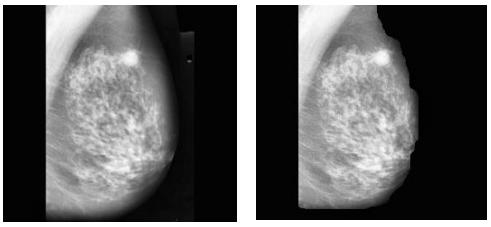


Fig 4. Result of image enhancement for the image

2.1.4. Noise Removal

A random variation in the brightness of an image is noise. The noise caused by external sources degrades the image signal. Noise may be introduced into the breast image during breast image acquisition and transmission, and processing. Speckle noise is multiplicative and this reduces mammographic image quality in diagnostic examinations. Noise breast images affect the performance of automatic systems. An adaptive median filter is used to de-noise the mammographic image. The pixels affected by noise are replaced with the median value of local pixels considering local variations over the entire mammographic image. The adaptive median filter procedure is given below:

Step 1: Compute the difference between the actual image and the median of the actual image.

$$D = \text{Image} - \text{Median (Image)} \quad (1)$$

Step 2: Smoothed image is obtained by convolution with the Gaussian kernel.

$$SI = \text{Image} * \text{Gaussian kernel (k) where } k < \text{size (Image)} \quad (2)$$

Step 3: Variability is the absolute difference between an Image and smoothed image.

$$V = |\text{Image} - SI| \quad (3)$$

Step 4: Smoothed variability is the product of variability and Gaussian kernel.

SV = V * Gaussiankernel (k)
(4)

Step 5: Find the ratio between D and SV.

R = D / SV
(5)

The mammographic image pixel value is changed by an average of the filter when R-value is larger than the threshold value, otherwise, preserve the original pixel value. The threshold value should be chosen so that only less than 10% of pixels in the original image were changed. ADMF reduce 80% of speckle from mammographic images.

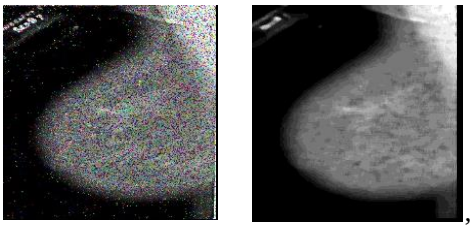


Fig 5. Result of noise removal for the image

2.1.5. Find The Area Of Interest

The varying size and shape of the pectoralis appear every time on the top south corner of the mammogram MLO view. The area of interest (AOI) is the rectangular area that contains pectoral muscles in mammographic images. The following procedure is used to determine the AOI.

Scan the mammographic image left to right starting from the 100th row until to get the first zero pixels.

The width of AOI is determined as the starting zero intensity pixel position rounded to a multiple of 8.

The AOI height is determined as 2/3 the height of the mammographic image rounded to a multiple of 8.

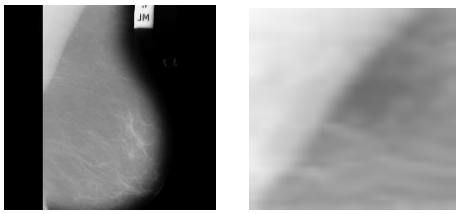


Fig 6. Result of selection of AOI of pectoral muscles

2.1.6. Pectoral Muscle Boundary Detection And Elimination

Interpretation of mammographic images is one of the most difficult in radiology. The difference in intensities between different tissue regions is not clear in low-contrast mammographic images. The top part of the pectoralis forms a sharp separation with the neighbour's low-concentration breast mass. The bottom part of the pectoralis forms an ill-defined boundary with the neighbouring breast tissue. Any automatic PMB detection algorithm deviates when mammograms with PMB have a fuzzy textural boundary with high-density glandular tissue. A major problem with unmanned PMB detection is having small size pectoral muscles. The performance of any PMB detection algorithm based on intensity is poor because of these reasons. The great promise of the Gabor filter is analysing images having specific frequency and direction. The bottom part of the pectoral muscle forms an ill-defined texture edge with high-density glandular tissue corresponding to relatively low-frequency information. A compact high-frequency MDGF assembly is designed here to cover all orientations where all PMB edges be in. The high pass-band purifiers are used to extract high- and mid-frequency information correlated to the strong and weak-intensity PMB fringes.

2.1.6.1 Multidirectional Gabor filter design

A 2-Dimensional Gabor purifier in the geometric domain is defined as a curved wave modified with Gaussian function. These purifiers are sensitive to occurrence and direction and can be expressed as:

$$h(a,b) = \frac{1}{2\pi\sigma_u\sigma_v}\exp\left(-\frac{1}{2}\left(\frac{a'^2}{\sigma_a{}^2} + \frac{b'^2}{\sigma_b{}^2}\right)\right)\exp(j2\pi Fa')$$
(6)

Where a' and b' are special coordinates rotated by an angle $\emptyset$

$$a' = a\cos\emptyset + b\sin\emptyset$$
(7)

$$b' = -a\sin\varphi + b\cos\emptyset$$
(8)

Here $$\sigma_a = \left(\frac{\lambda}{\pi}\right) * \left[\frac{(2^{SFB}+1)}{(2^{SFB}-1)}\right] * \sqrt{\frac{log2}{2}}$$
(9)

$$\sigma_b = \frac{\sigma_a}{SAR}$$
(10)

$\emptyset$ = Gaussian plane orientation concerning the x-axis.

