Hindawi
Journal of Healthcare Engineering
Volume 2020, Article ID 9162464, 21 pages
https://doi.org/10.1155/2020/9162464

Review Article

Methods Used in Computer-Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review

Saleem Z. Ramadan

Department of Industrial Engineering, German Jordanian University, Mushaqar 11180, Amman, Jordan

Correspondence should be addressed to Saleem Z. Ramadan; saleem.ramadan@gju.edu.jo

Received 10 May 2019; Revised 25 December 2019; Accepted 13 February 2020

Academic Editor: Angkoon Phinyomark

According to the American Cancer Society's forecasts for 2019, there will be about 268,600 new cases in the United States with invasive breast cancer in women, about 62,930 new noninvasive cases, and about 41,760 death cases from breast cancer. As a result, there is a high demand for breast imaging specialists as indicated in a recent report for the Institute of Medicine and National Research Council. One way to meet this demand is through developing Computer-Aided Diagnosis (CAD) systems for breast cancer detection and diagnosis using mammograms. This study aims to review recent advancements and developments in CAD systems for breast cancer detection and diagnosis using mammograms and to give an overview of the methods used in its steps starting from preprocessing and enhancement step and ending in classification step. The current level of performance for the CAD systems is encouraging but not enough to make CAD systems standalone detection and diagnose clinical systems. Unless the performance of CAD systems enhanced dramatically from its current level by enhancing the existing methods, exploiting new promising methods in pattern recognition like data augmentation in deep learning and exploiting the advances in computational power of computers, CAD systems will continue to be a second opinion clinical procedure.

1. Introduction

Cancer is a disease that occurs when abnormal cells grow in an uncontrolled manner in a way that disregards the normal rules of cell division, which may cause uncontrolled growth and proliferation of the abnormal cells. This can be fatal if the proliferation is allowed to continue and spread in such a way that leads to metastasis formation. The tumor is called malignant or cancer if it invades surrounding tissues or spreads to other parts of the body [1]. Breast cancer forms in the same way and usually starts in the ducts that carry milk to the nipple or in the glands that make breast milk. Cells in the breast start to grow in an uncontrolled manner and form a lump that can be felt or detected using mammograms [2]. Breast cancer is the most prevalent cancer between women and the second cause of cancer-related deaths among them worldwide [3–5]. According to the American Cancer Society's forecasts for 2019, there will be about 268,600 new cases in the United States of invasive breast cancer diagnosed in women, about 62,930 new noninvasive cases, and about 41,760 death cases from breast cancer [3]. The death rates among women dropped 40% between 1989 and 2016, and since 2007, the death rates in younger women are steady and are steadily decreasing in older women due to early detection through screening, increased awareness, and better treatment [3, 6].

Mammography, which is performed at moderate X-ray photon energies, is commonly used to screen for breast cancer [7, 8]. If the screening mammogram showed an abnormality in the breast tissues, a diagnostic mammogram is usually recommended to further investigate the suspicious areas. The first sign of breast cancer is usually a lump in the breast or underarm that does not go after the period. Usually, these lumps can be detected by screening mammography long before the patient can notice them even if these lumps are very small to do any perceptible changes to the patient [9].

Several studies showed that using screening mammography as an early detection tool for breast cancer reduces breast cancer mortality [10–12]. Unfortunately, mammography has a low detection rate and 5% to 30% of false-negative results depending on the lesion type, the age of the patient, and the breast density [13–19]. Denser breasts are harder to diagnose as they have low contrast between the cancerous lesions and the background [20, 21]. The miss of classification in mammography is about four to six times higher in dense breasts than in nondense breasts [17, 20–24]. Dense breast reduces the test sensitivity (increases false-positive value), hence requiring unnecessary biopsy, and decreases test specificity (increases false-negative value), hence missing cancers [25].

Radiologists try to enhance the sensitivity and specificity of mammography by double reading the mammograms by different radiologists. Some authors reported that double reading enhances the specificity and sensitivity of mammography [26–28] but with extra cost on the patient. A recent study [29] and an older study [30] showed that the detection rate of double reading was not statistically different from the detection rate of a single reading in digital mammograms and hence the double reading is not a cost-effective strategy in digital mammography. The inconsistency in the results shows that there is a need for further studies in this area. Recently, Computer-Aided Diagnosis (CAD) systems are used to assist doctors in reading and interpreting medical images such as the location and the likelihood of malignancy in a suspicious lesion [7]. CADe and CADx schemes are used to differentiate between two strands of CAD systems. The main difference between CADe and CADx is that CADe stands for Computer-Aided Detection system, in which CADe systems do not present the radiological characteristics of tumors but help in locating and identifying possible abnormalities in the image and leaving the interpretation to the radiologist. On the other hand, CADx stands for Computer-Aided Diagnosis system, in which CADx serves as decision aids for radiologists to characterize findings from radiological images identified by either a radiologist or a CADe system. CADx systems do not have a good level of automation and do not detect nodules.

CAD helped the doctors to improve the interpretations of images in terms of accuracy in detection and productivity in time to read and interpret the images [31–37]. A study regarding CAD systems showed an increase in the radiologists' performance for those who used CAD systems [38]. Another study indicated that the detection rate for double reading was not significantly different from the detection rate of a single reading accompanied by a CAD system [23]. Typically, a CAD session starts with the radiologist reading the mammogram to look for suspicious patterns in it followed by the CAD system scanning the mammograms and looking for suspicious areas. Finally, the radiologist analyzes the prompts given by the CAD system about the suspicious areas [7].

The two main signs for malignancy are the microcalcification and masses [24, 39, 40]. Microcalcifications can be described in terms of their size, density, shape, distribution, and number [41]. Microcalcification detection in denser breasts is hard due to the low contrast between the microcalcification and the surrounding tissues [21]. A valuable study on how to enhance contrast, extraction, suppression of noise, and classification of microcalcification can be found in [42]. Masses, on the other hand, are circumscribed lumps in the breast and are categorized as benign or malignant. Masses can be described by shape, margin, size, location, and contrast. The shape can be further classified as round, lobular, oval, and irregular. Margin can also be further classified as obscured, indistinct, and spiculated. Masses are harder to detect by the radiologists than microcalcification because of their similarity to the normal tissues [43, 44]. Many studies presented the usage of CAD systems in mammography diagnosis such as [7, 45–50]. Like any other algorithm for a classification problem, the CAD system can be divided into three distinct areas: feature extraction, feature selection, and classification methodologies. On top of these three major areas, CAD systems depend heavily on an image enhancement step to prepare the mammogram for further analysis. Figure 1 shows a flowchart for a typical CAD system schema.

