IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 27, 2021, accepted June 25, 2021, date of publication July 5, 2021, date of current version July 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094768

Unsupervised Method to Localize Masses in Mammograms

SAJIDA IMRAN[1], BILAL AHMED LODHI[2], AND ALI ALZAHRANI[1]
[1]Department of Computer Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[2]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K.

Corresponding author: Sajida Imran (skamran@kfu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Faisal University (Nasher Track) under Grant 216116.

ABSTRACT Breast cancer is one of the most prevalent types of cancer that mainly affects the women population. chances of effective treatment increase with early diagnosis. Mammography is considered one of the effective and proven techniques for the early diagnosis of breast cancer. Tissues around masses look identical in a mammogram, which makes the automatic detection process a very challenging task; they are indistinguishable from the surrounding parenchyma. In this paper, we present an efficient and automated approach to segment masses in mammograms. The proposed method uses hierarchical clustering to isolate the salient area followed by extraction of features to reject false detection. We applied our method to two popular publicly available datasets (mini-MIAS and DDSM). A total of 56 images from the mini-mias database and 76 images from DDSM were randomly selected. Results are explained in terms of ROC (Receiver Operating Characteristics) curves and compared with other state-of-the-art techniques. Experimental results demonstrate the efficiency and advantages of the proposed system in automatic mass identification in mammograms.

INDEX TERMS Breast mass detection, automatic mammogram segmentation, mass classification.

I. INTRODUCTION

Breast cancer is the most prevailing source of cancer-related deaths among women across the globe. Yearly, there are approximately 450, 000 deaths, out of which, breast cancer accounts for about 14% of all female cancer deaths [24]. Recent statistics say that 1 out of 10 women is affected by breast cancer in their lifetime. According to GLOBOCAN 2012, 1.7 million women were diagnosed with breast cancer and there were 6.3 million women alive who had been diagnosed with breast cancer in the previous five years [5]. Although the breast cancer rate is increasing in many parts of the world, however, the mortality rate is much higher in less developed countries, because of insufficient facilities available for diagnosis and treatment. Therefore, there is an urgent need for reliable and affordable approaches for the early diagnosis and treatment of breast cancer in less developed countries. It can have a significant impact on cancer treatment, faster recovery, and reducing mortality.

Mammography is considered the most effective technique as it can detect 85∼90% percent of all breast cancers [5]. A mass is an uncontrolled grown tumor and we classify them into malignant and benign by their size, shape, and other features. As described earlier that early diagnosis is a key for effective treatment. Therefore, the job of the radiologist becomes very important, who can interpret mammograms for early diagnosis. A mammogram does not have so much information imprinted on the film, therefore, a cancer diagnosis in this scenario becomes a subjective criterion where radiologist opinion depends on their experience. [37] states that radiologist's diagnosis inter-observer variation rate is 65 ∼ 75%). He can miss a significant proportion of abnormalities and in addition, a large number of masses come out to be benign after biopsy [37]. Computer-aided diagnosis (CAD) systems are helpful for radiologists in diagnosis. Detection accuracy can be improved by combining the expert knowledge with CAD scheme [25].

Due to the seriousness of accurate detection of abnormalities in mammograms, a wide-scale research has been performed using CAD systems to detect the masses in mammograms [1], [7], [11], [38], [39]. However, the complex nature of masses; variability in shape, size, margin, and occlusion within dense breast tissue, the accurate detection of masses becomes a very difficult task for computational

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

approaches [13]. Furthermore, [46] highlighted a standardization issue in their comparative study on mass classification. They performed a comparative study with a focus on; different extraction methods of ROIs, the use of different machine learning techniques using different datasets, and the use of different prediction accuracies. It was noted that mostly the research is done using a single data set. However, in such a case, the prediction model may perform well in one data set but fails to give good accuracy when tested on another data set. It is a challenging task to develop a generalized approach that performs well on more than one dataset.

To address the above-mentioned challenges, We proposed an unsupervised learning-based method that effectively and accurately detects masses for breast cancer diagnosis. The developed approach prevents the uncontrollable growth of candidate mass regions and helps to detect the masses occluded within dense tissue, which was a major limitation of earlier works [23]–[26]. Also, the proposed method is generalized to apply on differently sized datasets as compared to most AI techniques that require a bulk of data to achieve generalization. Moreover, the use of novel hierarchical clustering with its distance measure is the salient feature of this work. In addition, the paper introduces a distance measure to merge clusters for the reduction of false positives (FPs). Lastly, the method has been validated with cross-data-set performance. The proposed scheme is novel in the following ways:

- The application of the proposed detection algorithm is wider as it can detect multiple types of masses irrespective of their shape and size. The proposed algorithm was also tested on many ill-defined masses.
- The proposed mass identification method can accurately detect masses irrespective of their size and shape.
- We proposed an efficient and unsupervised approach to detect masses in mammogram images that segments the breast region and finds the candidate regions of interest (ROIs).
- A distance measure is proposed to merge the clusters in hierarchical clustering which also helps in reducing the FPs.
- Generalization of the proposed algorithm is tested by experimenting with cross-validation across two different datasets.

The organization of the paper is as follows. Section I presents the introduction and significance of the work. Section II discusses previous and related work. Section III briefly describes the proposed method for pre-processing. Section IV analyses the results and finally, Section V concludes the article.

II. RELATED WORK

To develop computer-aided breast cancer detection tools, researchers have used several approaches. A Particle Swarm Optimized Wavelet Neural Network (PSOWNN) based classification approach for detection of masses in digital mammograms is proposed in [23]. Their method is based on extracting Laws Texture Energy Measures from the mammograms and classifies the suspicious regions by PSOWNN. However, their method does not have any noise removal algorithm and also, they do not propose any automatic method of ROI detection. In [41], [44], authors used Latent Dirichlet Allocation (LDA) to mine the feature set of mammogram images. They presented the modified Morphological Component Analysis method to identify the mass region and then extracted morphological features. Finally, LDA is used to classify the masses. Simple Morphological approaches are very sensitive to noise. They also did not present any pre-processing for the collection of ROIs.