F = Gabor bandpass filter centre frequency = $\frac{1}{\lambda}$

Spatial Frequency Bandwidth (SFB) = It is the response limit of a purifier defined as the occurrences of a given image vary from the preferred of occurrences F.

All bandpass filters are compressed and contact each other in the frequency domain if the parameters satisfy the following two conditions.

$$2L\frac{2}{\sigma_a fk} \geq 2\pi f_k$$
(11)

And

$$\frac{1}{\sigma_a(f_{k-1})} + \frac{1}{\sigma_a(f_k)} \geq f_{k-1} - f_k$$

(12)

Here the number of Gabor purifiers in the range of [0, π] is L and the centre frequency of k[th] purifier f_k is given as $f_k = {f_0}/{2^k}$, where

$f_0 = {3\pi}/{4}$, and k = 0, 1, 2, and so on L-1

The spatial Aspect Ratio (SAR) value for L=10 orientations is obtained as:

$$SAR = \frac{\sigma_a}{\sigma_b} = \frac{3}{f_k} * \frac{\pi * f_k}{2*10} = 0.472$$

(13)

The centre frequency of each sub-band for an image size N × N is 1 √ 2, 2 √ 2, 4 √ 2, and so on. Choose $\frac{N\sqrt{(2)}}{4}$ as the maximum possible central frequency band to extract information from the high and medium frequency correlate to the strong and weak intensity fringes. For INbreast and DDSM the ROI is less than 1024 pixels wide. Therefore N = 512, and centre frequency $512 * \sqrt{2}/4$ = 181.019 is rounded to 180 pixels/cycle. MIAS ROI is less than 256 pixels wide and therefore N = 128, and centre frequency $128 * \sqrt{2} / 4 = 45.248$ is rounded to 44 pixels/cycle. Observing three datasets of mammograms used in this research, PMB lies in these cases between 45^0 to 90^0 orientations. To obtain edge information of PMB covering 45^0 to 90^0 range of orientation, three compact high-pass bandpass filters are developed. If the PMB orientation of any mammogram is less than 45^0, then an additional filter is used to cover it, but this will slightly increase the PMB detection algorithm's computational complexity.

Convoluting AOI of the mammographic image with the projected 2-D Gabor purifier in the geometric domain obtain complex Gabor transform.

$$GT(AOI) = Conv2[AOI, h(x, y)] \quad (14)$$

Using the complex GT (AOI) magnitude response (MR) and phase response (PR) are computed as follows:

$$MR(x, y) = \sqrt{(Real(GT(AOI)))^2 + (\text{Imagin}(GT(\text{AOI})))^2}$$

(15)

$$PR(x, y) = \tan^{-1}\left[\frac{\text{Imagin}(GT(\text{AOI}))}{\text{Real}(GT(\text{AOI}))}\right]$$

(16)

The mask is analogous to the original place of the PMB area is obtained using the MR of the MDGF pool. The MR value is high for edges with high intensity tuned in the direction of the Gabor filter. At each pixel location maximum intensity values of all MRs give a maximum MR (MMR), MR contains only incomplete details about high and mid-frequency texture fringes and the remaining texture information is present in PR. PR can detect low-intensity edges that can be used to find PMB correctly even if an edge is fuzzy. All strong and weak intensity fringes that fall within the pre-set range of orientations procure by coalescing the Laplacian of PRs corresponds to MDGFs. Two edges with different intensities are detected very close to each other in the combined PR. To merge these intensity edges morphological dilation followed by skeletonization was applied. To obtain the part of true PMB, combine MR and PR of true location PMB area. The Laplacian combine of PR yields an unbroken edge of the PMB if it has clear-cut intensity, otherwise gives broken edges. Broken edges are connected using boundary search and merge algorithms to obtain real PMB regions.

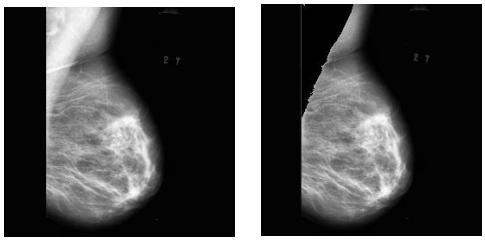


Fig 7. Result of pectoral muscles removal

2.2. Bank Of Gabor Filters For Features Extraction

Gabor filter is a texture descriptor used to obtain features by analysing the image frequency domain. The Gabor function is a product of a complex sinusoidal wave and a Gaussian function. The central frequency of each Gaussian function is very important to ensure that it covers all frequencies of the image. The 2-Dimensional Gabor function for each image point (x, y) is

$$g(x, y) = \exp\left(-\frac{x'^2 + y'^2\gamma^2}{2\sigma^2}\right) \qquad \exp\left(i\left(2\pi\frac{x'}{\lambda} + \varphi\right)\right)$$

(17)

Where $x' = x\cos\theta + y\sin\theta$, and $y' = -x\sin\theta + y\cos\theta$, λ represent the wavelength of sinusoidal factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, φ represents the phase offset, γ represent the spatial aspect ratio, and σ represent the standard deviation of the Gaussian envelope.

Smaller the value of σ higher the spatial resolution and the larger the value of σ lower the spatial resolution. The orientation parameter values are real in the range of (0, π). If the phase offset parameter is 0,π then the filter is Centro symmetric, if $-\frac{\pi}{2}, \frac{\pi}{2}$ then the filter is anti-Centro symmetric. The accurate value of the feature is decided by scale and orientation parameter values. Different values of the parameters are used to generate different filters. The

image features are coefficients of higher-order statistical parameters obtained from the filtered images.