In this study, we are presenting the developments of CAD methods used in breast cancer detection and diagnosis using mammograms, which include preprocessing and contrast enhancement, features extraction, features selection, and classification methods. The rest of the paper will be organized based on the schema in Figure 1 as follows: Section 2 presents the preprocessing and enhancement step, Section 3 discusses features selection and features extraction step, Section 4 is devoted to discussing classification through classifiers and combined classifiers, and Section 5 presents the conclusions.

2. Preprocessing and Contrast Enhancement

Mammograms do not provide good contrast between normal glandular breast tissues and malignant ones and between the cancerous lesions and the background especially in dense breasts [20–24, 51]. There are recognized poor contrast problems inherent to mammography images. According to the Beer-Lambert equation, the thicker the tissue is, the fewer the photons pass through it. This means that as the X-ray beam passes through normal glandular breast tissues and malignant ones in dense breast tissues, its attenuation will not differ much between the two tissues and hence there will be low contrast between normal glandular and malignant tissues [52]. Another well-known problem in mammograms is noise. Noise occurs in mammograms when the image brightness is not uniform in the areas that represent the same tissues as it supposes to be due to nonuniform photon distribution. This is called quantum noise. This noise reduces image quality especially in small objects with low contrast such as a small tumor in a dense breast. It is known that quantum noise can be reduced by increasing the exposure time. For health reasons, most of the time the radiologist prefers to decrease the exposure time for the patient at the expense of increasing the quantum noise, which will result in reducing the visibility of the mammogram. The presence of noise in a mammogram gives it a grainy appearance. The grainy appearance reduces the visibility of some features within the image especially for small objects with low contrast, which is the case for a small tumor in a dense breast [52, 53]. Because of this low-contrast

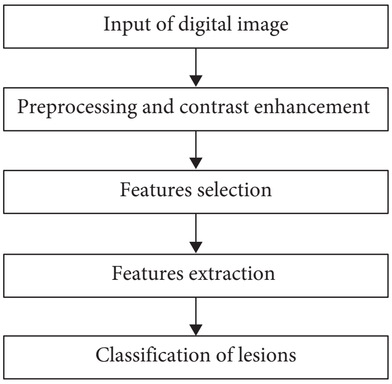
Input of digital image
Preprocessing and contrast enhancement
Features selection
Features extraction
Classification of lesions

Figure 1: Flow chart for a typical CAD system.

problem, contrast enhancement techniques were proposed in the literature. A good review of conventional contrast enhancement techniques can be found in [54]. Unfortunately, there is no unified metrics for evaluating the performance of the preprocessing techniques to enhance the low-contrast problem and the existing evaluations are still highly subjective [55, 56].

Image enhancement usually is done by changing the intensity of the pixels for the input image [57]. The conventional histogram equalization technique for image enhancement is an attractive approach for its traceability and simplicity. The conventional histogram equalization starts by collecting statistics about the intensity of the image's pixels and formulating a histogram for the intensity levels and their frequency as in

$$\mathbf{H} = [f_i], \tag{1}$$

where i is intensity index and f_i is the frequency for intensity index i.

Using equation (1), a cumulative density function is defined as in

$$\mathrm{CF}(i) = \sum_{i=0}^{L-1} \frac{f_i}{N}, \tag{2}$$

where L is the number of intensity levels and N is the total number of pixels in the image. A transformation function is defined based on equation (2) to give the output image I_{out} as follows:

$$I_{\text{out}}(i) = I_{\min} + (I_{\max} - I_{\min}) \times \mathrm{CF}(i), \tag{3}$$

where $I_{\min}$ and $I_{\max}$ are the minimum and the maximum intensity levels of the input image, respectively.

Unfortunately, the conventional histogram equalization tends to shift the output image brightness to the middle of the allowed intensity range. To overcome this problem, subimage based histogram equalization methods were proposed in the literature where the input image is divided into subimages and the intensity for each subimage is manipulated independently. A bi-histogram equalization method discussed in [58, 59] splits the image into two subregions of high and low mean brightness based on the average intensity of all pixels and then applies the histogram equalization to each subregion independently. The mathematics of bi-histogram equalization method starts by calculating the expected value for the intensity of the input image as in

$$E[I_{\text{inp}}] = \sum_{i=0}^{L-1} i \times \frac{f_i}{N}. \tag{4}$$

Based on $E[I_{\text{inp}}]$, two subimages, I_{low} and I_{high}, are created such that

$$\begin{aligned} EI_{\text{low}} &= I_{\text{inp}}(i), \quad \forall i \le E[I_{\text{inp}}], \\ I_{\text{high}} &= I_{\text{inp}}(i), \quad \forall i > E[I_{\text{inp}}]. \end{aligned} \tag{5}$$

Then I_{low} and I_{high} are equalized using their corresponding range of intensities to produce two enhanced images $I_{\text{low}_{\text{enh}}}$ and $I_{\text{high}_{\text{enh}}}$. The enhanced output image is then constructed as the union of these two subimages. The first limitation for this method is that the original brightness can be preserved only if the original image has a symmetric intensity histogram such that the number of pixels in I_{low} is equal to the number of pixels in I_{high}. If the histogram is not symmetric, the mean tends to shift toward the long tail and hence the number of pixels in each subimage will not be equal. The second limitation arises when pixels intensities tend to concentrate in a narrow range. This will be shown as peaks in the histogram and hence will generate artifacts in the output image. A third limitation is that the image may suffer from overenhancement especially if the brightness dispersion in the image is high such that there are regions of very high and very low brightness.