In [36], authors proposed the modified Fuzzy c-means clustering to cluster the masses, extracted morphological, textual, and spatial features form those masses, and classified the features using SVM (Support Vector Machine). Their method lacks noise removal and intelligent ROI segmentation. To aid segmentation and detection of masses in mammograms, a set of tools is presented in [31]. In this work, de-noising is applied after the top-hat morphological operator. Furthermore, image gray-level was enhanced by wavelet transform and a Wiener filter. And finally, the segmentation method was employed using multiple thresholding, wavelet transform, and genetic algorithm. They used a manual process to reduce the false positives generated by genetic algorithm. However, the authors did not do the automatic classification of the ROIs. A method for mass detection based on a saliency map is proposed in [2]. After the creation of the saliency map, a threshold is used to obtain the ROI. Several features were extracted and classified by SVM. A good threshold selection in the algorithm is significant as it will affect the overall accuracy; a low threshold will result in lots of FPs, while a higher threshold will miss the low contrast or occluded masses. Automated detection of malignant masses in screening mammography has been discussed in [34]. They developed a technique that used the presence of concentric layers which surround a focal area in the breast region, that has suspicious morphological characteristics and low relative incidence. The segmentation process in both of the earlier described algorithms is focused on the bright or salient parts of the image, which is always misled by the blood vessels resulting in the whole breast parenchyma as an ROI. Work in [28] is based on applying a one-dimensional recursive median filter to the different number of angles to each pixel. However, detection is failed when the structure of the mass and normal glandular look similar. So, the algorithm can only be detected if there is asymmetry between the left and right breasts.

The method proposed by [29] is based on the ISO-intensity analysis of groups to segment the skeptical masses. Adaptive flow orientation features are extracted from the ribbon around the masses to reduce the false positives. The procedure is tested on 56 images from the mini-MIAS database and false positives are then removed using features based on flow orientation in adaptive ribbons of pixels across the margins of masses. The algorithm achieved an 81% sensitivity rate

with 2.2 false positives per image. Furthermore, based on gray-level co-occurrence matrices (GCM) and using features on a logistic regression method, the classification of masses was performed as benign or malignant using five texture features. An accuracy of 0.79 is achieved as a result of classification by their algorithm, with 19 benign and 13 malignant lesions. The authors used the hard thresholds to get the contours of objects in the image. The contour is very sensitive to noise that result in the increase of false positives and poor segmentation results. The algorithm will fail to detect the mass if the boundary is ill-defined or even the mammogram is denser. Work in [43] used Morphological operation and scaled Reyni entropy to detect the masses in mammograms. To detect the mass, the mammogram is first pre-processed, then enhancement and mass segmentation is applied, and lastly, the mass detection is performed. The original mammogram is first filtered out from any artifacts using a median filter where thresholding is applied to obtain several binary objects. Then separation of objects is applied through Morphological erosion and the size of the object is computed based on the area of the objects, and only the larger objects kept. In the end, through Morphological dilation, border smoothing is performed to get an artifact-free mammogram. Before mass detection, mass segmentation is performed using Reynie's entropy. The mass detection system provides better results for all types of mammograms, with 93.2% and 93.9% TPF at the rate of 1.48 and 0.74 FP/Is for MIAS and DDSM datasets.

A wavelet-based breast lesions diagnosis is proposed in [14]. To detect the masses and reduce the false positive rate, multi-resolution features are extracted using a wavelet transform which serves as input to a binary tree classifier. The algorithm achieved 91.9% true positive detection accuracy. ROIs were manually cropped in the proposed system that is based on wavelet and curvelet coefficients, which are very high in numbers. Selecting the best coefficients is an optimization problem and also it is very sensitive to noise. A method that combines several artificial intelligence techniques with the discrete wavelet transform (DWT) is proposed in [47]. ROI's are determined through dimensional analysis using a multi-resolution Markov random field algorithm, the segmentation is performed that leads to the application of tree type classification strategy. When tested with Mini-MIAS data-set, the algorithm achieved a sensitivity of 97.3% with 3.9 false positives per image. Their proposed method works well with well-defined masses, but ill-defined masses are difficult to be classified by this method.

Recently, researchers are focusing on use of deep learning based approaches with a special focus on manually segmented regions of interest (ROIs) to detect mammogram masses [1], [11], [38], [45]. The medical domain requires high-performance algorithms and these algorithms require a bulk of data, which is a challenging task in medical imaging. Training such models in a low-data regime highly increase the risk of over-fitting. Although few-shot learning or domain adaptation has provided some good results in image classification, still they need a lot of research to be applied in the medical domain. A convolutional neural network for the mass classification is proposed in [45]. Their method requires bulk data to converge the network parameters. Moreover, localization of mass regions is not possible in their network.

The mentioned methods are effective in the detection of suspicious regions. However, several parameters need to be optimized. Also, when masses are occluded by dense tissues or when the density of a mass is similar to that of the surrounding normal dense tissues, the methods may not detect the exact positions of masses as they can only control the growth using intensity criteria (threshold).

III. METHODOLOGY

Female breast parenchyma is a multiplex biological structure and is composed of glandular, fatty, and lymphatic tissues (lymphovascular structures). Mammography imprints the texture information of breast tissue in the image. Although the composition components of mammograms are complex, still the lesions in a mammogram are recognized as higher intensity and texture. Figure 1 shows the process of a typical analysis system.

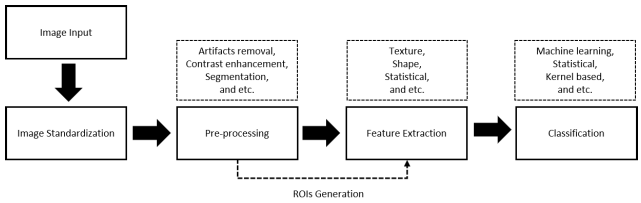


FIGURE 1. General phases of detection algorithm. The overall process is divided into four phases as input, pre-processing, feature extraction, and classification. Most of the diagnosis algorithms follow a similar procedure.

We propose an efficient and unsupervised approach to identify the suspicious regions in mammogram images. The proposed algorithm isolates the spatially interconnected structures in the image, which are concentrated around salient intensities. As a result, it is possible to extract high-level information to analyze further, to characterize the physical properties of mass regions, and to prepare a shortlist of skeptical ROIs. Figure 2 shows our proposed algorithm. Further explanation of the algorithm is explained in the following subsections. Subsections A and B represent the pre-processing of the algorithm.

A. IMAGE STANDARDIZATION

Data from different sources should be converted to one format. Proposed algorithm was tested on two data-sets: Digital Database for Screening Mammography (DDSM) [18], [19] and Mammographic Image Analysis Society Database (mini-MIAS) [35]. MIAS data-set is in Portable Gray Map (PGM) format while DDSM data-set contains images in LJPEG format. We converted the DDSM data-set TO 16-bit Portable

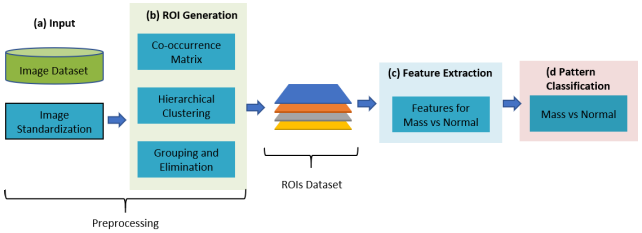


FIGURE 2. Overview of the proposed algorithm. The architecture is divided into four steps. The salient features map is generated in the first step which involves the pectoral muscle removal also. In the second step, hierarchical clustering is used where a proposed distance measure is used to merge the clusters which too helps in the removal of FPs. In the third step different features sets are extracted (FP removal and mass classification). The last step uses the extracted features to train an SVM classifier for FP identification and mass classification.