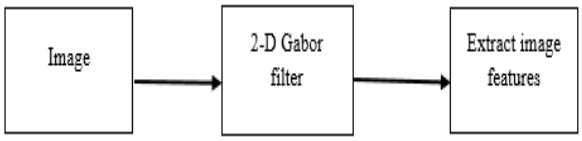


Figure 8: Flow diagram of the feature extraction method.

Figure 8 shows the feature extraction method using the Gabor filter bank. The various steps of the feature method are as follows:

1. Obtain compression of the input image in the spatial domain by performing skipping quantization such that there should be no loss of information.
2. Two-dimensional Gabor filters are obtained by setting values of scale, orientation, phase shift, threshold, and quantization. The number of filters depends on different values of these parameters.
3. Convoluting decomposed images with a Gabor filter obtains the filtered image.
4. Gabor features are obtained from these filtered images.
5. Similar features obtained from different orientations in the same scale are merged to reduce the number of features.

Extracted 32 Gabor features of each image from 32 Gabor filters obtained by varying orientation parameter values as $0, \frac{\pi}{4}, \frac{\pi}{2}, and\ \frac{3\pi}{4}$, varying σ parameter values as 1 and 3, varying λ parameter values as 0 and $\frac{\pi}{2}$, and varying γ parameter values as 0.05 and 0.5. In addition to Gabor features the following features are used to increase the accuracy further.

Sobel edge: 2-D spatial gradient of the image is obtained by using the Sobel operator. It identifies a high spatial frequency area corresponding to the edge. It finds at each point of the image a gradient magnitude.

The edge gradient magnitude is:

$$|G| = \sqrt{Gx^2 + Gy^2} \quad (18)$$

The angle of orientation of the edge is:

$$\theta = arctan\left(\frac{Gy}{Gx}\right) \quad (19)$$

Roberts's edge: The Roberts Cross operator finds the 2-D spatial gradient of an image. It identifies the area of the high frequency of an edge.

The edge gradient magnitude is:

$$|G| = \sqrt{Gx^2 + Gy^2} \quad (20)$$

The angle of orientation of the edge is:

$$\theta = arctan\left(\frac{Gy}{Gx}\right) - \frac{3\pi}{4} \quad (21)$$

Scharr edge: Scharr edge is used to identify the gradient using the first derivative. It detects changes in pixel intensity and used a gradient along the X-direction as G_x, and a gradient along the Y-direction as G_y.

$$G_x = \begin{bmatrix} +3 & 0 & -3 \\ +10 & 0 & -10 \\ +3 & 0 & -3 \end{bmatrix} * I \quad (22) \quad G_y = \begin{bmatrix} +3 & +10 & +3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} * I$$

(23)

The edge gradient magnitude is:

$$|G| = \sqrt{Gx^2 + Gy^2} \quad (24)$$

The angle of orientation of the edge is:

$$\theta = atan2(Gy, Gx) \quad (25)$$

Prewitt edge: Prewitt operator detects two edges, one along the direction of *the X*-axis and another along the direction *Y*-axis. It computes the approximate first derivative for changes in horizontal as G_x, and another for changes in the vertical direction as G_y.

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +1 & 0 & -1 \\ +1 & 0 & -1 \end{bmatrix} * I \quad (26) \quad G_y = \begin{bmatrix} +1 & +1 & +1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} * I \quad (27)$$

The edge gradient magnitude is:

$$|G| = \sqrt{Gx^2 + Gy^2} \quad (28)$$

The angle of orientation of the edge is:

$$\theta = atan2(Gy, Gx) \quad (29)$$

Gaussian filter: The Gaussian Smoothing Operator performs the weighted mean of pixels surrounding to pixels' Gaussian distribution. It generates a template of values, and these values are applied to a group of pixels

image. 2DGausian function defines template values.

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} \quad (30)$$

Sigma defines the amount of blurring. An approximation of a Gaussian function is the kernel:

$$G_{kernel} = \frac{1}{16}\begin{bmatrix}1 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 1\end{bmatrix} \quad (31)$$

Median filter: Median filters are non-linear filters used to reduce random noise in an image. The image is filtered by replacing the median of the values in the input window.

Median Img(x, y) = Intensity of middle pixel after arranging neighbour pixels in order by intensity value.

100	255	100
100	255	100
255	100	100

At location (1, 1) value is 255 after applying the median filter value at location (1, 1) is 100.

100	255	100
100	100	100
255	100	100

2.3. Machine Learning Models

Several Machine Learning algorithms are there to do the classification. Among these classification algorithms, ensemble learning algorithms are better to do the classification. The ensemble machine learning method is obtained by combining basic machine learning methods and producing an optimal predictive model. The breast masses are classified as normal, benign, or malignant in the proposed work ensemble ML methods random decision forest (RF), light gradient boosting machine (LightGBM), and extreme gradient boost (XGBoost) are used.

Random Forest is an ensemble ML classifier containing a large number of decision tree classifiers. Each decision tree is applied to a subset of the dataset RF. The accuracy of RF is the average accuracy of all decision trees. The predictor and response vectors of the training dataset are $X = x_1, x_2 \ldots, x_m$, and $Y = y_1, y_2 \ldots, y_m$ respectively. A random sample was selected with replacement from the training dataset repeatedly P times and fit in decision trees.

