

Modified Differential Box Counting in Breast Masses for Bioinformatics Applications

S. Sathiya Devi¹ and S. Vidivelli^{2,*}

¹University College of Engineering-BIT Campus, Trichy, India

²Anna University, Chennai, Tamilnadu, India

*Corresponding Author: S. Vidivelli. Email: vidieng@gmail.com

Received: 01 May 2021; Accepted: 17 June 2021

Abstract: Breast cancer is one of the common invasive cancers and stands at second position for death after lung cancer. The present research work is useful in image processing for characterizing shape and gray-scale complexity. The proposed Modified Differential Box Counting (MDBC) extract Fractal features such as Fractal Dimension (FD), Lacunarity, and Succolarity for shape characterization. In traditional DBC method, the unreasonable results obtained when FD is computed for tumour regions with the same roughness of intensity surface but different gray-levels. The problem is overcome by the proposed MDBC method that uses box over counting and under counting that covers the whole image with required scale. In MDBC method, the suitable box size selection and Under Counting Shifting rule computation handles over counting problem. An advantage of the model is that the proposed MDBC work with recently developed methods showed that our method outperforms automatic detection and classification. The extracted features are fed to K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) categorizes the mammograms into normal, benign, and malignant. The method is tested on mini MIAS datasets yields good results with improved accuracy of 93%, whereas the existing FD, GLCM, Texture and Shape feature method has 91% accuracy.

Keywords: Breast cancer; computer-aided diagnosis; K-nearest neighbour; mammograms; modified differential box counting; support vector machine

1 Introduction

About 1 in 28 women are expected to develop breast cancer during their lifetime. By 2030, breast cancer will cause most deaths among women in India than any other cancers. The survival rate of breast cancer is low because the detection takes place late. Early detection can not only improve the outcome but can remarkably cut down the costs of treatment [1]. Detection of breast cancer at an early stage is essential to reduce the mortality rate. Therefore, a breast cancer screening method is needed to facilitate early diagnosis of this potentially fatal disease [2]. Breast cancer medical imaging can be used to look inside the human body as a non-invasive method for helping



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

doctors for diagnose and treat. An early breast cancer diagnosis occurs with any of the available imaging methods, it cannot be confirmed that these images are malignant alone. There is a high risk of cancer cells being placed in the interstitial tissue veins or fluid until the microscopic exam of tissues from cancer to confirm their malignancy begins. Mammography related to clinical and self-breast examination is the practical and effective method for mass screening to identify breast cancer. It appears in women in the form of tumors [3]. Since mammograms are medical images, fractal geometry is an appropriate method to identify texture features in the mass region [4]. For normal tissues, dense breasts have intensities for the same to those in cancer regions and tumor regions must be successfully identified. Many imaging techniques have been developed for early detection and treatment of breast cancer reduce the number of deaths and many aided breast cancer diagnosis methods have been used to increase the diagnostic accuracy. Thus, in order to detect the mammograms using CAD techniques accurately and are used for determining the breast cancer which is the motivation of the research.

Basically, a CAD system plays a crucial role in early detection than the other methods like a biopsy. Though the CAD systems have desirable properties it poses inherent challenges. The main issue identified is, it is semi-automatic and needs expert radiologist for mass region identification [5]. On the other case when sudden grey scale variation at the borders of neighbouring box leads to box under counting situation. In this framework a modified DBC approach is proposed and is described in next section [6–8]. To address this challenge and reduce experts' overhead, we propose automatic mass region extraction using the maximum entropy principle and shape feature (Circularity). Since mass regions are identified successfully, the discriminate features are essential for efficient classification of a mammogram. The intermediate pixel value is considered with maximum and minimum grey level of box for the computation [9–11]. In this manner, the neighbor value of the pixel is effectively analyzed for suitable box size selection. Generally, the mammogram images are textural in nature, specifically the mass region [12,13]. And to extract this texture feature the methods like Local Binary Pattern (LBP), Co-occurrence matrix, shape, contours, and fractals have been considered in various existing methods for breast cancer classification were based on Active contour, Active region segmentation, Shape and texture feature, Hybrid method, Fractal method, SURF, Colour features, CNN, and Differential Box counting [14,15]. The traditional DBC model applying the same roughness of intensity surface with various gray-levels creates over counting and under counting problem which is a research gap. These works influence the proposed framework for automatic ROI extraction.

This paper extracts fractal features and classifies the mammogram with KNN and SVM classifiers. The rest of our paper is structured as follows. Section 2 describes about the existing methods involved for breast mass detection and Section 3 outlines the proposed MDBC framework. Section 4 discusses about the simulation results and discussions. Section 5 presents the conclusion and future works for the present research.

2 Related Work

In order to identify breast cancer, radiologist mostly depends on CAD screening setup. Few remarkable works are discussed in the kinds of literature and some of them are presented below.

Kaur et al. [16] applied k-means clustering for Speed-Up Robust Features (SURF) selection and Multiclass SVM with deep learning method is applied for classification. The mini-MIAS dataset was used to evaluate the performance of the proposed model. The developed model has the overfitting problem in the training data. Agnes et al. [17] develop Multiscale All CNN (MA-CNN) model for detection of breast cancer in the medical breast images. The multiscale

filter is applied to fuse the wider context of information to improve the classification of system accuracy. The mini-MIAS dataset was used to evaluate the performance of the MA-CNN model. The developed MA-CNN model has the overfitting problem that affects the performance of the method.

Rabidas et al. [18] applied Local Photometric Attributes (LPA) method to analysis the local information in the medical images. The mini-MIAS dataset was used to evaluate the performance of the developed method. The analysis shows that the developed model has the higher performance compared to existing models. The proposed model has the lower efficiency in the feature analysis. Ghasemzadeh et al. [19] developed a deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. The developed deep learning method had higher accuracy in detecting breast cancer mass due to effective feature analysis and increases the patient survival rate. The developed algorithm was too weak to classify the images and resulted in low margins and overfitting problem.