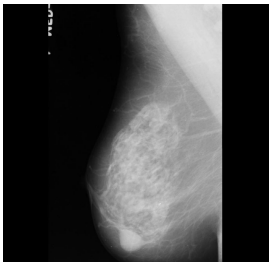


FIGURE 3. Original image from MIAS data-set.

Network Graphics (PNG) format by a wrapper program developed by us.[1]

B. ROI DETECTION PHASE

One of the main tasks is to get mass-candidate regions. The following subsections describe the way to get those regions.

1) SMOOTHING

It is assumed that malignant masses typically distort the surrounding tissues. So, the segmentation process can over-segment the image and it can't get those masses in a single entity. To overcome this problem, prior smoothing of the image is necessary. In the present work, the Gaussian pyramid is used to uniformly highlight the salient regions. Sampling to many levels results in over smoothing the image which converts the image regions as blobs. However, some researchers [32] have performed mass detection on reduced resolutions of 800m. Regions of mass are hyper-dense. We need to get the full mass area to extract meaningful features from the ROI. Abrupt changes in the intensity of the objects present in the image affect the segmentation process. Peaks in the image objects are smoothed by the above described pre-processing.

2) HIERARCHICAL CLUSTERING WITH GLCM (GRAY LEVEL CO-OCCURRENCE MATRIX) DATA

Before segmentation of the image, its contrast was enhanced by CLAHE (Contrast Limited Adaptive Histogram Equalization). Further, we calculate the gray-level co-occurrence matrix from image. GLCM is created with distance one and 4 directions [0 1; −1 1; −1 0; −1 −1] (0°, 45°, 90°, 135°). Other angles were not computed due to redundancy of the data. GLCM data from all directions are summed up and normalized. Figure 4 depicts the explanation of co-occurrence matrix.

Intensities in mass exhibit the glowing effect (intensities are propagated from the center of the masses). Hierarchical clustering can cluster image data according to propagated intensities while having a family structure of concentric objects. At each hierarchical level, a measure of dissimilarity is defined to differentiate clusters, and objects are merged as one if their dissimilarity is less than or equal to the acceptable dissimilarity measure.

Many researchers have proposed methods for multilevel thresholding by discriminant analysis [4], [30], and [33]. They thresholded the image by the cluster analysis irrespective of the physical location of the cluster. This idea works better if the image is multi-modal and we divide it into two clusters (background and foreground). However, it does not give fine results on low-level x-rays images which are mostly uni-modal. In this case, multi-thresholding does not give compact objects for ROI. We incorporated the discriminant analysis [4] with GLCM data to get compact objects. The proposed method clusters the image intensities in a hierarchy, according to their co-occurrence and similarity measure. Several thresholds are found by cutting the dendrogram at the desired level. Initially, each gray-level is designated to a different cluster i.e. g gray-levels in the image will generate q number of clusters and each cluster has its threshold T_i. The family hierarchy of the clustering process can be viewed as a dendrogram. The estimated thresholds for the image to segment can be obtained by cutting the branch in the dendrogram. The clustering algorithm is defined in algorithm 1.

a: DISTANCE METRIC

The distance measure between two clusters in the proposed algorithm is defined as the ratio between the measure of observed dispersement and the expected dispersement. it is calculated as:

$$dist_{(q_i,q_j)} = \frac{(1 - CP_{q_iq_j})(P_{q_i} - P_{q_j})^2\left[\overline{X_{q_i}} - \overline{X_{q_j}}\right]^2}{\sigma^2_{q_iq_j}} \quad (1)$$

where q is the total number of clusters, P_q is the probability density function of image histogram and it can be calculated as equation 2. $CP_{i,j}$ represents the normalized co-occurrence frequency of the cluster pair being merged. It is defined in equation 3. $\overline{X}$ is the mean value of the cluster and defined in equation 5. σ^2 is the variance of both clusters which are being merged. It is defined in equation 7.

$$P_q = \sum_{l=T_{q-1}+1}^{T_q} h(l) \quad (2)$$

[1] Utilities at http://microserf.org.uk/academic/Software.html were used to write a wrapper program.

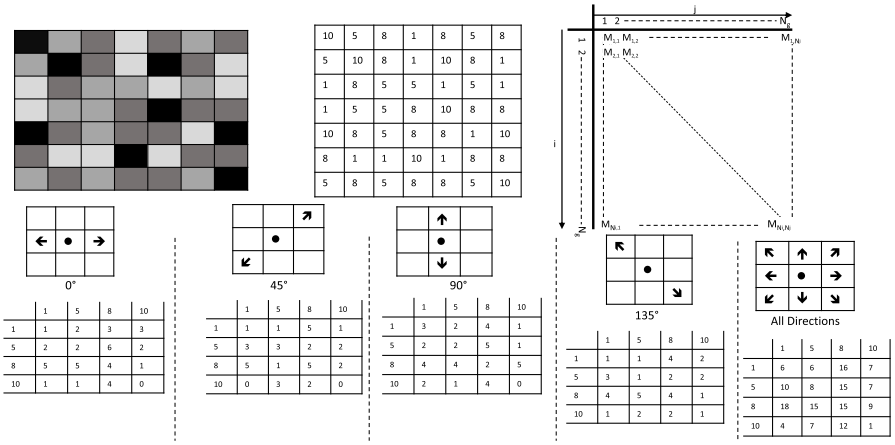


FIGURE 4. Process of co-occurrence matrix. The image shows the procedure to calculate the GLCM matrix. A random image with its pixel values is shown in the image and GLCM co-occurrence matrix is calculated in four directions/angles ((0°, 45°, 90°, 135°)) with a distance 1 (only the immediate neighbour is considered). Finally, all angle matrices are summed up as all direction matrix. Example angle matrices and all direction matrix is shown in image where the first row and column are pixel values and other values are the number of co-occurrences.