For P = 1, 2 ..., p:

1. Select a random sample (X_p, Y_p) from the training dataset (X, Y) with a replacement.
2. Train the tree f_p using the training sample (X_p, Y_p).

The Light Gradient Boosting Machine algorithm is based on advanced ensemble technique boosting. It takes less time even if the data set is huge. It requires a differential loss function to convert a weak learner to a strong learner.

For weak estimator F

$$\text{Loss} = \left(Y - F(X)\right)^2 \quad (32)$$

Negative gradient of loss $-\frac{\partial L}{\partial x} = -2 \times \left(Y - F(X)\right)$ (33)

Fit a weak estimator on $\left(X, \frac{\delta L}{\delta x}\right)$

The XGBoost is an extreme gradient-boosting ensemble ML algorithm, is scalable, distributed gradient-boosting. It is built upon decision trees. It has high predictive power and includes a variety of regularizations to reduce overfitting and improve overall performance. The training data set is $\{(x_i\ y_i)\}_{i=1}^{K}$, an initial loss function$\left(, f(x)\right)$, number of weak learners N, and a learning rate α

Initial model is $\hat{f}_{(0)}(x) = \underbrace{argmin}_{\theta} \sum_{i=1}^{k} L(y_i, \theta)$ (34)

Final model is $\hat{f}_{(N)}(x) = \sum_{n=0}^{N} \hat{f}_n(x)$ (35)

3. Experimental Results and Discussion

To analyse the proposed system performance publically available data set of all MIAS is considered for experimentation. The MIAS dataset contains 322 mammogram images, among them 207 normal, 62 benign, and 53 malignant. The breast images of all MIAS data set contains artefacts, and these artefacts affect the prediction performance of ML classification techniques. To improve the proposed method's performance, the Otsu threshold method is used to remove artefacts from mammographic images. Some mammographic images of the dataset are in low contrast. The low-contrast breast images affect the features extracted and these features may reduce the performance of ML algorithms. This problem can be overcome in the proposed by applying the CLAHE method to breast images. Breast images of all MIAS datasets have speckle noise and this noise reduces breast image quality in diagnostic examinations. In this proposed method ADMF is used to reduce speckle noise. All most all mammographic images in the data set have pectoral muscles, due to the presence of these muscles, it becomes difficult for radiologists to interpret mammographic images. In this proposed work the pectoral muscles were detected and removed using MDGF. The performance of classification methods used in the proposed system is done using a confusion matrix (CM). A CM is a table of rows and columns, where rows

represent actual values and columns represent predicted values.

True Positive: Predicted positive and it's true.

True Negative: Predicted negative and it's true.

False Positive: Predicted positive and it's false.

False Negative: predicted negative and it's false.

The CM is used to calculate the following parameters, and these are used to analyse the classification method performance.

Accuracy: The ratio of the sum of real positives and real negatives to all observations in test data. This can be expressed as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (36)$$

The accuracy is used to judge the efficiency of the classification method.

Error rate: The ratio of all incorrect predictions to all observations of test data. This can be expressed as

$$\text{Error rate} = \frac{FP+FN}{TP+TN+FP+FN} \quad (37)$$

The model is best when the fault rate is 0.0 and worst when the fault rate is 1.0.

Precision: The ratio of true positives over the total positives. This can be expressed as

$$\text{Precision} = \frac{TP}{TP+FP} \quad (38)$$

This measure assesses the model rate of positive prediction.

Recall: The ratio of true positives over actual positives. This can be expressed as

$$\text{Recall} = \frac{TP}{TP+FN} \quad (39)$$

This measure assesses the ability of the model to identify the actual true results.

F1score: The Pythagorean mean between recall and precision. This can be expressed as

$$\text{F1score} = \frac{2(p*r)}{p+r} \quad (40)$$

This score can be used as an overall metric that incorporates both precision and recall.

Sensitivity: The ratio of correct positive prediction over total positives. This can be expressed as

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (41)$$

Specificity: The ratio of negative predictions over total negatives. This can be expressed as

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (42)$$

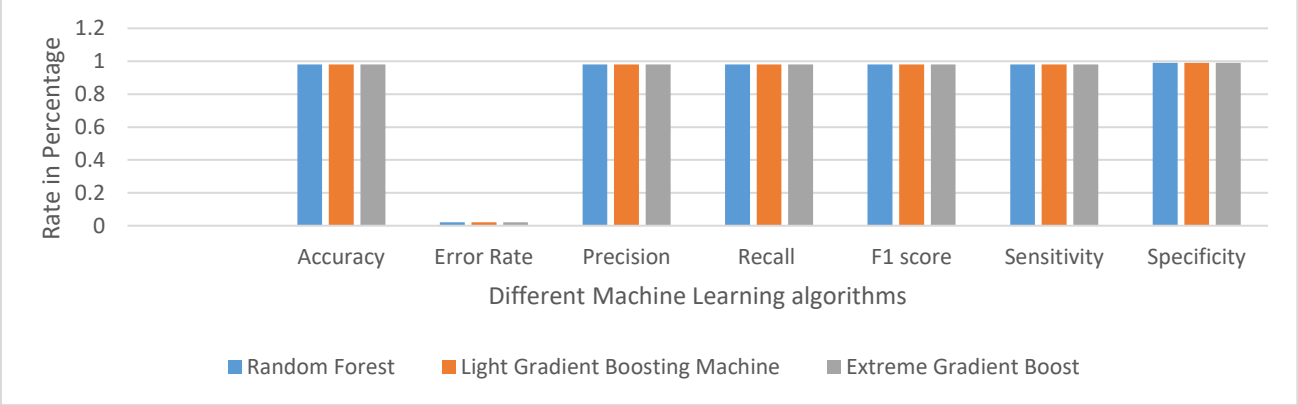
Table 1 illustrate the performance metrics like accuracy, error rate, precision, recall, F1 score, sensitivity, and specificity of three ensemble machine learning methods RF. LGBM, and XGBoost. These metrics are taken for the sample size of 40, which gives the best metrics comparison with other sample sizes. Plot 1 illustrates the accuracy of the three ensemble machine learning methods RF, LGBM, and XGBoost, showing that all three ensemble ML techniques give the same best accuracy of 0.9898 because optimised features are extracted using the Gabor filter and fed to these ensemble ML algorithms. Table 2 illustrates the various performance metrics of the RF ensemble learning method for different sample sizes. As the sample size increases the performance measures accuracy, precision, recall, F1 score, sensitivity, and specificity also increase, but the error rate decrease as the sample size increases.