To overcome the limitation generated from the unequal number of pixels in the subimages, dualistic subimage histogram equalization discussed in [60] incorporated the median value as the divider threshold in the process instead of the mean to have an equal number of pixels in each subimage. This method maximizes Shannon's entropy of the output image [61]. In this method, the input image is divided into two subimages just like with equal number pixels and hence the original brightness of the input image can be preserved to some extent. To overcome the limitation generated from pixels concentrated in a narrow range, the histogram peaks are clipped prior to cumulative density calculations [62]. The procedure used is similar to bi-histogram equalization method discussed earlier in which two subimages are created along with their histograms, but on the top of that, the mean values for the two subimages are calculated as the clipping limits for the subimages as in

$$\begin{aligned} \mathrm{CL}_{\text{low}} &= \sum_{i=0}^{E[I_{\text{inp}}]} i \times \frac{f_i}{N}, \\ \mathrm{CL}_{\text{high}} &= \sum_{i=E[I_{\text{inp}}]+1}^{L-1} i \times \frac{f_i}{N}. \end{aligned} \tag{6}$$

The clipping limits are used to generate the histograms for the subimages as in equation (7) and the procedure continues after that as in the bi-histogram equalization method.

$$\mathbf{H}_{\text{low}} = \left\{ \begin{matrix} H_{\text{low}}(i), & H_{\text{low}}(i) < \text{CL}_{\text{low}} \\ H_{\text{low}}, & \text{Otherwise} \end{matrix} \right\},$$
$$\mathbf{H}_{\text{low}} = \left\{ \begin{matrix} H_{\text{high}}(i), & H_{\text{high}}(i) < \text{CL}_{\text{high}} \\ H_{\text{high}}, & \text{Otherwise} \end{matrix} \right\}. \tag{7}$$

A difficulty that sometimes arises in this method is that the mean brightness for the resulting image is hard to calculate, as it sometimes does not have a closed-form expression to evaluate it [63].

An attempt to mitigate the limitation of over-enhancement is by multiple division method proposed in [63] in which dynamic quadrants histogram equalization plateau limit method divides the image into four sub-images. Another attempt is by median-mean based sub-image clipped histogram equalization proposed in [64], which is an enhancement for the dynamic quadrants histogram equalization plateau limit method. In this method, the mean brightness of the input image was used first to divide the input image into subimages, these subimages are then further divided using the mean brightness for the subimages, and then the peaks of the subimage histogram were clipped using median values. The minimum mean brightness error bi-histogram equalization method was used in [65], which is an exhaustive search method, to determine the best separation threshold. This method, like the subimage based histogram equalization method, cannot guarantee a match between the input and the output brightness. Another drawback of this method is that it may need a considerable amount of computational time, as it is exhaustive in nature.

A dynamic stretching strategy is adopted in [63] for contrast enhancement instead of histogram equalization with the mean or median value for the separation threshold in which the efficient golden section search approach was used to find the optimal threshold, which preserves the mean brightness in the output image. Equation (8) gives the function used for the golden section search. The golden section search method is basically searching for the mean output image intensity $E[I_{\text{out}}]$ that will minimize the absolute deviation between the output image mean intensity and the input image mean intensity $E[I_{\text{inp}}]$.

$$f(E[I_{\text{out}}]) = \left|E\left[I_{\text{inp}}\right] - E\left[I_{\text{out}}\right]\right|. \tag{8}$$

There are many methods in literature to reduce the noise level in X-ray images. The traditional solution to reduce noise in X-rays images is to use a Wiener filter that computes a statistical estimate of the desired output image by filtering out the noise from the input image utilizing the second-order statistics of the Fourier decomposition [66, 67]. The Wiener filter minimizes the overall mean square error between the output image and the input image as in

$$\min\left(E\left[(f(x,y) - \hat{f}(x,y))^2\right]\right). \tag{9}$$

By taking the derivatives of equation (9), the Fourier transformation, $\hat{F}(u,v)$, of the constructed image can be derived as in

$$\hat{F}(u,v) = \frac{H^*(u,v)}{H^2(u,v) + (S_\zeta/S_f) \times G(u,v)}, \tag{10}$$

where $f(x,y)$ is the original image, $\hat{f}(x,y)$ is the constructed image, $H^*(u,v)$ is the complex conjugate of the Fourier transform of the degradation filter, $H^2(u,v)$ is the momentum square of the degradation filter, S_ζ is the power spectrum of the noise, S_f is the power spectrum of the original image, $G(u,v)$ is the observation, and $(H^*(u,v)/(H^2(u,v) + (S_\zeta/S_f)))$ is the Wiener filter. The output image is just the input image multiplied by the Wiener filter.

The Bayesian estimator is an extension to Wiener filter to exploit the higher-order statistics found in the point statistics of the subband decomposition of natural images, which cannot be captured by Fourier based techniques [68]. The Bayesian estimator needs to have the probability density function of the noise and the prior probability density function of the signal and hence usually parametrization model is needed to estimate the parameters for those functions. Let y be a scalar x with additive noise n such that $y = x + n$. The least-square estimator of x as a function of y can be derived using Bayes' rule as in

$$\begin{aligned} \hat{x}(y) &= \int \mathrm{d}x P_{x|y}(x \mid y)x \\ &= \hat{x}(y) = \frac{\int \mathrm{d}x P_{y|x}(y \mid x)P_x(x)x}{\int \mathrm{d}x P_{y|x}(y \mid x)P_x(x)} \\ &= \hat{x}(y) = \frac{\int \mathrm{d}x P_n(y - x)P_x(x)x}{\int \mathrm{d}x P_n(y - x)P_x(x)}, \end{aligned} \tag{11}$$

where P_n is the *pdf* for the noise, P_x is the prior *pdf* for the signal, and the denominator is the *pdf* for the noise observations. Both probability density functions must be known to estimate the original signal x. A generalized Laplacian distribution was used by [69] as a parameterization model for these densities.

The wavelet transformation is a signal processing technique used to represent real-life nonstationary signals with high efficiency [70]. Continuous and discrete wavelet transformations are used extensively in image processing especially in microcalcification enhancement methods in mammograms [71].

Wavelet decomposes the signal into subbands using a mother wavelet function to generate other window functions. The mother wavelet function is a scaling and translation function of the form

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-a}{b}\right), \tag{12}$$

where a is the scaling factor and b is the translation parameters. The mother wavelet function is applied to the original function $f(x)$ as in

$$\left(W_{\psi}f\right)(a,b) = \left\{f, \psi_{a,b}(x)\right\} = \int f(x)\cdot \psi*_{a,b}(x)\mathrm{d}x. \tag{13}$$

Images are 2D and hence a 2D discrete wavelet transform is needed, which can be computed using 2D wavelet filters followed by 2D downsampling operations for one level decomposition [72].