Algorithm 1 Clustering Algorithm

Result: Return n thresholds
Given: A set of gray-levels $\{x_1, x_2, \ldots., x_q\}$;
A distance function dist(c1,c2);
m number of threshold levels;
for *i=1 to q* **do**
  $c_i = \{x_i\}$;
  $t_i = \{x_i\}$;
end
C = $\{c_1,\ldots., c_q\}$;
T = $\{t_1,\ldots., t_q\}$;
for *k=1 to q-m* **do**
  - make adjacent cluster pairs;
  - (c_{min1}, c_{min2}) = minimum dist(c_i, c_j) for all c_i, c_j in C;
  - remove c_{min1} and c_{min2} from C;
  - remove t_{min1} and c_{min2} from T;
  - add $\{c_{min1}, c_{min2}\}$ to C;
  - add $\{t_{min1}, c_{min2}\}$ to T;
end

where l represents the gray-level in image (value: [0 255]) such that $\sum_{i=1}^{q} P_i = 1$.

$$CP_{i,j} = \sum_{t=T_{q_{j-1}}+1}^{T_{q_j}} \frac{\sum_{s=Tq_{i-1}+1}^{T_{qi}} CM_{s,t}}{T_{qi} - T_{q_{i-1}}} \tag{3}$$

where $CM_{s,t}$ is the co-occurrence probability of gray-level s and t.

Mean is also called as the expectation of the cluster and can be represented as:

$$\mu = E(q_i) = \sum_{i=1}^{q} l_i P(l_i) \tag{4}$$

so we calculated the mean as:

$$\overline{X}_q = \frac{1}{P_q} \sum_{l=T_{q_1}+1}^{T_q} lh(l) \tag{5}$$

Variance of the distribution is defined as:

$$\sigma^2 = \sum_{i=1}^{q} (l_i - \mu)^2 P(l_i) \tag{6}$$

This formulates the variance into the following equation.

$$\sigma^2_{q_i q_j} = \sum_{l=T_{q-1}+1}^{T_{q2}} \left[l - \overline{CX}_{q_i q_j}\right]^2 h(l) \tag{7}$$

where $\overline{CX}$ is defined as average mean of the cluster pair. It is calculated as the weighted average between the cluster means of the pair being merged:

$$\overline{CX}_{qiqj} = \frac{P_{qi}\overline{X}_{qi} + P_{qj}\overline{X}_{qj}}{P_{qi} + P_{qj}} \tag{8}$$

We imposed a restriction that only the adjacent clusters are allowed to merge. The similarity measurement is adapted by [30]. Pair having the minimum distance value is the best candidate to merge.

The saliency of a region is measured by the nesting depth of hierarchical clustering which identifies nested objects. One statistical parameter LevelParameter is introduced that represents the levels in hierarchical clustering. LevelParameter

(a) Detected ROIs (Objects) (b) Merged ROIs (Objects)

FIGURE 5. The figure shows the intermediate results of detected regions and their merging process in hierarchical clustering.

value of 5 is used in the study. Figure 5 shows the number of objects found in mammogram by segmentation process.

3) GROUPING AND ELIMINATION

Segmentation process described in the previous section results in a large number of segmented objects. We devised an algorithm to reduce the number of objects and extract only the relevant data for analysis. The first step in this process is grouping and elimination. As previously described, masses exhibit the glowing effect, therefore, we first find the dense-core portions and then go to the next threshold level to find objects which encircle the previously detected object. The idea of prestige in link analysis is used along with the hierarchical clustering nodal relation. Every possible region is given a prestige score of 1. When these regions are encircled by other immediate lower densities, they forward their prestige score to the parent. Sum of Euclidean distance between the higher density objects and lower density objects. Lower density objects should cover at least 80% of higher density objects. Algorithm 2 describes the process of merge score. This process is repeated for all the segmented regions at every selected hierarchical level. As Hierarchical clustering also gives a parent-child relationship of clusters, we can use this relationship to avoid unacceptable merging of objects. Objects having at least 3 prestige score from each level is up-sampled to full resolution image. Result of merging process is shown in Figure 5, where 5a represents the detected ROIs and 5b shows the merged objects.

Algorithm 2 Merge Score

```
Result: Merge Score
Given;
Labels = {L_1, L_2,...., L_n};
for i = 1 to n do
    currentLabel = {L_i};
    Objects = Object by current current label;
    numObjects = number of Objects by current label;
    for j = 1 to numObjects do
        mergeScore[i][j] = 1;;
    end
    dist = distanceL2 (Objects[i], Objects[i-1] );
    if dist < 0.2 then
        mergeScore[i][j] + = mergeScore [i-1][j] ;
    end
end
```

C. FEATURES FOR FALSE POSITIVE (FP) ANALYSIS

The following set of features are extracted to classify objects into true mass and breast tissue (false positive). These features are well-established statistical features and are finalized by radiologists too after analyzing the prominent patterns of masses on mammograms.

1) REGION CONTRAST

Generally, mass is imprinted on the mammogram as a dense object as compared to its surroundings, having at least a uniform density. We used this property for classification between true mass and breast tissue. Region Contrast is computed as a difference between mean intensities of foreground and background in ROI. The foreground area is the selected mass or object while the background represents the background area surrounding this object. Regions that results in negative values of region contrast are rejected for further processing.

2) MEAN GRADIENT

Gradient monitors the directional change in intensity. Gradient magnitude describes the velocity of change in the image. We calculated the mean gradient of the boundary pixels which strengthens the compactness of the region (described later).

3) ENTROPY

The concept of entropy is in information theory which states the probabilistic behavior of the information sources. This statistical measure is a measure of randomness that is used to characterize the texture of the image.

4) STANDARD DEVIATION

It is a popular term in statistics that give a measure of the spread of data. This represents the measure, that how close the points are in the given region of the image.

TABLE 1. Mammographic image analysis society (MIAS) data-set.

Benign			Malignant		
Dense	**Fatty-Glandular**	**Glandular**	**Dense**	**Fatty-Glandular**	**Glandular**
8	12	10	2	6	5

5) COMPACTNESS

The value of compactness gives the ratio of contour which encloses an area. it is defined as:

$$compactness = 1 - \frac{4 * pi * A}{P^2} \tag{9}$$

where A is the Area of object enclosed by perimeter P. Usually, benign masses have a higher value of compactness, because it defines that a small perimeter is enclosing a bigger area. We have used this feature in benign vs. malignant classification too.

D. CLASSIFICATION MODEL

SVM (Support Vector Machine) is used to classify the masses. We selected SVM because it gives good results for binary classification. The basic idea behind SVM is to separate the input data by the optimal method. As our data is not linearly separable, we used Gaussian RBF (Radial basis function) kernel. Sigma and C are two important factors for the RBF kernel. Optimal values for RBF were grid-searched between 10^{-3} to 10^{3}. Harmonic Mean (HM) is calculated to compare the C and sigma pairs. Harmonic Mean is defined as:

$$HM = \frac{2 * sens * spec}{sens + spec} \tag{10}$$

where *sens* is sensitivity and *spec* represents the specificity of the system. We adopted a repeated 10-fold cross-validation technique to train, test, and validate the data. The results in Table 3 are given as mean results by repeating the folds 10 times.