Table 1. Performance metrics of ensemble learning methods RF, LGBM, and XGBoost.

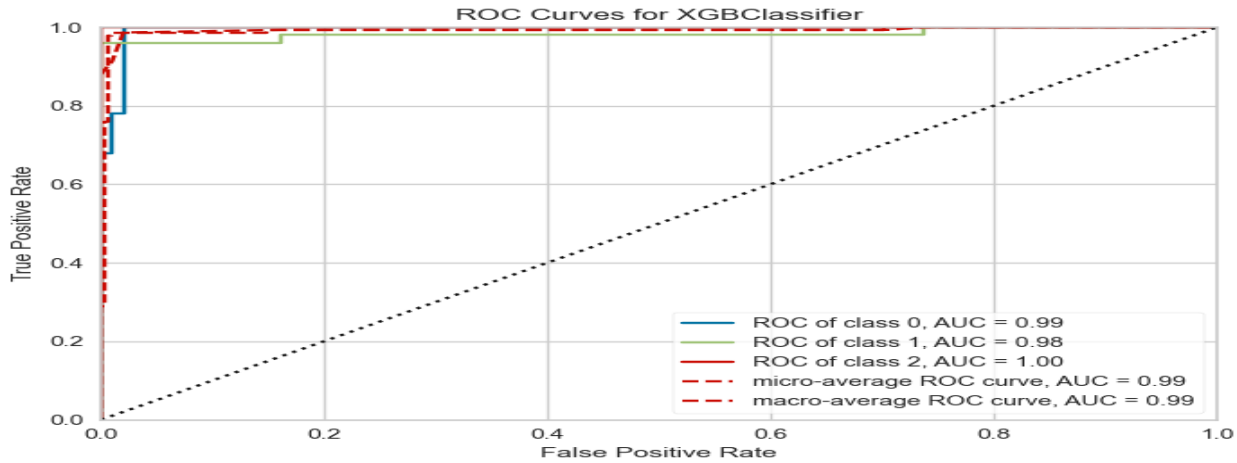
Metrics/Models	Accuracy	Error Rate	Precision	Recall	F1 score	Sensitivity	Specificity
RF	0.9798	0.0202	0.9798	0.9798	0.9798	0.98	0.9898
LGBM	0.9798	0.0202	0.9798	0.9798	0.9798	0.98	0.9898
XGBoost	0.9798	0.0202	0.9798	0.9798	0.9798	0.98	0.9898

Table 2. Performance metrics of ensemble learning method RF for different samples.

Data samples	Accuracy	Error Rate	Precision	Recall	F1 score	Sensitivity	Specificity
15	0.95	0.05	0.95	0.95	0.95	0.95	0.975
20	0.953	0.047	0.953	0.953	0.953	0.953	0.9764
25	0.96	0.04	0.96	0.96	0.96	0.96	0.98
30	0.9666	0.0334	0.9666	0.9666	0.9666	0.9666	0.98333
35	0.9714	0.0286	0.9714	0.9714	0.9714	0.9714	0.9857
40	0.9798	0.0202	0.9798	0.9798	0.9798	0.98	0.9898

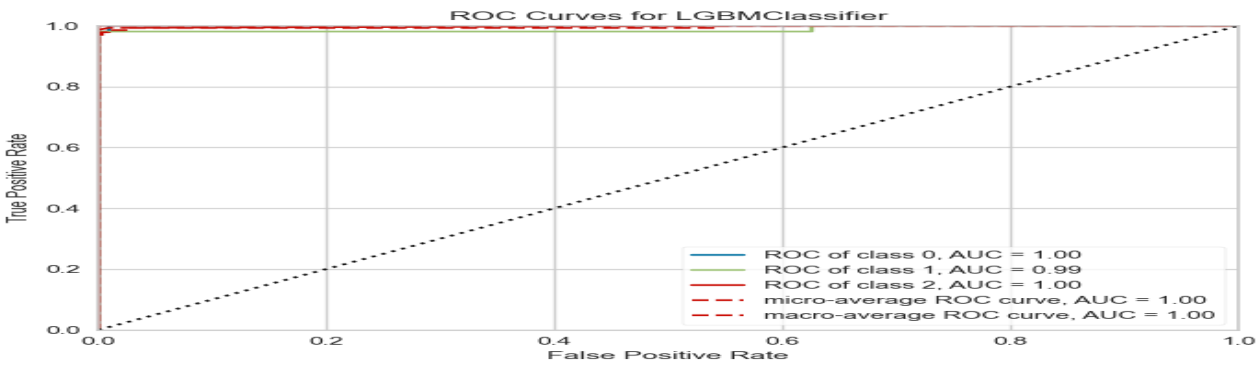


Plot 1. Accuracy comparison for three ensemble learning Methods RF, LGBM, and XGBoos