Microcalcification in mammograms was detected using a wavelet transformation with supervised learning through a cost function in [73]. The cost function represented the difference between the desired output image and the reconstructed image, which is obtained from the weighted wavelet coefficient for the mammogram under consideration. Then, a conjugate gradient algorithm was used to modify the weights for wavelet coefficients to minimize the cost function. In [74], the continuous wavelet transform was used to enhance the microcalcification in mammograms. In this method, a filter bank was constructed by discretizing a continuous wavelet transform. This discrete wavelet decomposition is designed in an optimal way to enhance the multiscale structures in mammograms. The advantage of this method is that it reconstructs the modified wavelet coefficients without the introduction of artifacts or loss of completeness.

Reference [75] presented a tumor detection system for fully digital mammography that detects the tumor with very weak contrast with its background using iris adaptive filter, which is very effective in distinguishing rounded opacities regardless of how weak its contrast with the background. The filter uses the orientation map of gradient vectors. Let Q_i be any arbitrary pixel and let g be gradient vector toward the pixel of interest P; then, the convergence index can be expressed as

$$f(Q_i) = \left\{\begin{array}{ll} \cos\theta, & |g| \neq 0 \\ 0, & |g| = 0 \end{array}\right\}, \tag{14}$$

where θ is the orientation of the gradient vector g at Q_i with respect to the ith half line. The average of convergence indexes over the length PQ_i, i.e., C_i is calculated by

$$C_i = \frac{\int_P^{Q_i} f(Q)\mathrm{d}Q}{PQ_i}. \tag{15}$$

The output of the iris filter $C(x,y)$ at the pixel (x,y) is given by equation (16), where C_{im} is the maximum convergence degree deduced from equation (15):

$$C(x,y) = \frac{1}{N}\sum_{i=0}^{N-1} C_{im}. \tag{16}$$

Reference [54] lists a number of other contrast enhancement algorithms with their advantages and limitations. For example, manual intensity windowing is limited by its operator skill level. Histogram-based intensity windowing has the advantage of improving the visibility of the lesion edge but at the expense of losing the details outside the dense area of the image. Mixture-model intensity windowing enhances the contrast between the lesion borders and the fatty background but at the expense of losing mixed parenchymal densities near the lesion. Contrast-limited adaptive histogram equalization improves the visibility of the edges but at the expense of increasing noise. Unsharp masking improves visibility for lesion's borders but at the expense of misrepresenting indistinct masses as circumscribed. Peripheral equalization represents the lesion details well and keeps the peripheral details of the surrounding breast but at the expense of losing the details of the nonperipheral portions of the image. Trex processing increases visibility for lesion details and breast edges but at the expense of deteriorating image contrast.

3. Feature Selection and Feature Extraction

Pattern, as described in [76], is the opposite of chaos, i.e., regularities. Pattern recognition is concerned with automatic discovering of these regularities in data utilizing computer algorithms in order to take action like classification under supervised or unsupervised setup [77, 78]. Pattern recognition has been studied in various frameworks but the most successful framework is the statistical framework [79–82]. A good reference for discussing statistical tools for features selection and features extraction can be found in [83]. In a statistical framework, a pattern is described by a vector of d features in d-dimensional space. This framework aims to reduce the number of features used to allow the pattern vector, which belongs to different categories, to occupy compact and disjoint regions in m-dimensional feature space to improve classification, stabilize representation, and/or to simplify computations [78]. A preprocessing and contrast step, which includes outlier removal, data normalization, handling of missing data, and enhancing contrast, is usually performed before the selection and extraction step [84]. The effectiveness of the selection step is measured by how successful the different patterns can be separated [76]. The decision boundaries between the patterns are determined by the probability distributions of the patterns belonging to the corresponding class. These probability distributions can be either provided or learned [85, 86]. Because selecting an optimal subset of features is done offline, having an optimal subset of features is more important than execution time [85].

The selection step involves finding the most useful subset of features that best classifies the data into the corresponding categories by reducing the d-dimensional features vector into an m-dimensional vector such that $m \leq d$ [84]. This can be done by features selection in measurement space (i.e., features selection) or transformation from the measurements to lower-dimensional feature space (i.e., features extraction). Features extraction can be done through a linear or nonlinear combination of the features and can be done under supervision or no supervision [83]. The most important features that are usually extracted from the mammograms are *spectral* features, which correspond to the variations in the quality of color and tone in an image; *Textural* features, which describe the spatial distribution of the color and tone within an image, and *contextual* features, which contain the information from the area surrounding the interest region [67]. Textural features are very useful in mammograms and can be classified into fine, coarse, or smooth, rippled, molled, irregular, or lineated [67]. Several linear and nonlinear features extraction techniques based on textural features are used in mammograms analysis such as in [72, 87–95].

Considering M classes and corresponding feature vectors distributed as $p(x \mid w_i)$, the parametric likelihood functions, and the corresponding parameters vectors θ_i, we can find the corresponding probability density function $p(x \mid \theta_i)$. A maximum log-likelihood estimator, or other methods like Bayesian inference, can be used to estimate the unknown parameters giving the set of known feature vectors. The expected maximization algorithm can be used to handle missing data. Having the probability density functions for the data available, we can extract meaningful features from them.

Gray-level cooccurrence matrix (GLCM) proposed by [67] is a well-established method for texture features extraction and is used extensively in the literature [80, 96–104]. Fourteen features were extracted from GLCM by the same author who originally proposed it [105]. The basic idea of the GLCM is as follows: let I be an N-greyscale level image, and then the gray-level cooccurrence matrix G for I is an N square matrix with its entries being defined as the number of occasions a pixel with intensity i is adjacent (on its vertical, horizontal, right, or left diagonals) to a pixel with intensity j. The features are calculated on each possible combination of adjacency and then the average is taken. G can be normalized by dividing each element of G by the total number of cooccurrence pairs in G. For example, consider the following gray-level image:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 2 \\ 1 & 1 & 0 & 1 & 1 \\ 2 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{17}$$

If we apply the rule "1 pixel to the right and 1 pixel down," the corresponding gray-level cooccurrence matrix is

$$C = \frac{1}{16}\begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 2 \\ 0 & 2 & 0 \end{bmatrix}. \tag{18}$$

For example, the first entry comes from the fact that there are 4 occasions where a 0 appears below and to the right of another 0, whereas the normalization factor $(1/16)$ comes from the fact that there are 16 pairs entering into this matrix.