Dhahri et al. [20] developed an infrared high classification accuracy hand held machine learning based method for breast cancer detection. The developed method showed effective performance in terms of sensitivity and specificity for the detection of breast cancer. The computational complexity of the developed method was more and consumed more time in classification. Wang et al. [21] diagnosed breast cancer Using an Efficient CAD System Based on Multiple Classifiers. First, the mammogram images were enhanced to increase the contrast. Second, the pectoral muscle was eliminated and the breast was suppressed from the mammogram. Next, k-nearest neighbor (k-NN) and decision trees classifiers were used to classify the normal and abnormal lesions. However, the developed CAD system could be considered as a powerful tool to detect and classify abnormalities in the breast

Indra et al. [22] developed a dual mode deep transfer learning system for breast cancer detection by using contrast enhanced digital mammograms. The developed model used deep transfer learning method effectively that classified the benign and malignant tumors using deep transfer learning system. However, the optimization problem used to generate the reconstructed graphs and rigorous criteria for evaluating the graphs was the limitations of visualization approach. Pezeshki et al. [23] developed Texture Analysis of Gradient Images for Benign-Malignant Mass Classification. In addition to the local texture feature, Local Binary Pattern, approximation coefficients have been extracted from the gradient images using wavelet transform to evaluate their efficiency in a Computer-Aided Diagnosis (CADx) system. However, other texture features along with different classifiers can be incorporated in future which may enhance the efficiency of the system.

In this section, the works of literature related to ROI segmentation and feature extraction methods under the fractal domain are discussed separately with their advantages and disadvantages. Since most of the work discussed above is semi-automatic, this paper discussed fractal features computation and extraction. It concluded that the fractal dimension is one of the prominent features of fractals. The next section discusses the proposed framework MDBC work with recently developed methods showed the model performs automatic detection and classification. The proposed model has the advantages of applying the suitable box selection to improve the performance

3 Proposed Approach

The proposed automated CAD system framework comprises of four phases (i) Pre-processing, (ii) Automatic mass region Identification and Extraction, (iii) Feature extraction using the

proposed MDBC which uses box over counting and undercounting that covers the whole image with required scale. (iv) Classification is performed for the extracted features that are fed to K-Nearest Neighbour (KNN) classifier and Support Vector Machine (SVM) which categorize the mammograms into normal, benign, and malignant. The diagrammatic representation of the framework is shown in Fig. 1.

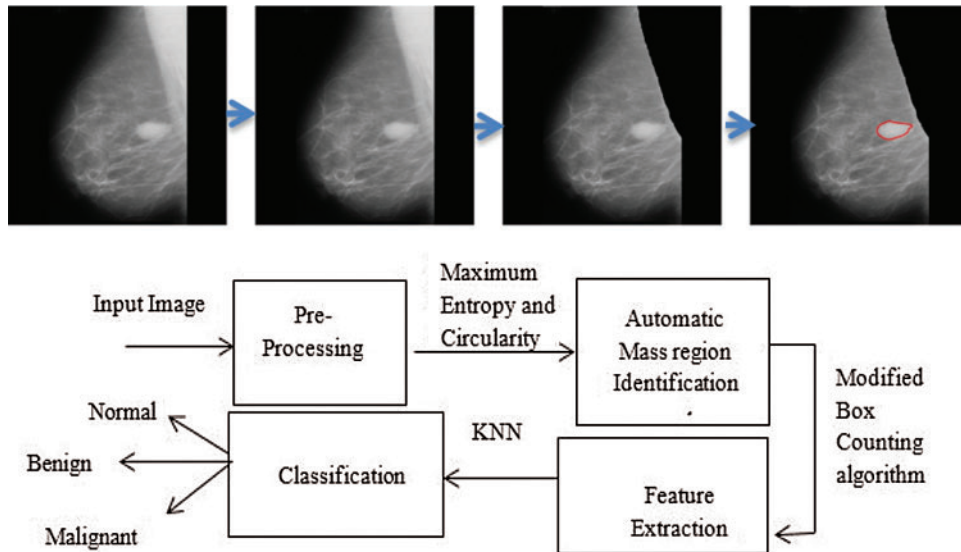


Figure 1: The proposed framework for automated CAD system

3.1 Pre-Processing

The mammogram images are always noisy and it contains artifacts and labels which affects the results of classification. Pre-processing is the initial step of automatic analysis of mammograms. It involves the segmentation of the breast region and removing pectoral muscles that can minimize the search area for abnormalities and make it limited to the relevant region of the breast without excessive influence from the mammogram's background. Pectoral muscle appears as a triangular opacity across the upper posterior margin of the image and the pectoral muscle can bias and affect the result of any mammogram processing system, so it is necessary to identify and segment the pectoral muscle automatically. The present research performs two stages of the pre-processing methods such as Breast Region segmentation and Removal of Pectoral muscle. The median filter and morphological operation is applied to remove the speckle noise in the medical image.

3.2 Mass Region Identification and Segmentation

After pre-processing, the automatic ROI segmentation is to be performed. For the dense breast tissues, the mass region frequently merges with them, and it makes the CAD system complex to locate and segment the masses accurately. So that most of the CAD systems are designed for the selection of mass regions with radiologist. Hence experienced and trained persons are required in the successful operation of CAD systems.

To address the issue, the proposed MDBC method is performed that comprises of two steps (i) Binarization and (ii) ROI extraction. For effective binarization, the maximum entropy principle is applied and hence the suspected regions are identified in this step. The circularity value is

computed for all suspected regions, which will help differentiate the mass region from other tissue regions.

3.3 Binarisation

The detection of the complex mass segment region in mammogram images because complicated tissues and ambiguous shape margin surround them. In order to differentiate normal over-complicated tissues, the foreground and background objects are to be separated. Binarization is the process of separating the mass region from the surrounding tissues based on threshold value. A threshold can be calculated by global or adaptive methods [24]. Generally, adaptive thresholds are preferable for mammogram images. The proposed framework computes the optimal adaptive threshold with maximum entropy principle model.