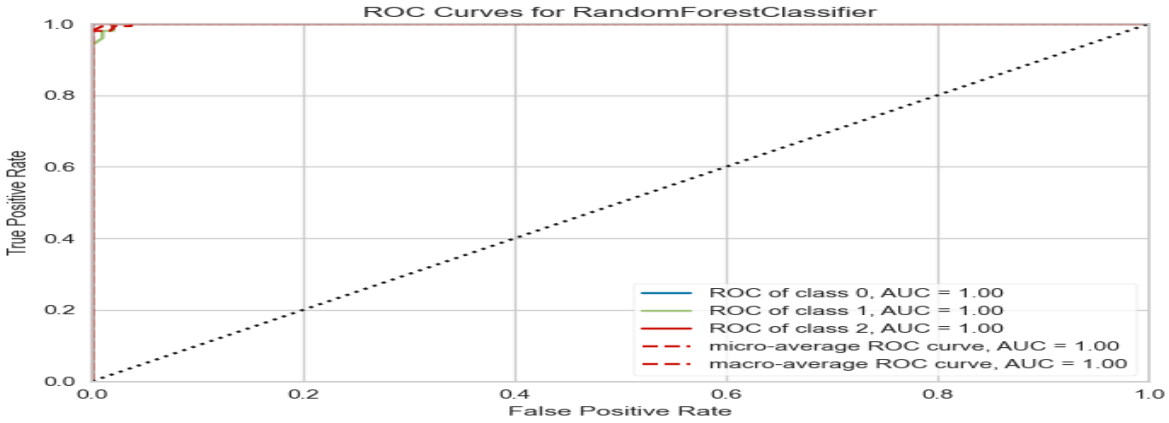
Plot 2. ROC curves of the XGBoost ensemble learning method.

Discussion of XGBoost ensemble ML algorithm: It is highly flexible, faster than gradient boosting, uses parallel processing power, and supports regularization. However, its performance is not good on sparse and unstructured data, is outlier sensitive, and is hardly scalable.



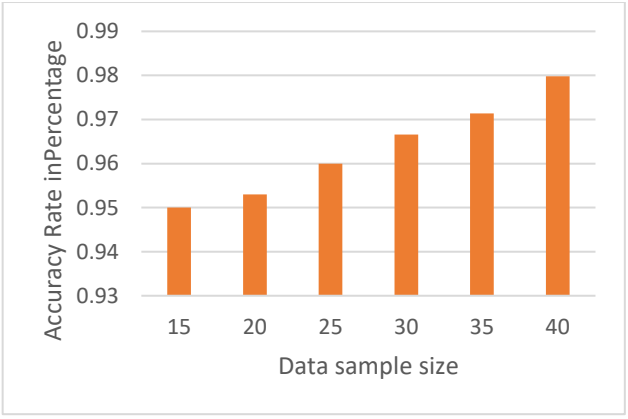
Plot 3. ROC curves of the LGBM ensemble learning method.

Discussion of LGBM ensemble ML algorithm: Its training speed is fast and has higher efficiency, and requires low memory. Its accuracy is better than other boosting algorithms. It is compatible with a huge amount of data and supports parallel learning. However, it can split the tree leaf-wise and produce complex trees and leading to overfitting.

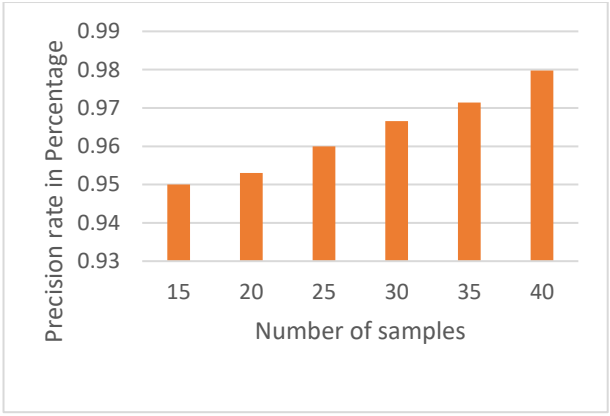


Plot 4. ROC curves of the RF ensemble learning method.

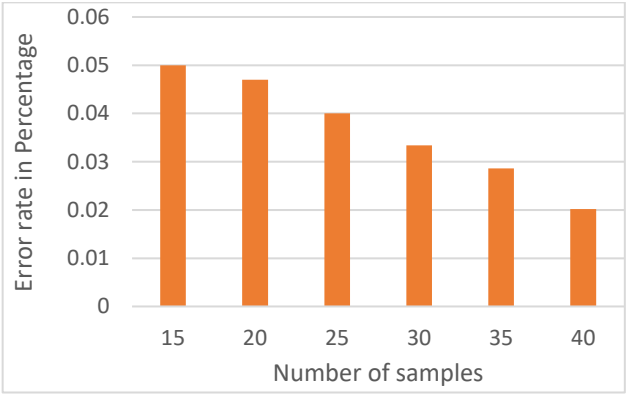
Discussion of RF ensemble ML algorithm: It is robust and stable and is less impacted by noise. RF can be used to solve the problems of regression and classification. It can work on both categorical and continuous data. It does not require feature scaling and handles non-linear parameters, missing values, and outliers. However, it creates more trees and combines their output and so requires more computational power and resources, and takes a longer time to train the model.