The 14 Haralick's texture features are contrast, correlation, sum of squares, homogeneity, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measure of correlation 1, information measure of correlation 2, and maximum correlation coefficient. Probably the most used are energy, entropy, contrast, homogeneity, and correlation [105].

Let $P(i, j, d, \theta)$ denote the probability of how often one gray-tone i in an N-greyscale level image will appear in a specified spatial relationship determined by the direction θ and the distance d to another gray-tone j in a mammogram. The probability can be calculated as

$$P(i, j, d, \theta) = \frac{C(i, j)}{\sum_{i,j=0}^{N-1} C(i, j)}, \tag{19}$$

where $C(i, j)$ are the values in cell (i, j).

This definition will be used to define five of the most commonly used measures of the 14 Haralick's texture features: energy, entropy, contrast, homogeneity, and correlation.

Energy, also known as angular second moment, measures the homogeneity of the image such that if the texture of the image is uniform, there will be very few dominant gray-tone transitions and hence the value of energy will be high. The energy can be calculated as

$$\text{energy} = \sum_{i,j=0}^{N-1} P^2(i, j, d, \theta). \tag{20}$$

Entropy measures the nonuniformity in an image or complexity of an image. Entropy is strongly but inversely correlated to energy and can be calculated from the second-order histogram as

$$\text{entropy} = -\sum_{i,j=0}^{N-1} P(i, j, d, \theta) \times \log P(i, j, d, \theta). \tag{21}$$

Contrast measures the variance of the gray level in the image, i.e., the local gray-level variations present in an image. It detects disorders in textures. For smooth images, the contrast value is low, and for coarse images, the contrast value is high.

$$\text{contrast} = \sum_{i,j=0}^{N-1} (i - j)^2 \times P(i, j, d, \theta). \tag{22}$$

Homogeneity, which is also known as inverse difference moment, is a measure of local homogeneity and it is inversely related to contrast such that if contrast is low, the homogeneity is high. Homogeneity is calculated as

$$\text{homogeneity} = \sum_{i,j=0}^{N-1} \frac{P(i, j, d, \theta)}{i + |i - j|^2}. \tag{23}$$

Correlation is used to measure the linear dependencies in the gray-tone level between two pixels.

$$\text{Corr} = \sum_{i,j=0}^{N-1} \frac{(i - \mu_x) \times (j - \mu_y) \times P(i, j, d, \theta)}{\sigma_x \times \sigma_y}, \tag{24}$$

where μ_x is the mean value of pixel intensity i, σ_x is the standard deviation value of pixel intensity i, μ_y is the mean value of pixel intensity j, and σ_y is the standard deviation value of pixel intensity j.

Figure 2 shows the mdb028 mammogram (a) from the MIAS database [106] and the corresponding region of interest (b).

The corresponding contrast, correlation, energy, entropy, and homogeneity values calculated based on the GLCM matrix are 0.0813, 0.9617, 0.2656, 6.53, and 0.9594, respectively. One can see that this image has high homogeneity and low contrast as expected.

Gradient-based methods are widely used in analyzing mammograms [107–109]. The basic idea of the gradient is as follows: let Y be an $\mathbb{R}$-valued random variable, and then

$$\frac{\partial}{\partial x}E[Y \mid X = x] = B\int y\frac{\partial}{\partial z}\tilde{p}(y \mid z)\Big|_{z=B^T x}\mathrm{d}y, \tag{25}$$

where (X, Y) is a random vector such that $X = (X^1, \ldots, X^m) \in \mathbb{R}^m$, and B is a projection matrix, such that $B^T B = I_d$, into d-dimensional subspace such that $d \leq m$. Equation (25) implies that the gradient $\partial/\partial x E[Y \mid X = x]$ at any x is contained in the effective direction for regression.

Traditional gradient methods often could not reveal a clear-cut transition or gradient information because malignant lesions usually fill a large area in the mammogram [110]. This limitation was addressed by [95] where directional derivatives were used to measure variations in intensities. This method is known as acutance, A, and is calculated as

$$A = \frac{\sum_{i=1}^{N}\sqrt{\sum_{j=0}^{n_i=1}\left(f_i(j) - f_i(j+1)\right)^2}}{f_{\max} - f_{\min}}, \tag{26}$$

where $f_{\max}$ and $f_{\min}$ are the local maximum and minimum pixel values in the region under consideration, respectively, N is the number of pixels along the boundary of the region, and $f_i(j), j = 0, 1, \ldots, n_i$ are $(n_i + 1)$ number of perpendicular pixels available at the ith boundary point including the boundary point [110].

The evaluation of the spiculations of the tumor's edges through pattern recognition techniques is widely used among scholars to classify masses into malignant and benign [87, 89, 111–114]. Morphological features can help in distinguishing between benign and malignant masses. Benign masses are characterized by smooth, circumscribed, macrolobulated, and well-defined contours, while malignant masses are vague, irregular, microlobulated, and spiculated contours. Based on these morphological features of the mass, scholars defined certain measures and indicators to classify the masses into benign and malignant like the degree of compactness, the spiculation index, the fractional factor, and fractal dimension [115].

Differential analysis is also used to compare the prior mammographic image with the most current one to find if the suspicious masses have changed in size or shape. The relative gray level is also compared between the old mammogram and the current one to deduce the changes in the breast since the last mammogram by comparing the cumulative histograms of prior and current images [90]. The bilateral analysis is also used to compare the left and right mammograms to see any unusual differences between the left and right breasts [116].

The classification metrics for a classifier depends on the interrelation between sample size, number of features, and type of classifier [78]. For example, in naïve table-lookup, the number of training data points increases exponentially with the number of features [117]. This phenomenon is called the curse of dimensionality and it leads to another phenomenon. As [78] argued that as long as the number of training samples is arbitrarily large and representative of the underlying densities, the probability of misclassification of a decision rule does not increase as the number of features increases because under this condition the class-conditional densities are completely known. In practice, it has been observed that adding features will degrade the metrics of the classifier if the size of the training data used is small compared to the number of features. This inconsistent behavior is known as the peaking phenomenon [118–120]. The peaking phenomenon can be explained as follows: most of the parametric classifiers estimate the unknown parameters for the classifier and then plug them into the class-conditional densities. At the same sample size, as the number of features increases (consequently the number of unknown parameters increases), the estimation of the parameters degrades, and consequently, this will degrade the metrics of the classifier [78]. Because of the curse of dimensionality and peaking phenomena, features selection is an important step to enhance the overall metrics of the classifier. Many methods have been discussed in the literature for features selection [78, 83]. Class separability measure can help in a deep understanding of the data and in determining the separability criterion of various features classes along with suggestions for the appropriate classification algorithms.