IV. RESULTS AND DISCUSSION

A. IMAGE DATABASE

This study was carried out on images from two databases. We selected 56 images from the mini-MIAS database [35]. It includes 13 normal, 13 malignant, and 30 benign cases. The dataset includes all types of masses from both classes (benign and malignant). Table 1 shows the overview of the number of cases used in experiments from MIAS-data-set. We also selected 76 cases from DDSM database [18], [19]. Table 2 shows the summary of DDSM database.

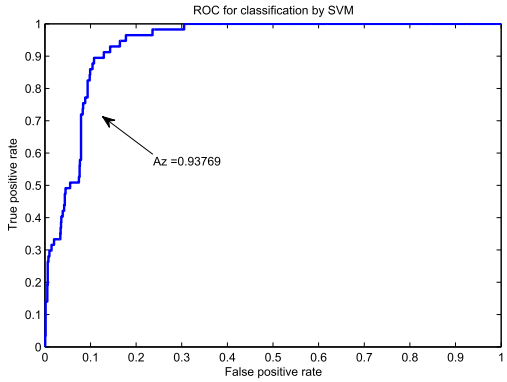
B. DETECTION OF ROIs

Our proposed pre-processing technique detected almost all masses in the dataset. Through careful examination of ROIs, we found that our algorithm missed two cases in the MIAS database. One from Malignant and the other from Benign

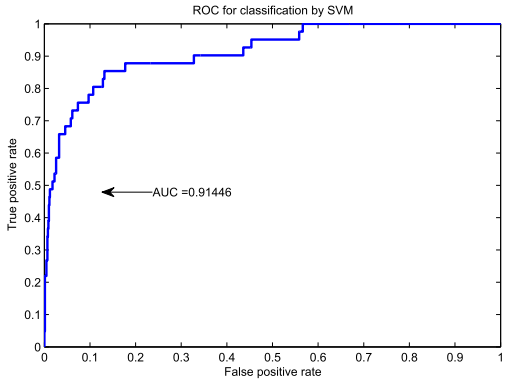
TABLE 2. Digital database for screening mammography (DDSM) dataset information.

Property	Description	Case Count
Density	1	12
	2	29
	3	32
	4	2
Shape	Fine Linear Branch	2
	Irregular	16
	Irregular Architecture	8
	Lobulated	6
	Oval	2
	Pleomorphic	5
	Round	2
Margin	Circumscribed	4
	Circumscribed ill Defined	1
	ill Defined	7
	ill Defined Spiculated	1
	Microlobulated	1
	Clustered	7
	Obscured ill defined	2
	Obscured ill defined spiculated	3
	Spiculated	15

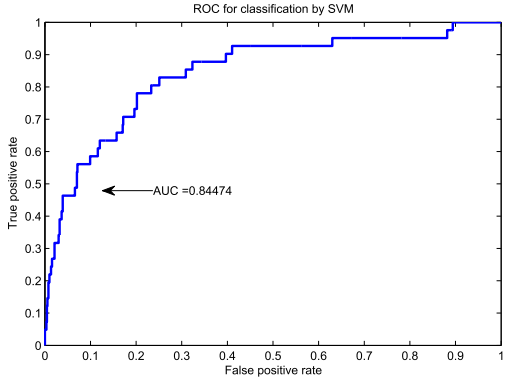
case (mdb179 and mdb191), Dense-glandular and Fatty Glandular. The contrast in these two images was very high and distributed, making it difficult to detect isolated regions. All other masses were successfully detected. This results in the detection accuracy of 95.3%. The detection accuracy on the DDSM dataset was 97.3%. We missed 2 cases. Detected ROIs were carefully compared with the given ground truth data. To quantify the ROI detection, we employed the Jaccard similarity measure between the ground truth and the detection using the proposed method. We consider it a correct detection



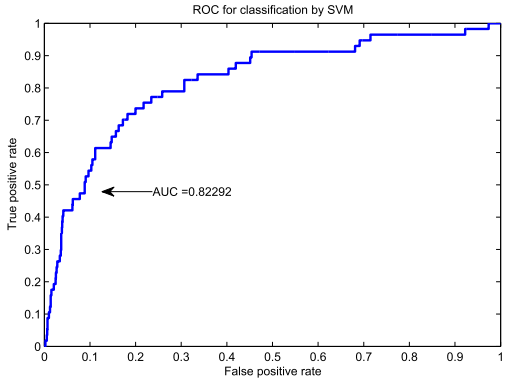
(a) ROC of Normal and Mass (MIAS data-set)



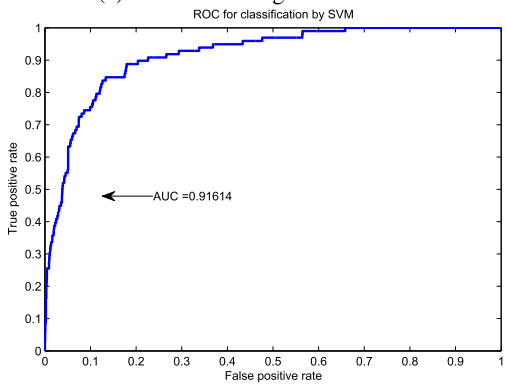
(b) ROC of Normal and Mass (DDSM Data-set)



(c) ROC of training on MIAS and Test on DDSM



(d) ROC of training on DDSM and Test on MIAS



(e) ROC of MIAS and DDSM combined

FIGURE 6. **Breast tissue vs. mass classification results (ROC plots).**

if the intersection region between the two ROIs is bigger than 90%.

C. NORMAL AND MASS DIFFERENTIATION

Our algorithm detected all the malignant masses except one (mdb0186) on the MIAS dataset. Our detection rate for benign masses is however not so prominent. Out of 30 tested cases, it missed 6 cases. Three of these missed masses were Fatty (mdb069, mdb080, and mdb195), two were Dense-glandular (mdb193 and mdb290), and one was Fatty-glandular (mdb190). The total accuracy of the system was 83.43%. Figure 8 shows the example ROI which is classified as mass.

We further investigated the missed cases and found the following observations. In the first missed case (mdb069), the margin and boundary with the wide transition zone, if we compare with opposite side breast, the lesion could be detectable, and in clinical practice, we describe it as architectural distortion. In the case of mdb080, the tumor lesion is a subtle ill margined, non-mass-like parenchymal asymmetric pattern. In the case of mdb195, the malignant lesion is almost isodensed to the normal breast fatty parenchyma. So, the detection is not feasible. In mdb186 we found that

TABLE 3. Average specificity and sensitivity of mass vs. normal classification by proposed method.