3.4 Maximum Entropy Principle

In information theory, entropy is used to measure the amount of information [25]. In the proposed framework the partition of the mass region from the background tissue is extracted with entropy based on gray distribution. Suppose a random variable of discrete type x with possible outcomes $\{x_1, \dots, x_n\}$ is assumed, n is the number of gray level and then $P(x_k)$ be the probability of the outcome X_k where k ranges from n to $k - 1$ and the entropy is defined in Eq. (1).

$$H(x) = - \sum_{k=1}^n p(x_k) \log p[x_k] \tag{1}$$

Let us consider the task to partition input image into mass (A) and background tissues (B) and the probability distribution of grey level in the given input images are $\{p_0, p_1, \dots, p_n\}$.

Then the Probability Distributions of mass (p_A) and background (p_B) are given in Eqs. (2), (3)

$$p_A = \frac{p_0}{P_A}, \frac{p_1}{P_A}, \dots, \frac{p_s}{P_A}$$

$$p_B = \frac{p_{s+1}}{P_B}, \frac{p_{s+2}}{P_B}, \dots, \frac{p_n}{P_B} \tag{2}$$

where

$$P_A = \sum_{i=0}^s p_i, \quad P_B = \sum_{i=s+1}^n p_i \tag{3}$$

To obtain the optimum threshold values the total entropy has to be maximized as in Eq. (4).

$$s = \arg(\max(H(A) + H(B))) \tag{4}$$

where the entropy $H(A)$ and $H(B)$ can be calculated by the Eqs. (5) and (6)

$$H(A) = - \sum_{i=0}^s \frac{p_i}{P_A} \log \frac{p_i}{P_A} \tag{5}$$

$$H(B) = - \sum_{i=s+1}^{n-1} \frac{p_i}{P_B} \log \frac{p_i}{P_B} \tag{6}$$

Now using this optimal threshold s the image is Binarized and the pixels higher than this s value are considered to be a mass region. As improving the entropy in the method, the gray level complexity of the images are decreases.

These mass regions are having holes and are connected with nearby objects that can be corrected by applying morphological operations opening and closing.

3.5 ROI Extraction Using Circularity

Generally, mammograms are characterized by circular, lobulated or speculated shape [26]. As many regions of different shapes are extracted in a previous step, but not all regions are masses. To detect the mass region from normal one the proposed framework applies the circularity method as one of the shape features. This method considers two parameters such as circumference and area of the mass to identify the circularity as defined in the Eq. (7).

$$C_r = \frac{C^2}{4N\pi} \quad (7)$$

where C_r = circularity, C = circumference, and N = Area.

From the extracted region the total number of 1's is considered as area. And the parameter circumference can be calculated by summing the total number of 1's in the boundary region. Based on the C_r value, the shape of the mass is identified. The methods like Hough transform and Template matching failed to identify the mass region when the size varies. But this Eq. (7) is derived so that the mass region of any size can be identified from normal because the circularity value of blood vessel region must be higher than the mass region.

3.6 Feature Extraction

In the previous section, the ROI is extracted and the features in the ROI are extracted in this section. As described in Sections 1 and 2 the mass regions are of texture in nature. Basically texture is described by fractal geometry using three aspects: (i) FD, (ii) Lacunarity and (iii) Succolarity. FD is one of the features which groups self-similarity and roughness in medical images. Lacunarity is another fractal feature that measures gap distribution in mammograms and this feature is useful in representing the inner structure of the tumor. The two features like FD and Lacunarity is more studied and well used in mammogram image analysis [27], where Succolarity has not been considered widely. This succolarity is one of the fractal features which are used to discriminate images with flow information allied with it. Hence this framework extracts the fractal features such as FD, Lacunarity and Succolarity and the method to compute FD in modified way is described in next section.

3.7 Fractal Dimension (FD)

Traditionally FD is computed by using methods wiz, (i) Ruler, (ii) Blanket, (iii) Box counting, (iv) Differential Box counting, (v) Triangle prism surface area, and (vi) Power spectral analysis. Among these methods, Box counting is a simple and frequently used method in the estimation of FD. Let us consider a bounded set A in Euclidean space and the FD of A can be estimated by the Eq. (8).

$$FD = \lim_{r \rightarrow 0} \frac{\log(N_r)}{\log\left(\frac{1}{r}\right)} \quad (8)$$

where $FD =$ Fractal dimension,

$N_r =$ Number of Copies of A, Scaled down by ratio ‘r’

$r =$ Scaled ratio of A

Consider a gray scale image of size $(M \times M)$ pixels in a 3D spatial surface (x, y, z) , where (x, y) denote an image plane and z denotes grey level pixels. The (x, y) plane is divided into grids of size $(s \times s)$ where $M/2 \geq s > 1$ and s is an integer. Let an estimation of $r = s/M$ each grid is a column of boxes of size $(s \times s \times s)$, where s the height of each box is computed by the Eq. (9) and G the total number of grey levels.

$$\left[\frac{G}{s'} \right] = \left[\frac{M}{s} \right] \tag{9}$$

The number of box count n_r covering on each grid is counted in Eq. (10).

$$n_r(i, j) = 1 - k + 1 \tag{10}$$

where 1 is Maximum grey level intensities, and k is Minimum grey level intensities. The 1 and k can be calculated as in Eqs. (11) and (12).

$$l = \frac{I_{max}}{s} \tag{11}$$

$$k = \frac{I_{min}}{s'} \tag{12}$$

The total number of boxes of $(M \times M)$ the image is computed with Eq. (13).

$$N_r = \sum_{i,j} n_r(i, j) \tag{13}$$

Then the FD (D) of grey scale image is estimated using Eq. (8) by substituting value obtained from Eq. (13).