Class separability measures are based on conditional probability. Given two classes C_i and C_j and a features vector v, C_i will be chosen if the ratio between $P(C_i \mid v)$, and $P(C_j \mid v)$ is more than 1. The distance D_{ij} between C_i and C_j can be calculated as

$$D_{ij} = \int_{-\infty}^{\infty} p(v \mid C_i)\ln\frac{p(v \mid C_i)}{p(v \mid C_j)}\mathrm{d}v. \tag{27}$$

And D_{ji} can be calculated in the same way. The total distance d_{ij} is a measure of separability for multiclass problems. This distance is referred to as divergence or Kullback-Leibler distance measure. Kullback-Leibler distance can be calculated as discussed in [84, 121] as

$$d_{ij} = D_{ij} + D_{ji} = \int_{-\infty}^{\infty}\left(p(v \mid C_i) - p(v \mid C_j)\right)\ln\frac{p(v \mid C_i)}{p(v \mid C_j)}\mathrm{d}v. \tag{28}$$

Another well-known distance is Mahalanobis distance. Consider two Gaussian distributions with equal covariance matrices; then, the Mahalanobis distance is calculated as

$$d_{ij} = \left(\mu_i - \mu_j\right)^T \Sigma^{-1}\left(\mu_i - \mu_j\right), \tag{29}$$

where Σ is the covariance matrix for the two Gaussian distributions and μ_i and μ_j are their two means.

Class separability measures are important if many features are used. If up to 3 features, the analyst can see the class

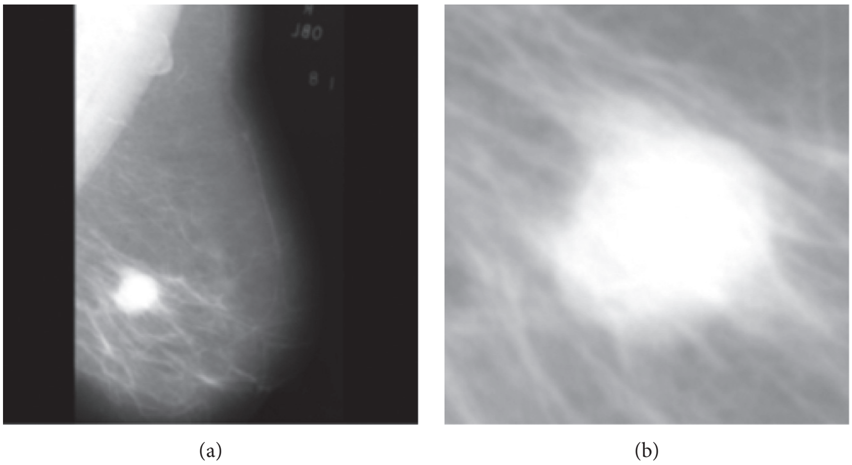
(a)
(b)

Figure 2: (a) Original mdb028 mammogram for a malignant patient. (b) The corresponding region of interest.

scatter. Three class separability measures are widely used: class scatter measure, Thornton's separability index, and direct class separability measure [122].

In class scatter measure, an unbounded measure J is defined as the ratio of the *between*-class scatter and the *within*-class scatter such that the larger the value of J is, the smaller the within-class scatter compared to the between-class scatter is. J can be calculated as

$$J = \frac{\sum_{i=1}^{C} (m_i - m)^t (m_i - m)}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \left(x_{ij} - m_i\right)^t \left(x_{ij} - m_i\right)}, \tag{30}$$

where C is the number of classes, n_i is the number of instances in class i, m_i is the mean of instances in class i, m is the overall mean of all classes, and x_{ij} is the jth instance in class i.

Separability index (SI) reports the average number of instances that share the same class label as their nearest neighbors. SI is calculated as

$$\mathrm{SI} = \frac{\sum_{i=1}^{n} \left(f\left(x_i\right) + f\left(\acute{x}_i\right) + 1\right) \bmod 2}{n}. \tag{31}$$

Direct class separability DCSM takes into consideration the compactness of the class compared to its distance from the other class. DCSM can be calculated as

$$\mathrm{DCSM} = \left[\sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \left\| x_i - x_j \right\| - \sum_{i=1}^{n_i} \sum_{i=1}^{n_i} \left\| x_i - x_j \right\| \right], \tag{32}$$

where x_i and x_j are the instances in classes i and j, respectively.

Class separability measures aim to choose the best set of features to increase the metrics of the classifier. Without this insight choice of features, two different datasets can look alike if the features were selected in the wrong way. This phenomenon is known as Ugly Duckling Theorem [123]. Many methods are used in literature to select features and can be categorized into three main methods: filter methods, wrapper methods, and hybrid methods (which is composed of both the filter and wrapper methods).

Filter methods assign ranks to the features to denote how useful each feature is for the classifier. Once these ranks are computed and assigned, the features set is then composed with the highest N rank features. Pearson's correlation coefficient method, as a filter method, looks at the strength of the correlation between the feature and the class of data [124]. If this correlation is strong, then this feature will be selected as it will help in separating the data and will be useful in classification. The mutual information method is another filter method. The mutual information method measures the shared information between a feature and the class of the labeled data. If there is a lot of shared information, then this feature is an important feature to distinguish between different classes in the data [125]. The relief method looks for the separation capability of randomly selected instances. It selects the nearest-class instant and the opposite-class instance and then calculates a weight for each feature. The weights are updated iteratively with each random instance. This method is known for its low computational time with respect to other methods [126]. Ensemble with the data permutation method is another filter method; the concept is to combine many weak classifiers to give a better classifier than any of the single classifiers used. The same concept is used to features selection where many weak rankings are combined to give much better ranking [127].