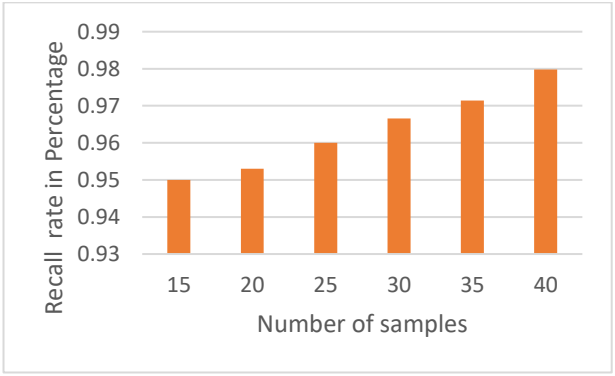
Plot 5. Accuracy of the RF ensemble learning method for different data sample sizes.



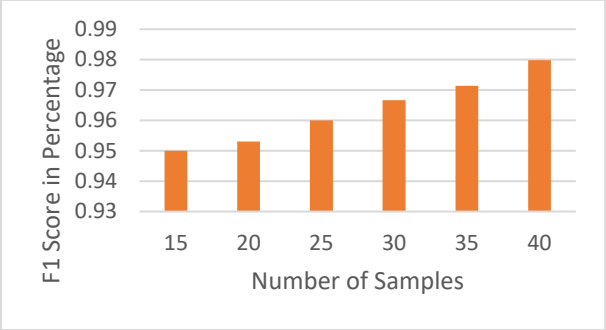
Plot 7. The precision of the RF ensemble learning method for various data sample sizes.



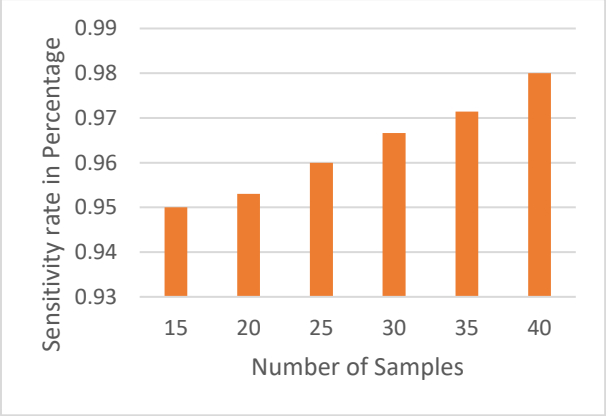
Plot 6. The error rate of the RF ensemble learning method for various data sample sizes.



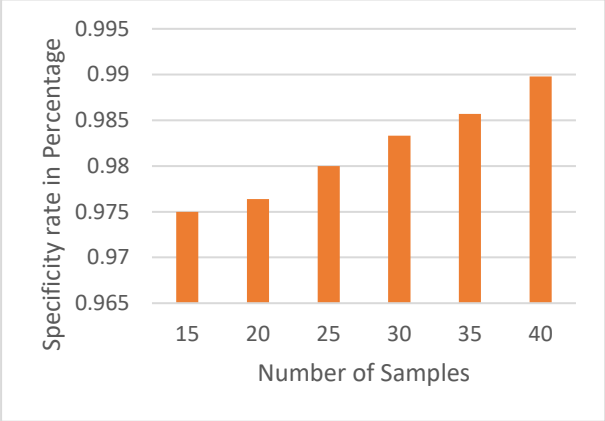
Plot 8. The recall of the RF ensemble learning method for various data sample sizes.



Plot 9. The F1 score of the RF ensemble learning method for various data sample sizes.



Plot 10. The sensitivity of the RF ensemble learning method for various data sample sizes.



Plot 11. The specificity of the RF ensemble learning method for various data sample sizes.

Plot 5 to plot 11 shows the relationship between different data sample sizes and various performance metrics for RF ensemble machine learning algorithms. Plot 6 shows that as data sample size increases error rates are inversely proportional. Plot 5, plot 7- plot 11 show that the performance measure accuracy, precision, recall, F1 score, sensitivity, and specificity are directly proportional to data sample size. Plot 2, plot 3, and Plot 4 illustrate the ROC curves of XGBoost, LGBM, and RF. The accuracy of all three ensemble machine learning algorithms is the same. When we compare individual class performance, the Random forest gives the best.

4. Conclusion

An efficient approach to detecting breast tumours at an early stage is screening mammograms. It is difficult for the radiologist to analyse the mammograms due to the presence of artefacts, noise, and pectoral muscles. This work effectively eliminated artefacts, noise, and pectoral muscles. This proposed GFEML technique used XGBoost, LGBM, and RF ensemble machine learning algorithms to categorise breast tissue as normal, benign, or malign. The experimental result shows that all three algorithms have the same accuracy. When comparing the individual class performance of three machine learning algorithms, the RF is the best.