Then the FD of an image or the slope of a line is computed by fitting all the points $(1/r, Nr)$ using Linear Least squares.

3.8 Proposed Modified DBC (MDBC)

In the MDBC method the over counting problem is encountered by Selection of suitable box size and Modified way of $n_r(i, j)$ computation for Under Counting Shifting rules are formulated. The proposed MDBC derives two assumptions, firstly increases the box-count precision based on the unequal triangle box partition. Secondly, the weights of the box count the size of triangle box partition proportions and based on the assumptions, squat box in each of the grid are divided to 4 asymmetric triangle box patterns. Each of the patterns will calculate the counts of boxes using box-counting technique. Maximum is the number of box counts better will be the estimation. The MDBC follows the Under Counting Shifting rules that outperforms in terms of fitting error that are as follows:

i) Selection of Suitable Box Size

Selecting box size is also an important issue in DBC method because if M cannot be appropriately partitioned by s then zero will be taken as values in that partition this may affect

accuracy of the method. So while choosing the box size the divisor of M can be used. Example for image of size (256×256) , the box sizes must be 2, 4, 8, 16, 32, 64 and 128 and the proper partitioning will increase the accuracy.

ii) Modified Way of $n_r(i,j)$ Computation

As $n_r(i,j)$ computed in the traditional method using Eqs. (10) and (11) considered only the maximum and minimum grey level of box, the importance of intermediate pixel is omitted. This may affect the accuracy of the system and cause an over-counting problem. To avoid this situation, in our work the average values in the $(i,j)^{th}$ block is calculated as I_{avg} . The minimum and maximum values are computed as I_{nmin} and I_{nmax} . This method reduces the box count and in turn, increases the accuracy.

If there are grey scaled variations at neighbouring boxes' borders, then the undercounting of boxes may occur at z direction. So that shifting of boxes along x and y direction and finding the maximum n_r value into consideration will improve accuracy of method and avoid undercounting of boxes.

While finding n_r value the boxes of size $(s \times s)$ is shifted along (x,y) plane with α pixels and then find the new n_r value (new_n_r) and compare it with the n_r value obtained without shifting (old_n_r) and select the maximum of two as in Eq. (14).

$$N_r = \max(new_n_r, old_n_r) \quad (14)$$

3.9 Shifting Rules

- (1) Read the image from top left starting pixel and shift it along (x,y) plane as $(i + \alpha, j + \alpha)$
- (2) To avoid the box out of the image at the end-use the below three conditions

if $i < M$ and $j = N$ then $(i,j) \rightarrow (i + \alpha, j)$

if $i = M$ and $j < N$ then $(i,j) \rightarrow (i, j + \alpha)$

if $i = M$ and $j = N$ then $(i,j) \rightarrow (i - \alpha, j - \alpha)$

This method is used to catch the borders of neighboring boxes so that the undercounting problem can be avoided. Here the value for α is taken as 1, because an enormous value α will result from inappropriate FD values.

In traditional DBC method the unreasonable results may obtained when FD is computed for tumour regions with the same roughness (tumour) of intensity surface but different gray-levels. But this can be avoided using our proposed MDDBC method because this method overcomes the problem of box over counting and undercounting. Although this FD is a significant feature, it yields better results when combined with Lacunarity.

3.10 Lacunarity

Lacunarity [28] is the counterpart to the FD that describes the texture of a fractal. The higher lacunarity indicates that the area is more heterogeneous nature. It is defined as the ratio of the variance over the mean value of the function and shown in Eq. (15).

$$L_r = \frac{\sum_M N_r^2 Q(N_r, s)}{[\sum_M N_r Q(N_r, s)]^2} \quad (15)$$

where, M is the sizes of the FD processed image.

$Q(N_r, s)$ is probability of N in box size s ,

L_r is the lacunarity of box size s N_r and is computed using MDBC method.

Lacunarity unambiguously characterize the spatial organization of the tumor region and the feature yields good results when combined with other fractal features with improved accuracy.

Succolarity

The fractal feature in researches pay less attention in mammogram image analysis but having a wide area of application in texture analysis [29]. A Succolarity is defined as an estimation of the degree of filaments that allow percolation. Before implementing the algorithm, our input grey scale image is converted into binary image as described in Section 3.2. The algorithm is as follows,

Algorithm

1. Consider all black pixels as empty and white as obstacles.
2. Divide the image into equal box size $BS(k)$ where k is the divisor of image size like box-counting algorithm.
3. Compute occupation percentage $OP(BS(k))$ of each box and pressure above the centroid of box to evaluate Eq. (16).

$$\sum_{k=1}^n OP(BS(k)) \times PR(BS(k), pc) \tag{16}$$

4. To make succolarity value dimensionless like FD and Lacunarity divide (17) by large possible value as in Eq. (17).

$$\sigma(BS(k), direction) = \frac{\sum_{k=1}^n OP(BS(k)) \times PR(BS(k), pc)}{\sum_{k=1}^n PR(BS(k), pc)} \tag{17}$$

This feature evaluates the percolation capacity of fluid in the tumour region at all four directions. The region with benign type tumour has smooth contour but it occupies a major region so that the percolation capacity is low compared with irregular rough contours malignant region. Though the fractal features FD and lacunarity measure the inner complexity and roughness of the tumour effectively, this succolarity characterize the nature of the tumour in terms of roughness of contours. The combination of these three features shows the improved result when combined with an effective classifier.

3.11 Classification of Breast Cancer

The one major advantage of the SVM is the use of convex quadratic programming, which provides only global minima hence avoid being trapped in local minima. The binary classification is performed using the below Eq. (18).

$$\{x_i, y_i\}, \quad i = 1, \dots, l, \quad y_i \in \{-1, 1\}, \quad x_i \in R_d \tag{18}$$

From the equations, x_i are known as the data points and y_i corresponds to the labels. The labels present in the hyper plane separate the data using the hyper plane equation $W_T x + b = 0$

where, w is known as the d-dimensional coefficient vector that is normal with the hyper plane and the value b is known as the offset from the origin. Based on the optimal separable margin, the optimization problem is solved by using the Eq. (18).