The general idea in wrapper methods is to calculate the efficacy for a certain set of features and to update this set until a stopping criterion is reached. A greedy forward search is an example of wrapper methods. This method calculates the efficacy of the set of features on hand and replaces the current set with this set only if its efficacy is better than the efficacy of the current set; otherwise, it keeps the current set. The greedy forward method starts with one feature and keeps adding features one at a time. The classifier is evaluated each time a new feature is added, and only if the efficacy of the classifier improved, the feature is maintained [128]. This method does not guarantee an optimal solution but it picks up the features that work best together. The exhaustive search method is considered a wrapper method. Exhaustive search features selection method, also known as the brute force method, looks for every possible combination of features and selects the combination that gives the best metrics for the classifier [128]. Of course, this can be only done within a reasonable computational time with a small number of features or if the number of possible combinations is reduced by searching certain combinations only. To see how large the number of possible combinations could be, consider a dataset with 300 features, and then the number of possible sets of features is $2^{300} \approx 2 \times 10^{90}$ which is a huge number. To reduce this number, we may specify that the number of features that we want is 20 features, and then the number of possible sets is $300C20 = 7.5 \times 10^{30}$ (300 choice 20) which is much lower than the first case but still impractical in terms of computational time. If we are able somehow to reduce the 300 features into 50 features and we want to choose a set of 20 features out of them, the number of sets is $50C20 = 2.7 \times 10^{13}$ which is still a very huge number. This illustration shows that the exhaustive search method for features selection works only when the method is highly constrained.

The principal component analysis is a widely used method for dimensionality reduction. The principal component analysis is a statistical procedure used to transfer a set of possibly correlated features into a set of linearly uncorrelated features by orthogonal transformation using eigenvalue decomposition on covariance matrices of the observed regions to determine their principal components. The transformation is carried out such that the first principal component has the highest variance, which means that it accounts for the largest amount of variability in the data, and the second principal component has the second highest variance under the constraint that it is orthogonal to all other components and so on. Principal component analysis reveals the internal structure of the data. It reveals how important each feature is in explaining the variability in the data. It simply shows the higher dimensional data space onto a shadow of lower-dimensional data space [127, 129–133].

Another method in dimensionality reduction is factor analysis. Factor analysis is a statistical method that describes variability in correlated observed factors (features) in terms of a lower number of unobserved new factors (new features). The principal component analysis is often confused with factor analysis. In fact, the two methods are slightly different. The principal component analysis does not involve any new features and it only ranks the features according to their importance in describing the data while the factor analysis involves creating new features by replacing a number of correlated features with a linear combination of them to create a new feature that does not exist originally [127].

4. Classification

Classification is the process of categorizing observations based on a training set of data. Classification predicts the value of a categorical variable, i.e., class of the observation, based on categorical and/or numerical variables, i.e., features. In mammograms, classification is used to predict the type of mass based on the extracted set of features. Classification algorithms can be grouped into four main groups according to their ways of calculations: frequency table based, covariance matrix based, similarity functions based, and others.

ZeroR classifier is the simplest type of frequency table classifiers that ignores the features and does classification based on the class only. The class of any observation is always the class of the majority [134, 135]. The ZeroR classifier is usually used as a baseline for benchmarking with other classifiers.

OneR classifier algorithm is another type of frequency table classifier. It generates a classification rule for each feature based on the frequency and then selects the feature that has the minimum classification error. This method is simple to construct and its accuracy is sometimes comparable to the more sophisticated classifiers with the advantage of easier results interpretation [136, 137].

Naïve Bayesian (NB) classifier is also a frequency classifier based on Bayes' theorem with a strong independence assumption between the features. NB classifier is especially useful for large datasets. Its performance sometimes outperforms the performance of the more sophisticated classifiers as discussed in [138–141]. Unlike ZeroR classifier that does not use any features in the prediction and OneR classifier that uses only one feature, the NB classifier uses all the features in the prediction.

NB classifier calculates the posterior probability of the class c given the set of features $X = \{x_1, x_2, \ldots, x_n\}$ as

$$P(c \mid x) = P\left(x_1 \mid c\right) \times P\left(x_2 \mid c\right) \times \cdots \times P\left(x_n \mid c\right) \times P(c), \tag{33}$$

where $P(x_i \mid c)$ is the probability of the feature x_i given the class c and $P(c)$ is the prior probability of the class [135]. Both probabilities can be estimated from the frequency table.

One problem facing the NB classifier is known as the zero-frequency problem. It happens when a combination of a feature and a class has zero frequency. In this case, one is added for every possible combination between features and classes, so no feature-class combination has a zero frequency.

The decision tree (DT) classifier is widely used in breast cancer classification [142–144]. The strength of DT is that it can be translated to a set of rules directly by mapping from the root nodes to the leaf nodes one by one and hence the decision-making process is easy to interpret. DT can be built based on a frequency table. It develops decision nodes and leaf nodes by repetitively dividing the dataset into smaller and smaller subsets until a stopping creation is reached or a pure class with single entry is reached. DTs can handle both categorical or numerical data. The core algorithm for DTs is called ID3, which is a top-down, greedy search algorithm with no backtracking that uses entropy and information gain to divide the subset with dependency assumption between the features [145].

Entropy was introduced in the context of features selection as one of Haralick's texture features earlier. In the ID3 algorithm, entropy has the same meaning as before and it is a measure of homogeneity in the sample. It has the same basic formula of

$$E(S) = -\sum_{x\in X} P(i) \times \log P(i), \tag{34}$$

where S is the current dataset under consideration which changes at each step, X is the set of classes in S, and $P(i)$ is the probability of class i, which can be estimated from the frequency table based on the dataset as the proportion of the number of elements in class x to the number of elements in S. A zero entropy for a dataset indicates a perfect classification.

Information gain is the change of entropy after a dataset is split based on a certain feature, i.e., the reduction of uncertainty in S after splitting it using the feature.

$$\mathrm{IG}(S,F) = E(S) - \sum_{t\in T} p(t)E(t) = E(S) - E(S \mid F), \tag{35}$$

where T is the subsets created by splitting S with F such that $S = \cup_{t\in T} t$ and $p(t)$ is the cardinality of t divided by the cardinality of S.