References

[1] Yan Wang , Zizhou Wang, Yangqin Feng , and Lei Zhang, "WDCCNet: Weighted Double-Classifier Constraint Neural Network for Mammographic Image Classification" IEEE Transactions on Medical Imaging, vol. 41, no. 3, March 2022.

[2] Md. Akhlaqur Rahman, and Rajib Kumar Jha , "Multidirectional Gabor Filter-Based Approach for Pectoral Muscle Boundary Detection" IEEE Transactions on radiation and plasma medical sciences, vol. 6, no. 4, April 2022.

[3] M. Broeders *et al.*, "The impact of mammographic screening on breast cancer mortality in Europe: A review of observational studies," *J. Med. Screening*, vol. 19, no. 1, pp. 14–25, Sep. 2012.

[4] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic imageclassification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2246–2255, Jun. 2020.

[5] L. Wang, L. Zhang, M. Zhu, X. Qi, and Z. Yi, "Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101665.

[6] M. Mustra, M. Grgic, and R. M. Rangayyan, "Review of recent advances in segmentation of the breast boundary and the pectoral muscle in mammograms,"*Med. Biol. Eng. Comput.*, vol. 54, no. 7, pp. 1003–1024, 2016.

[7] M. A. Rahman, R. K. Jha, and A. K. Gupta, "Gabor phase response based scheme for accurate pectoral muscle boundary detection," *IET Image Process.*, vol. 13, no. 5, pp. 771–778, 2019.

[8] R. Gupta and P. E. Undrill, "The use of texture analysis to delineate suspicious masses in mammography," *Phys. Med. Biol.*, vol. 40, no. 5, pp. 835–855, 1995.

[9] P. K. Saha, J. K. Udupa, E. F. Conant, D. P. Chakraborty, and D. Sullivan, “Breast tissue density quantification via digitized mammograms,” *IEEE Trans. Med. Imag.*, vol. 20, no. 8, pp. 792–803, Aug. 2001.

[10] R. B. Gunderman, Essential Radiology. Clinical Presentation, Pathophysiology, Imaging, 2ed. New York, NY, USA: Georg Thieme 2006.

[11] S. M. Kwok, R. Chandrasekhar, Y. Attikiouzel, and M. T. Rickard, “Automatic pectoral muscle segmentation on mediolateral oblique view mammograms,” *IEEE Trans. Med. Imag.*, vol. 23, no. 9, pp. 1129–1140, Sep. 2004.

[12] R. J. Ferrari, R. M. Rangayyan, J. E. L. Desautels, R. A. Borges, and A. F. Frere, “Automatic identification of the pectoral muscle in mammograms,” *IEEE Trans. Med. Imag.*, vol. 23, no. 2, pp. 232–245, Feb. 2004.

[13] V. B. Bora, A. G. Kothari, and A. G. Keskar, “Robust automatic pectoral muscle segmentation from mammograms using texture gradient and euclidean distance regression,” *J. Digit. Imag.*, vol. 29, no. 1, pp. 115–125, 2016.

[14] P. Shi, J. Zhong, A. Rampun, and H. Wang, “A hierarchical pipeline for breast boundary segmentation and calcification detection in mammograms,” *Comput. Biol. Med.*, vol. 96, pp. 178–188, May 2018.

[15] A. Rampun *et al.*, “Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network,”*Med. Image Anal.*, vol. 57, pp. 1–17, Oct. 2019.

[16] W. Cheung and G. Hamarneh, “N-SIFT: N-dimensional scale invariant feature transform for matching medical images,” in *Proc. 4th IEEE Int. Symp. Biomed. Imag., Nano Macro*, Apr. 2007, pp. 720–723.

[17] R. Hupse and N. Karssemeijer, “Use of normal tissue context in computer-aided detection of masses in mammograms,” *IEEE Trans. Med. Imag.*, vol. 28, no. 12, pp. 2033–2041, Aug. 2009.

[18] G. Carneiro, J. Nascimento, and A. P. Bradley, “Unregistered multiview mammogram analysis with pre-trained deep learning models,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015, pp. 652–660.

[19] K. Sankar and K. Nirmala, “An enhanced mammogram diagnosis using shift-invariant transform,” *ICTACT J. Image Video Process.*, vol. 5, no. 2, pp. 920–925, Nov. 2014.

[20] S. A. Taghanaki, Y. Liu, B. Miles, and G. Hamarneh, “Geometry-based pectoral muscle segmentation from MLO mammogram views,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2662–2671, Nov. 2017.

[21] A. G. Gale, "The Mammographic Image Analysis Society digital mammogram database," *Proceedings of the 2nd International Workshop on Digital Mammography*, vol. 1069, pp. 375--378, 1994.

[22] I. A. Inês C Moreira, Inês Domingues, António Cardoso, Maria JoãoCardoso, and J. S. Cardoso, "INbreast: toward a full-field digital mammographic database," *Acad Radiol*, vol. 19, no. 2, pp. 236-48, Feb 2012.

[23] Sagenela Vijaya Kumar, C Nagaraju “Support vector neural network based fuzzy hybrid filter for impulse noise identification and removal from gray-scale image” Journal of King Saud University - Computer and Information Sciences, Volume 32, Issue 10, December 2020, Pages 1210-1211.

[24] C Naga Raju, A Hima Bindhu “Primary Screening Technique for Detecting Breast Cancer” i-manager's Journal on Image Processing; **Nagercoil** Vol. 6, Iss. 2, (Apr/Jun 2019): 21-27.