The present research work uses K-Nearest Neighbors (KNN) algorithm that does not require a learning phase. During the training phase, the distance function is used as a class choice function that works on the basis of classes in KNN model. The KNN considers the class which appears as one among them and is assigned to element neighbors that needs to be classified. The neighbors present are weighted based on the distance that separate to the new elements for the classification. The parameter K is used in the KNN Algorithm that chooses to assign the class for each new element which is calculated by using the Eq. (18).

In this step the k-NN [30] and SVM [31] classify the extracted features.

In the KNN algorithm, the testing set is identified by assigning it to the nearest point's class label in the training set. Euclidean distance metric is chosen to measure the distance between data points in KNN and it is given by Eq. (19).

$$d(x, y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \dots + (X_n - Y_n)^2} \quad (19)$$

From the equation, X, Y are the two points in Euclidean space, $(X_1 \dots X_n)$ and (Y_1, \dots, Y_n) are the Euclidean vectors starting from the space to the origin. The n is the n-space values.

SVM is broadly used in mammogram image classification, though it was designed to solve binary classification problem, it has been extended to multiclass classification problems. In this work, Radial Basis Function (RBF) kernel is used to perform mapping of data from input space to feature space. Decision intelligence design is about making decisions based on objective principles that may or may not apply. It's making the most objective decision you can with an understanding that in the end those decisions are all subjective. It produces good results with high accuracy and low error rate than polynomial kernel. The KNN is implemented in Weka and the libSVM library is used for SVM classifier that classifies the breast mass image as Benign, malignant or normal.

Pseudo Code for The Proposed Modified DBC

```

Input H with x samples, y lines, and m bands
Given window size M, grid size s, and total gray level G.
R = M/s, s' = G/r
for n = 0 to m - 1
  for k = 0 to s - 1
    for l = 0 to s - 1
      for j = 0 to (y + M - 1 - k)/s - 1
        for i = 0 to (x + M - 1 - l)/s - 1
           $g_u = \max(H[n, k + j * s : k + j * s + s - 1, l + i * s : l + i * s + s - 1])$ 
           $g_l = \min(H[n, k + j * s : k + j * s + s - 1, l + i * s : l + i * s + s - 1])$ 
           $B[i, j] = g_u / s' - g_l / s' + 1$ 
        endfor
      endfor
    endfor
  endfor
  for j = 0 to (y - 1 - k)/s - 1
    for i = 0 to (x - 1 - l)/s - 1

```

```

    N[n,j, i] = sum(B[j: j + r - 1, i: i + r - 1])
  end for
end for
end for
end for
end for
Output: Image N

```

4 Result and Discussion

The proposed methodology is simulated with respect to two benchmark dataset (i) Mini-MIAS [32] and (ii) INbreast [33]. CAD, or computer-aided design and drafting (CADD), is technology for design and technical documentation, which replaces manual drafting with an automated process. If you are a designer, drafter, architect or engineer, you have probably used 2D or 3D CAD programs such as Auto CAD or AutoCAD software. These widely used software programs help you draft construction documentation, explore design ideas, visualize concepts through photo realistic renderings and simulate how a design performs in the real world. System Requirement: The proposed simulations are conducted on an Intel (R) core (TM) i7 CPU 965@3.20 GHz system with 4.00 GB RAM.

4.1 Dataset

The mini-MIAS are grey scale images of size (1024×1024) , totally it contains 322 images from left and right breast of 161 patients. This is a reduced version of the original MIAS database reduced to 200-micron pixel edge. Among them, 266 images are considered for simulation, in that 207 are normal and 59 images contain masses of benign and malignant type. There are three characteristics of background tissues are present in the images such as fatty, fatty-glandular, and dense-glandular. Two types of severity are present in the dataset such as Benign and Malignant.

Another publicly available INbreast dataset is also used with total 82 images, in which 45 are benign cases, 8 are normal and 29 are malignant. Several types of lesions such as distortions, asymmetries, calcifications, and masses are present in the images. The Groundtruth of the lesions are present in the XML format. For conducting an experiment, an image of size (1024×1024) is taken from min-MIAS database and it is pre-processed as described in Section 3.1 and shown in Fig. 2a. Then the pre-processed image is shown in Fig. 2b, mass identification image is shown in Fig. 2c, and classified image is shown in Fig. 2d.

In this step, circularity values for suspected regions are computed using the Eq. (6).

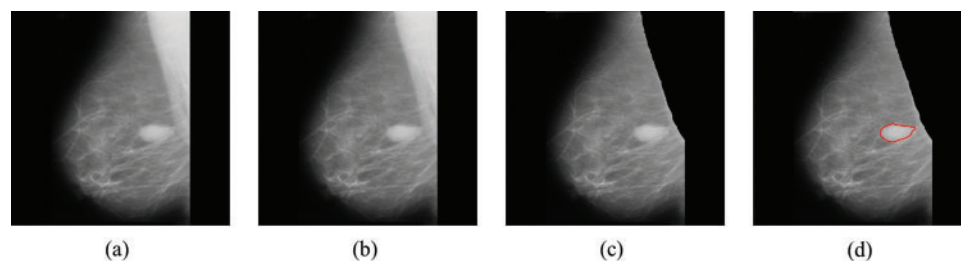


Figure 2: (a) Input images, (b) Pre-Processing, (c) Mass region identification, and (d) Classification images

The result shows that all the circular mass regions have value ranges from 0.9 to 1.5 and speculated mass regions from 1.5 to 2.0. This region is taken as a mask and overlapped on the original image for mass segmentation as in Fig. 2b, the same process is repeated for all images in the databases and ROI are extracted successfully. Then the performance of automatic mass segmentation is computed with Eq. (20).