Training Dataset	Testing Dataset	Sensitivity (%)	Std. Dev	Specificity (%)	Std. Dev.
MIAS	MIAS	83.06	0.17	87.25	0.25
DDSM	DDSM	83.79	0.14	76.63	0.28
MIAS + DDSM	MIAS + DDSM	82.50	0.01	74.83	0.60
MIAS	DDSM	87.80	0.09	64.88	0.11
DDSM	MIAS	74.39	0.03	84.67	0.05

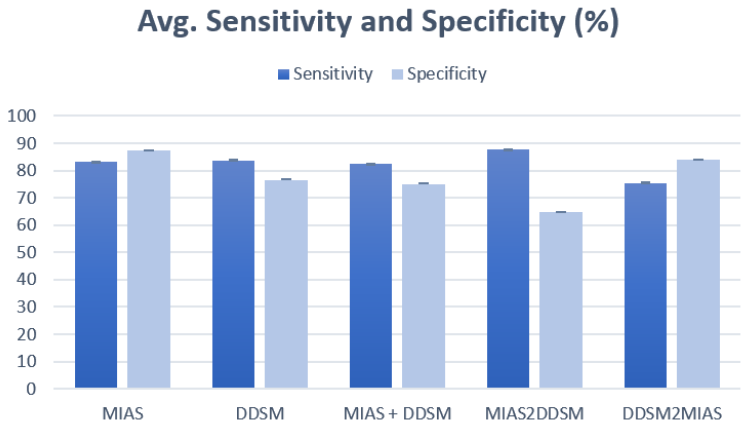


FIGURE 7. Average sensitivity and specificity (%) plot with standard deviation.

the mass has poor contrast as well as it lacks the dense region. Itś contrast with respect to the surrounding was very poor. In benign cases, where the algorithm was unable to classify masses, we observed that in three fatty and one fatty glandular cases (mdb069, mdb080, mdb190, and mdb195) the masses were not clear. They do not have a central core region and their contrast with respect to their surrounding was poor too. We are confident that if we add some good contrast enhancement techniques, our algorithm performance will be improved by classifying the above-described cases as well. The remaining two dense-glandular cases (mdb193 and mdb290) do not follow the assumption we made in this paper (they do not have a glowing effect), so features values were not good in these cases to classify them. To successfully detect masses in these cases, it may require additional methods or include more features. In the present work, we did not reject any region because of its size, this results in generating a large number of false positives. Although our classification phase reduces the number of FPs; we aim to reduce the number of FPs by the improved algorithm in future work. We also believe that automatic breast density assessment before applying our method will improve the performance [24].

We validated the results by plotting the receiver operating characteristic (ROC) curve, which illustrates the performance of the binary classifier system as its discrimination threshold is varied. Figure 6 shows the ROC curve of classification between normal and mass data, which is obtained by varying the threshold on the probabilities by the classifier (SVM). AUC refers to the Area Under Curve. Table 3 shows the classification results in terms of specificity, sensitivity, and their standard deviation which can also be seen in Figure 7. In the medical domain, sensitivity alone is not enough, the algorithm should also yield good specificity results. As previously described, we used harmonic mean (equation 10) to get the best pair of specificity and sensitivity.

Algorithm missed 2 cases from malignant category and 6 from the benign category of DDSM data-set. The maximum sensitivity and specificity pair, we achieved is 91.32% and 85.05% respectively. Average sensitivity and specificity are 76.19% and 87.05% respectively.

We also tested our algorithm for its generality by training it on one data-set and testing on the other. The algorithm was trained on MIAS data-set, tested on DDSM and vice versa. Algorithm results in table 3 confirm our claim that the proposed algorithm is not bounded to some limited type of masses or abnormalities. It covers a wide spectrum of masses. The distribution of the data-set is uneven, which degrades the performance of the learning algorithm.

Investigation of the missed cases confirms the reasons described earlier. Case0004 from DDSM shows poor contrast around the mass, making it difficult to be detected. Case0005, case0006, and case008 do not follow the assumption we made in the paper. More features may be required to detect those

masses. We also calculated the number of false positives per image which was 4.67 FP/Image. This number is calculated only on Normal Images to give a fair view of the system.

D. COMPARISON WITH EXISTING ALGORITHMS

Table 4 compares the related mass detection techniques. In [15], the mammograms are enhanced and detected based on the improved MCL with the MCA model. The outer contours of masses detected by the model are extended to a rectangular region. Then the rectangular masses are set as the ROIms to perform spatial LDA analysis. These ROIs with different sizes are then grouped into two sets. [26] used the bilateral similarity analysis to reduce the FPs. They tested their method on a set of 332 mammograms, which shows a 34% FP reduction in comparison to, single-view CAD, with a detection sensitivity of 85%. A dual-stage adaptive thresholding (DuSAT) [3] has been successfully applied to selected mass regions in mammograms. For removal of background, Thresholding is applied in [21]. [20] have proposed intensity and gradient-based method with Abnormality detection classifier (ADC) for the classification of normal and abnormal mammograms. Feature weights are determined using Principal Component Analysis (PCA) and have obtained a sensitivity of 86%. [34] stated their results of mass detection phase where they achieved 84.4% detection accuracy. Their algorithm is based on image enhancement where Gaussian Markov Random Field (MRF) is used for mass segmentation. However, they did not classify the ROIs into mass and non-mass regions. [23] also reported their detection accuracy as 94.44%. They presented a particle swarm optimization (PSO) based detection technique.

Work presented by [9], [14], [17], and [40] can be consider as the baseline in recent work in this domain. [9] implemented a fully automated system by extracting local binary pattern LBP features. They used SVM for classification. A feature selection technique is also proposed. [9] reported their performance in terms of sensitivity and reported 75.86% for overall CAD performance on the MIAS database. [16] reported their results on already selected 305 ROIs and achieved a sensitivity of 76.53%. They extracted features from Grey-level. They extracted features from Grey-level co-occurrence matrices (GLCM) and then classify features into mass and non-mass regions. [14] proposed the technique of curvelet transformation, feature selection, and then classification by SVM. They manually cropped the ROIs and then applied their algorithm. Their reported accuracy is higher than 90%, but their algorithm is not fully automated, they lack a mass detection phase. All methods were tested on a separate dataset, cross-validation between the datasets was never performed.

TABLE 4. Result comparison with state-of-the-art methods.