$$X_A = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (20)$$

where X_A is Accuracy, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

In research work simulation the performance of the mini-MIAS is computed for 266 images. Among which it produces 58, 200, 5 and 3 images corresponding to TP, TN, FP and FN, respectively. Accuracy of the proposed technique is compared with existing methods and the images from INbreast are also tested and the results are shown in the Tab. 1. Although [5,6] produced good result compared to our method it is not proven to be a reliable method for varying mass size. All the template methods depend on the template's size is the major drawback in automatic mass segmentation. The method proposed by [7] is robust but complexity is high and also provides low result. But the proposed MDBC method uses box over counting and under counting that covers the whole image with required scale. MDBC selects a suitable sized box and Under Counting Shifting rule computation handles over counting problem. Thus, the proposed MDBC obtains 1% to 8% improvement in Accuracy values compared to existing models. Dataset insights are visualized graphically to describe the pros and cons of data and reaching the aim defined is shown in the Fig. 3. But our method yields satisfactory results irrespective of mass size. This method yields good results for both the mini-MIAS and INbreast dataset.

Table 1: Accuracy of ROI segmentation with proposed and other methods

Method	Accuracy (%)	Dataset
1 Mass template method [5]	90	Mini-MIAS
2 Entropy with template matching [6]	97.2	DDSM
3 Hough transform with snake model [7]	90.4	MIAS
4 Proposed MDBC method	97	Mini-MIAS
	98	INbreast

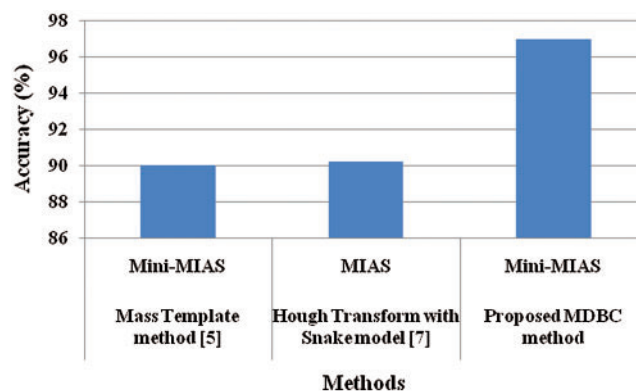


Figure 3: Graphical representation of the proposed and the existing models using MIAS dataset with respect to accuracy

4.2 Quantitative Analysis for the Proposed MDBC

The exact FD values are extracted as result with less computational time. Hence the simulation values of FD, Lacunarity and Succolarity obtained from MDBC method is shown in [Tab. 2](#). The succolarity computation evaluates the degree of percolation in the region of interest. For region with smooth contours, the information flow is very low because mass region inside the rectangle occupies more space. So that in our work, the benign region is having low succolarity value than the malignant one.

Table 2: Sample features for (a) Benign images, (b) Malignant images, (c) Normal images of Mini-MIAS and INbreast databases

Benign (Mini MIAS)				Benign (INbreast)			
ID	FD	Lac	Succ	ID	FD	Lac	Succ
mdb001	1.10	0.21	0.10	22579847	1.24	0.27	0.11
mdb002	1.17	0.22	0.10	20587758	1.23	0.23	0.15
mdb005	1.15	0.19	0.20	20588458	1.25	0.19	0.10
mdb010	1.15	0.22	0.21	20588654	1.18	0.21	0.04
mdb012	1.20	0.15	0.21	22427682	1.19	0.25	0.10
mdb091	1.20	0.18	0.10	22427728	1.29	0.22	0.05
mdb015	1.14	0.25	0.10	22580393	1.19	0.29	0.14
mdb017	1.20	0.24	0.21	22580706	1.22	0.19	0.08

Malignant (Mini MIAS)				Malignant (INbreast)			
ID	FD	Lac	Succ	ID	FD	Lac	Succ
mdb023	1.30	0.41	0.47	24055445	1.32	0.31	0.60
mdb028	1.28	0.31	0.42	20588046	1.34	0.39	0.30
mdb058	1.25	0.33	0.41	20587612	1.33	0.43	0.90
mdb072	1.50	0.32	0.51	20587664	1.29	0.34	0.40
mdb075	1.30	0.49	0.71	20587994	1.37	0.40	0.42
mdb202	1.39	0.31	0.30	20588190	1.31	0.41	0.41
mdb206	1.50	0.41	0.70	20588216	1.37	0.31	0.61
mdb190	1.40	0.50	0.40	20588536	1.47	0.40	0.53

Normal				
ID	Database	FD	Lac	Succ
mdb074	Mini-MIAS	0.90	0.09	0.80
mdb078		1.02	0.12	0.79
mdb082		1.12	0.11	1.06
mdb280		1.00	0.10	0.80
mdb322		1.09	0.09	0.90
20588138	INbreast	1.11	0.19	0.71

The Fig. 4 shows the region with the smooth contour of benign type and speculated mass of malignant type is flooded in all directions using Eq. (17). After pre-processing, the PSNR range of the image is 46 dB and effectively removes the noise. The change in the slightest PSNR range of the images doesn't much affects the performance of the model.











Directions	Benign	Malignant
Binary ROI		
Left to Right		
Right to Left		
Bottom to Top		
Top to Bottom		

Figure 4: Intermediate images from succolarity computation

This sample features in Tab. 2 shows that the mean value for benign category FD is 1.25, Lacunarity is 0.21 and Succolarity is 0.1 and for Malignant images the mean value for FD is 1.45, Lacunarity is 0.45 and Succolarity is 0.4. But for the normal image, value of FD is 1.01, Lacunarity is 0.1 and succolarity be 0.9. In Mini-MIAS database, the 5 normal images are misclassified in the previous phase, but the fractal features correctly classify it. Overall accuracy of the system is computed by Eq. (20).

For mini-MIAS database among 63 instances 27, 27 and 5 images from Benign, Malignant and Normal are correctly classified using KNN with $k = 1$ at 10-folds cross-validation. On substituting these values on Eq. (20) the accuracy of 93.6% is produced as result and its corresponding Receiver Operating Curve (ROC) is shown in Fig. 5 in that curve the x-axis indicates False Positive Rate and y-axis indicate True Positive Rate. On the SVM classifier with RBF kernel function

the result of 90.4% is achieved on 10-folds cross-validation. For INbreast database the KNN has produced 95% of accuracy and the model takes 0.08 s to build hence it shows the low computational cost in terms of time and SVM classifier yields 93% of result.

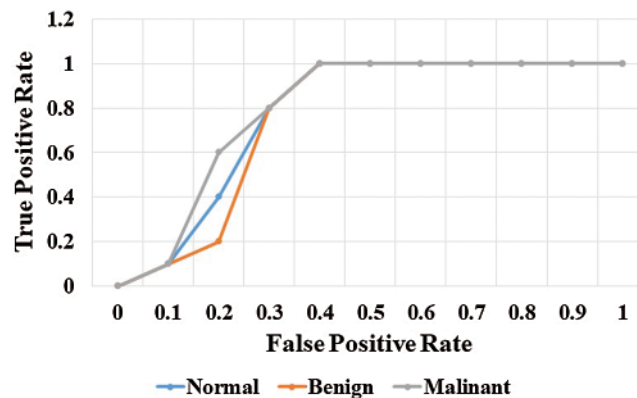


Figure 5: ROC for normal, benign and malignant

4.3 Comparative Analysis

The proposed MDBC result is compared with existing work and result analysis is shown in Tab. 3. Although [16,17] have achieved an excellent result, they have manually segmented the mass region. In [18] they have attained good result with Local Photometric Attributes method. In [34] they have extracted 34 features in various categories viz., Intensity, GLCM (texture), Shape, Texture and Margin. The proposed model applies suitable box sizes and Under Counting Shifting rule to effectively handle the over counting problem. Finally, they have obtained good result with SVM classifier. Our proposed automatic method has reached good result on both the databases. However, the proposed MDBC model failed to detect masses with blurred edges and ill-defined shapes, which impacted on the feature extraction step also. So combining some other contour detection techniques in future produce improved result.

Table 3: Classification accuracy of our proposed method compared with other methods

Sl. No.	Method	Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Multiscale—CNN [17]	Mini-MIAS	91	89	90
2	Deep learning and Multiclass SVM [16]	Mini-MIAS	92	91	90
3	Local photometric attributes [18]	Mini-MIAS	90	87	90
4	FD, GLCM, shape and texture [34]	Mini-MIAS	91	90	91
5	Proposed method	Mini-MIAS	93	94	93
		INbreast	95	94	95

The results obtained in Tab. 1 shows the efficiency of our proposed MDBC framework in automatic mass region detection. Then the feature extraction and classification accuracy of mammogram is discussed in Tab. 3. Comparing the proposed MDBC work with recently developed

methods showed that our method outperforms in automatic detection and classification. The proposed model has the advantages of applying the suitable box selection to improve the performance. The AUC of the proposed model for normal, Benign, and Malignant is shown in Fig. 5. The area under ROC of the proposed model for normal is 1, Benign is 0.89, and Malignant is 0.89.

5 Conclusion

In the present research, an automated CAD system framework without human interventions was proposed. Generally, the majority of the system depends on the radiologist to select the mass region which requires much computation time. Therefore, the proposed framework was used for automatic mass region identification and segmentation which proved to be better for all sizes of masses. Second part of the framework proposes MDBC for fractal feature computation, and this method solved the problem of the box over counting and undercounting simply. The extracted features are fed to K-Nearest Neighbour (KNN) classifier and the Support Vector Machine (SVM) categorizes the mammograms into normal, benign, and malignant. The KNN model has the higher efficiency when number of training data is more and SVM has the capacity to effectively analysis the features. The method is tested on mini MIAS datasets yields good results with improved accuracy of 93%, whereas the existing FD, GLCM, Texture and Shape feature method has 91% accuracy. Future work of this model involves in detecting micro-calcification in early stage of breast cancer that act as an expert system.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] Z. Momenimovahed and H. Salehiniya, "Epidemiological characteristics of and risk factors for breast cancer in the world," *Breast Cancer: Targets and Therapy*, vol. 11, pp. 151–164, 2019. <https://dx.doi.org/10.2147%2FBCTT.S176070>.
- [2] T. Berber, A. Alpkocak, P. Balci and O. Dicle, "Breast mass contour segmentation algorithm in digital mammograms," *Computer Methods and Programs in Biomedical*, vol. 110, no. 2, pp. 150–159, 2013.
- [3] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan and A. A. Basha, "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform," *Measurement*, vol. 146, no. 1, pp. 800–805, 2019.
- [4] H. Li, Y. Wang and K. J. R. Liu, "Computerized radiographic mass detection—Part II: Decision support by featured database visualization and modular neural networks," *IEEE Transaction of Medical Imaging*, vol. 20, no. 4, pp. 302–313, 2001.
- [5] S. Ozekes, O. Osman and A. Y. Camurcu, "Mammographic mass detection using a mass template," *Korean Journal of Radiology*, vol. 6, no. 4, pp. 221–228, 2005.
- [6] V. P. Singh and R. Srivastava, "Automated and effective content-based mammogram retrieval using wavelet based CS-LBP feature and self-organizing map," *Bio Cybernetics and Biomedical Engineering*, vol. 38, no. 1, pp. 90–105, 2018.
- [7] Y. Ma, X. Lu, M. Dong and K. Wang, "Automatic mass segmentation method in mammograms based on improved VFC snake model," in *Proc. Emerging Trends in Image Processing, Computer Vision, and Pattern Recognition*, pp. 201–217, 2015. <https://doi.org/10.1016/B978-0-12-802045-6.00013-2>.
- [8] T. Sadad, A. Munir, T. Saba and A. Hussain, "Fuzzy c-means and region growing based classification of tumor from mammograms using hybrid texture feature," *Journal of Computational Science*, vol. 29, no. 3, pp. 34–45, 2018.