Method	Training	Testing	Sensitivity (%)
Gao et al. [15]	Mias	mias	80
	ddsm	ddsm	90
Varela et al. [42]	ddsm	ddsm	80
Jayasree et al. [8]	mias	mias	80
	ddsm	ddsm	80
Dominguez and Nandi [12]	mias	mias	80
Kozegar et al. [22]	mias	mias	90
Campanini et al. [6]	ddsm	ddsm	86
Martins et al. [10]	ddsm	ddsm	76.8
Liu and Feng [27]	ddsm	ddsm	90
Tai et al. [39]	ddsm	ddsm	75
Li et al. [26]	mias	mias	85
Anitha et al. [3]	ddsm	ddsm	92.5
	mias	mias	93.5
Kashyap et al. [21]	ddsm	ddsm	91.76
	mias	mias	94.63
Jen et al. [20]	mias	mias	88
	ddsm	ddsm	86

FIGURE 8. Figure shows the shape of detected mass ROI.

V. CONCLUSION

This paper proposes a new mass detection algorithm in mammogram images. The proposed method is fully automated. It finds the candidate regions by segmenting the salient regions in a mammogram and then extract features to differentiate between the breast tissue and mass. Promising results are obtained in mass identification and normal vs. mass tissue classification. Classification results confirm that the segmentation process extracts enough information to find masses and localize them in a mammogram. Experiments were performed on mini-MIAS and DDSM databases to show the usefulness and generalization of the proposed algorithm. Correlating the full image set (CC and MLO) is considered as future work that can also help to identify the architectural distorted mammograms.

REFERENCES

[1] Q. Abbas, "DeepCAD: A computer-aided diagnosis system for mammographic masses using deep invariant features," *Computers*, vol. 5, no. 4, p. 28, Oct. 2016.
[2] P. Agrawal, M. Vatsa, and R. Singh, "Saliency based mass detection from screening mammograms," *Signal Process.*, vol. 99, pp. 29–47, Jun. 2014.
[3] J. Anitha, J. D. Peter, and S. I. A. Pandian, "A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms," *Comput. Methods Programs Biomed.*, vol. 138, pp. 93–104, Jan. 2017.
[4] A. Z. Arifin and A. Asano, "Image thresholding by histogram segmentation using discriminant analysis," in *Proc. Indonesia-Japan Joint Sci. Symp.*, 2004, pp. 169–174.
[5] *Breast Care and You*, Brigham, F. H. Women's Hospital, Boston, MA, USA, 2015.

[6] R. Campanini, D. Dongiovanni, E. Iampieri, N. Lanconelli, M. Masotti, G. Palermo, A. Riccardi, and M. Roffilli, "A novel featureless approach to mass detection in digital mammograms based on support vector machines," *Phys. Med. Biol.*, vol. 49, no. 6, p. 961, 2004.

[7] P. Casti, A. Mencattini, M. Salmeri, A. Ancona, F. Mangeri, M. L. Pepe, and R. M. Rangayyan, "Contour-independent detection and classification of mammographic lesions," *Biomed. Signal Process. Control*, vol. 25, pp. 165–177, Mar. 2016.

[8] J. Chakraborty, A. Midya, S. Mukhopadhyay, R. M. Rangayyan, A. Sadhu, V. Singla, and N. Khandelwal, "Computer-aided detection of mammographic masses using hybrid region growing controlled by multilevel thresholding," *J. Med. Biol. Eng.*, vol. 39, no. 3, pp. 352–366, Jun. 2019.

[9] J. Y. Choi and Y. M. Ro, "Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms," *Phys. Med. Biol.*, vol. 57, no. 21, p. 7029, 2012.

[10] L. de Oliveira Martins, G. B. Junior, A. C. Silva, A. C. de Paiva, and M. Gattass, "Detection of masses in digital mammograms using k-means and support vector machine," *ELCVIA Electron. Lett. Comput. Vis. Image Anal.*, vol. 8, no. 2, pp. 39–50, 2009.

[11] N. Dhungel, G. Carneiro, and A. P. Bradley, "A deep learning approach for the analysis of masses in mammograms with minimal user intervention," *Med. Image Anal.*, vol. 37, pp. 114–128, Apr. 2017.

[12] A. R. Domínguez and A. K. Nandi, "Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection," *Comput. Med. Imag. Graph.*, vol. 32, no. 4, pp. 304–315, Jun. 2008.

[13] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby, "A concentric morphology model for the detection of masses in mammography," *IEEE Trans. Med. Imag.*, vol. 26, no. 6, pp. 880–889, Jun. 2007.

[14] M. M. Eltoukhy and I. Faye, "An optimized feature selection method for breast cancer diagnosis in digital mammogram using multiresolution representation," *Appl. Math. Inf. Sci.*, vol. 8, no. 6, pp. 2921–2928, Nov. 2014.

[15] X. Gao, Y. Wang, X. Li, and D. Tao, "On combining morphological component analysis and concentric morphology model for mammographic mass detection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 266–273, Mar. 2010.

[16] S. J. S. Gardezi, I. Faye, and M. M. Eltoukhy, "Analysis of mammogram images based on texture features of curvelet sub-bands," in *Proc. 5th Int. Conf. Graph. Image Process. (ICGIP)*, Jan. 2014, Art. no. 906924.

[17] F. B. Garma, M. A. E. Almoon, M. M. Bakry, M. E. Mohamed, and E. Osman, "Detection of breast cancer cells by using texture analysis," *J. Clin. Eng.*, vol. 38, no. 2, pp. 79–83, 2013.

[18] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, Jr., R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography*. Dordrecht, The Netherlands: Springer, 1998, pp. 457–460.

[19] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proc. 5th Int. Workshop Digit. Mammogr.*, 2000, pp. 212–218.

[20] C.-C. Jen and S.-S. Yu, "Automatic detection of abnormal mammograms in mammographic images," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3048–3055, Apr. 2015.

[21] K. L. Kashyap, M. K. Bajpai, and P. Khanna, "An efficient algorithm for mass detection and shape analysis of different masses present in digital mammograms," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 9249–9269, Apr. 2018.

[22] E. Kozegar, M. Soryani, B. Minaei, and I. Domingues, "Assessment of a novel mass detection algorithm in mammograms," *J. Cancer Res. Therapeutics*, vol. 9, no. 4, p. 592, 2013.

[23] W. C. Lin, S. C. Hsu, and A. C. Cheng, "Mass detection in digital mammograms system based on PSO algorithm," in *Proc. Int. Symp. Comput., Consum. Control (ISC)*, Jun. 2014, pp. 662–668.

[24] Y. N. Law, M. K. Lieng, J. Li, and D. A.-A. Khoo, "Automated breast tissue density assessment using high order regional texture descriptors in mammography," *Proc. SPIE*, vol. 9035, Mar. 2014, Art. no. 90351Q.