- [9] A. Niaz, A. A. Memon, K. Rana, A. Joshi, S. Soomro *et al.*, “Inhomogeneous image segmentation using hybrid active contours model with application to breast tumor detection,” *IEEE Access*, vol. 8, pp. 186851–186861, 2020.
- [10] T. M. Cabral and R. M. Rangayyan, “Fractal analysis of breast masses in mammograms,” *Morgan & Claypool*, vol. 7, no. 2, pp. 1–18, 2012.
- [11] T. G. Nguyen, T. V. Phan, D. T. Hoang, T. N. Nguyen and C. So-In, “Efficient SDN-based traffic monitoring in IoT networks with double deep Q-network,” in *Proc. Int. Conf. on Computational Data and Social Networks*, Cham, Springer, pp. 26–38, 2020.
- [12] M. M. R. Krishnan, S. Banerjee, C. Chakraborty, C. Chakraborty and A. K. Ray, “Statistical analysis of mammographic features and its classification using support vector machine,” *Expert Systems with Applications*, vol. 37, no. 1, pp. 470–478, 2010.
- [13] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty *et al.*, “Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset,” *SN Computer Science*, vol. 2, no. 1, pp. 1–13, 2021.
- [14] C. Chakraborty, B. Gupta and S. K. Ghosh, “Chronic wound characterization using bayesian classifier under telemedicine framework,” *Proc. Medical Imaging: Concepts, Methodologies, Tools, and Applications*, vol. 7, no. 1, pp. 76–93, 2016.
- [15] A. Kishor, C. Chakraborty and W. Jeberson, “Reinforcement learning for medical information processing over heterogeneous networks,” *Multimedia Tools and Applications*, vol. 80, pp. 23983–24004, 2021.
- [16] P. Kaur, G. Singh and P. Kaur, “Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification,” *Informatics in Medicine Unlocked*, vol. 16, no. 1, pp. 100151, 2019.
- [17] S. A. Agnes, J. Anitha, S. I. A. Pandian and J. D. Peter, “Classification of mammogram images using multiscale all convolutional neural network (MA-CNN),” *Journal of Medical Systems*, vol. 44, no. 1, pp. 1–9, 2020.
- [18] R. Rabidas and W. Arif, “Characterization of mammographic masses based on local photometric attributes,” *Multimedia Tools and Applications*, vol. 79, no. 29–30, pp. 21967–21985, 2020.
- [19] A. Ghasemzadeh, S. S. Azad and E. Esmaeili, “Breast cancer detection based on gabor-wavelet transform and machine learning methods,” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 7, pp. 1603–1612, 2019.
- [20] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani and M. Faisal Nagi, “Automated breast cancer diagnosis based on machine learning algorithms,” *Journal of Healthcare Engineering*, vol. 2019, no. 12, pp. 1–11, 2019.
- [21] K. Wang, B. K. Patel, L. Wang, T. Wu, B. Zheng *et al.*, “A dual-mode deep transfer learning (D2TL) system for breast cancer detection using contrast enhanced digital mammograms,” *IISE Transactions on Healthcare Systems Engineering*, vol. 9, no. 4, pp. 357–370, 2019.
- [22] P. Indra and M. Manikandan, “Multilevel tetrolet transform based breast cancer classifier and diagnosis system for healthcare applications,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 1–10, 2020.
- [23] H. Pezeshki, M. Rastgarpour, A. Sharifi and S. Yazdani, “Extraction of speculated parts of mammogram tumors to improve the accuracy of classification,” *Multimedia Tools and Application*, vol. 78, no. 14, pp. 1–25, 2019.
- [24] H. Li, X. Meng, T. Wang, Y. Tang and Y. Yin, “Breast masses in mammography classification with local contour features,” *Bio-Medical and Engineering*, vol. 16, no. 2, pp. 16–44, 2017.
- [25] W. Tatsuaki and S. Talesjo, “When nonextensive entropy becomes extensive,” *Physica A*, vol. 301, no. 1–4, pp. 284–290, 2001.
- [26] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger *et al.*, “Inconsistent performance of deep learning models on mammogram classification,” *Journal of the American College of Radiology*, vol. 17, no. 6, pp. 796–803, 2020.

- [27] Q. Guo, J. Shao and V. F. Ruiz, "Characterization and classification of tumor lesions using computerized fractal based texture analysis and support vector machines in digital mammograms," *International Journal of Computer Assistant and Radiology Surgery*, vol. 4, no. 1, pp. 11–25, 2009.
- [28] R. E. Plotnick, "Lacunarity indices as a measure of landscape texture," *Landscape Ecology*, vol. 8, no. 3, pp. 201–211, 1993.
- [29] J. I. R. Cojocaru, D. Popescu and I. E. Nicolae, "Texture classification based on succolarity," *Telecommunications Forum*, pp. 498–501, 2013. <https://doi.org/10.1109/TELFOR.2013.6716275>.
- [30] R. Duda, P. Hart and D. Stork, *Pattern Classification*, 2nd ed., New York: John Wiley & Sons, 2001.
- [31] F. Girosi, M. Jones and T. Poqgio, "Regularization theory and neural network architectures," *Neural Computation Cambridge*, vol. 7, no. 2, pp. 217–269, 1995.
- [32] J. P. Suckling, "MIAS—the mammographic image analysis society digital mammogram database," *Exerptational Medication of International Congress Series*, vol. 1069, pp. 375–378, 1994.
- [33] [Online]. Available: <https://www.kaggle.com/ramanathansp20/inbreast-dataset>.
- [34] J. Zheng, D. Lin and Z. Gao, "Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis," *IEEE Access*, vol. 8, pp. 96946–96954, 2020.