[25] J. Lesniak, R. Hupse, M. Kallenberg, M. Samulski, R. Blanc, N. Karssemeijer, and G. Székely, "Computer aided detection of breast masses in mammography using support vector machine classification," *Proc. SPIE*, vol. 7963, Mar. 2011, Art. no. 79631K.

[26] Y. Li, H. Chen, Y. Yang, L. Cheng, and L. Cao, "A bilateral analysis scheme for false positive reduction in mammogram mass detection," *Comput. Biol. Med.*, vol. 57, pp. 84–95, Feb. 2015.

[27] X. Liu, X. Xu, J. Liu, and Z. Feng, "A new automatic method for mass detection in mammography with false positives reduction by supported vector machine," in *Proc. 4th Int. Conf. Biomed. Eng. Informat. (BMEI)*, vol. 1, Oct. 2011, pp. 33–37.

[28] M. Mavroforakis, H. Georgiou, N. Dimitropoulos, D. Cavouras, and S. Theodoridis, "Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines," *Eur. J. Radiol.*, vol. 54, no. 1, pp. 80–89, 2005.

[29] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels, "Detection of breast masses in mammograms by density slicing and texture flow-field analysis," *IEEE Trans. Med. Imag.*, vol. 20, no. 12, pp. 1215–1227, Dec. 2001.

[30] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, nos. 285–296, pp. 23–27, 1975.

[31] D. C. Pereira, R. P. Ramos, and M. Z. do Nascimento, "Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm," *Comput. Methods Programs Biomed.*, vol. 114, no. 1, pp. 88–101, Apr. 2014.

[32] N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Trans. Med. Imag.*, vol. 15, no. 1, pp. 59–67, Feb. 1996.

[33] R. Rodrigues, R. Braz, M. Pereira, J. Moutinho, and A. M. Pinheiro, "A two-step segmentation method for breast ultrasound masses based on multi-resolution analysis," *Ultrasound Med. Biol.*, vol. 41, no. 6, pp. 1737–1748, Jun. 2015.

[34] V. Rodríguez-López, R. Miranda-Luna, and J. A. Arias-Aguilar, "Detection of masses in mammogram images using morphological operators and Markov random fields," in *Advances in Artificial Intelligence and Its Applications*. Berlin, Germany: Springer, 2013, pp. 558–569.

[35] J. Suckling, "The mammographic image analysis society digital mammogram database," in *Proc. Exerpta Medica Int. Congr. Ser.*, vol. 1069, 1994, pp. 375–378.

[36] W. Sun, B. Zheng, F. Lure, T. Wu, J. Zhang, B. Y. Wang, E. C. Saltzstein, and W. Qian, "Prediction of near-term risk of developing breast cancer using computerized features from bilateral mammograms," *Comput. Med. Imag. Graph.*, vol. 38, no. 5, pp. 348–357, 2014.

[37] B. Surendiran, A. Vadivel, and H. Selvaraj, "A soft-decision approach for microcalcification mass identification from digital mammogram," *Sat*, vol. 1, p. 2, Jan. 2008.

[38] S. Suzuki, X. Zhang, N. Homma, K. Ichiji, N. Sugita, Y. Kawasumi, T. Ishibashi, and M. Yoshizawa, "Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis," in *Proc. 55th Annu. Conf. Soc. Instrum. Control Eng. Jpn. (SICE)*, Sep. 2016, pp. 1382–1386.

[39] S.-C. Tai, Z.-S. Chen, and W.-T. Tsai, "An automatic mass detection system in mammograms based on complex texture features," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 2, pp. 618–627, Mar. 2014.

[40] J. Torrents-Barrena, D. Puig, M. Ferre, J. Melendez, L. Diez-Presa, M. Arenas, and J. Marti, "Breast masses identification through pixel-based texture classification," in *Breast Imaging*. Cham, Switzerland: Springer, 2014, pp. 581–588.

[41] N. Vállez, G. Bueno, O. Déniz, J. Dorado, J. A. Seoane, A. Pazos, and C. Pastor, "Breast density classification to reduce false positives in CADe systems," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 569–584, 2014.

[42] C. Varela, P. G. Tahoces, A. J. Méndez, M. Souto, and J. J. Vidal, "Computerized detection of breast masses in digitized mammograms," *Comput. Biol. Med.*, vol. 37, no. 2, pp. 214–226, 2007.

[43] P. S. Vikhe and V. R. Thool, "Morphological operation and scaled Réyni entropy based approach for masses detection in mammograms," *Multimedia Tools Appl.*, vol. 77, no. 18, pp. 23777–23802, 2018.

[44] Y. Wang, J. Li, and X. Gao, "Latent feature mining of spatial and marginal characteristics for mammographic mass classification," *Neurocomputing*, vol. 144, pp. 107–118, Nov. 2014.

[45] P. Xi, C. Shu, and R. Goubran, "Abnormality detection in mammography using deep convolutional neural networks," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2018, pp. 1–6.

[46] S. B. Y. Tasdemir, K. Tasdemir, and Z. Aydin, "A review of mammographic region of interest classification," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 10, p. e1357, Sep./Oct. 2020.

[47] L. Zheng, A. K. Chan, G. McCord, S. Wu, and J. S. Liu, "Detection of cancerous masses for screening mammography using discrete wavelet transform-based multiresolution Markov random field," *J. Digit. Imag.*, vol. 12, no. S1, pp. 18–23, May 1999.

SAJIDA IMRAN received the Ph.D. degree from Ajou University, in 2018. She worked as an Assistant Professor at the Department of Computer Engineering, The University of Lahore, Pakistan. She is currently working as an Assistant Professor with the Department of Computer Engineering, King Faisal University, Saudi Arabia. She has authored a number of international journals. Her research interests include wireless internet technologies for localization, detection and tracking of objects, and applications of Internet of Things using various machine learning techniques.

BILAL AHMED LODHI received the Ph.D. degree from Korea University, South Korea. He is currently working as a Research Fellow at Queen's University Belfast, U.K. His research interests include visual recognition, natural language description of visual contents, and real-world applications of computer vision and machine learning, specifically robust representative learning and its interpretations, social media analysis, and predictions.

ALI ALZAHRANI received the B.E. degree in computer engineering from Umm Al-Qura University, Mecca, Saudi Arabia, and the M.Sc. and Ph.D. degrees in computer engineering from the University of Victoria, BC, Canada, in 2015 and 2018, respectively. He is currently an Assistant Professor with the Department of Computer Engineering, King Faisal University. His research interests include hardware security, encryption processors, image processing, and systems-on-chip.