Overfitting is a significant problem in DTs. Overfitting is the problem of enhancing the prediction based on the training data on the expense of the prediction based on the test data. Prepruning and postpruning are used to avoid overfitting in DTs. In prepruning, the algorithm is stopped earlier before it classifies the training set perfectly while, in postpruning, the algorithm is allowed to perfectly classify the training data but then the tree is postpruned. Postpruning is more successful than prepruning because it is hard to know when exactly to stop the growth of the tree [146].

Linear Discriminant Analysis (LDA) is widely used in analyzing mammograms [51, 147] for breast cancer. LDA is a simple classifier that sometimes produces classification that is as good as the classification of the complex classifiers. It searches for a linear combination Z of features X that best separates two classes c_1 and c_2 such that

$$Z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d, \tag{36}$$

where β_i is the coefficient corresponding to feature i and $i = 1, 2, \ldots, d$ and d is the number of features. The coefficients are determined such that the score function $S(\beta)$ in equation (37) is maximized.

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta}, \tag{37}$$

where β is a vector of coefficients for the linear model given in equation (36) and can be calculated as

$$\beta = C^{-1}(\mu_1 - \mu_2), \tag{38}$$

where μ_1 and μ_2 are the mean vectors of the two classes, and C is pooled covariance matrix given as

$$C = \frac{1}{n_1 + n_2}(C_1 n_1 + C_2 n_2), \tag{39}$$

where C_1, n_1, C_2, and n_2 are covariance matrix for the first class, the number of elements in the first class, the covariance matrix for the second class, and the number of elements for the second class, respectively. A new point x is classified as C_1, i.e., class 1, if the inequality (40) stands:

$$\beta^T\left(x - \left(\frac{\mu_1 + \mu_2}{2}\right)\right) > -\log\frac{P(C_1)}{P(C_2)}, \tag{40}$$

where $P(C_1)$ is the first-class probability and $P(C_2)$ is the probability of the second class. These probabilities can be estimated from the data.

The logistic regression classifier is another covariance classifier that is used to analyze mammograms for breast cancer prediction [148–151]. It can be used only with binary classification where there are only two classes just like classifying the masses into benign or malignant in a mammogram. It uses categorical and/or numerical features to predict a binary variable (the class either 0 or 1). Linear regression is not appropriate to predict a binary variable because the residuals will not be normal and the linear regression may predict values outside the permissible range, i.e., 0 to 1 while logistic regression can only produce values between 0 and 1 [152].

Logistic regression uses the natural logarithm of the odds of the class variable. The logistic regression equation is written in terms of the odd ration as in

$$\frac{p}{1-p} = \exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n), \tag{41}$$

where p is the logistic model predicted probability, and b_0, $b_1, \ldots, b_n$ are the estimations of the coefficients in the logistic regression model for the n features, i.e., x's [151, 153]. The estimation of the model coefficients is carried out using maximum likelihood estimation. The predicted probability p by the logistic model can be calculated as

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}. \tag{42}$$

One way to do classification is to calculate p for the data instance, and if its probability is below 0.5, it will be assigned to class 1, and if it is above or equal to 0.5, it will be assigned to class 2.

K nearest neighbors (KNN) classifier is used in literature to diagnose mammograms [154–157]. This classifier is a type of majority vote and a nonparametric classifier based on a similarity function. It stores all available cases and then classifies a new data instance based on its similarity to other points in the nearest K classes measured by distance. If $K = 1$, then the new data instance will be assigned to the nearest neighbor's class. Generally speaking, increasing the number of classes, i.e., the value of K, increases the precision as it reduces the overall noise. Cross-validation is one way to determine the best value of K by using an independent dataset to validate the value of K. Also, the cross-validation technique can reduce the variance in the test error estimate calculations. A good practice is to have K between 3 and 10.

There are three well-known distance functions used in KNN classifier [158] for continuous features: Euclidean, Manhattan, and Minkowski distance functions. Their equations are as in

$$\begin{aligned}\text{Euclidean} &= \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2},\\ \text{Manhattan} &= \sum_{i=1}^{k}|x_i - y_i|,\\ \text{Minkowski} &= \left(\sum_{i=1}^{k}\left(|x_i - y_i|\right)^q\right)^{1/q},\end{aligned} \tag{43}$$

where k is the number of features, x_i is the value of feature i for object x, y_i is the value of feature i for object y, and q is the order of the Minkowski metric.

These three distances are valid for continuous features only. If the features are categorical, the Hamming distance given in equation (44) is used, which basically measures the number of mismatches between two vectors.

$$\text{Hamming} = \sum_{i=1}^{k} 1_{x_i \neq y_i}. \tag{44}$$

If features are mixed, then numerical features should be standardized between 0 and 1 before the distance is calculated.

There are three types of classifiers that are not based on the frequency table, covariance matrix, or similarity functions. These classifiers are support vector machines, artificial neural networks, and recently deep learning.

Support vector machines (SVM) classifier is first proposed by [159] and is used extensively in breast cancer detection and diagnosis using mammograms [160–166].

A linear SVM classifies a linearly separable data by constructing a linear hyperplane in N-dimensional feature space (N is the number of features) to maximize the margin distance between two classes [160]. Figure 3 shows a linear hyperplane for 2-dimensional feature space, the support vectors, and the marginal width for a linear SVM [167].

If the data is not linearly separable, it is mapped into higher-dimensional feature space by various nonlinear mapping functions like sigmoid and radial basis functions. The strength of the SVM classifier is that it does not need to have a priori density functions between the input and the output like some other classifiers and this is very important because, in practice, these prior densities are not known and there are not enough data to estimate them precisely.

The linear SVM classifier uses the training data to find the weight vector $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ and the bias b for the decision function [161, 168] in

$$d(\mathbf{X}, \mathbf{w}, b) = \sum_{i=1}^{n} w_i x_i + b. \tag{45}$$

The optimal hyperplane is the hyperplane that satisfies $d(\mathbf{X}, \mathbf{w}, b) = 0$.

In the testing phase, a vector y is created such that

$$y = \operatorname{sign}(d(\mathbf{X}, \mathbf{w}, b)). \tag{46}$$

Equation (46) is used to classify a new point $\mathbf{X}_{\text{new}}$ such that if $y(\mathbf{X}_{\text{new}})$ is positive, then $\mathbf{X}_{\text{new}}$ belongs to class 1 and to class 2 otherwise.

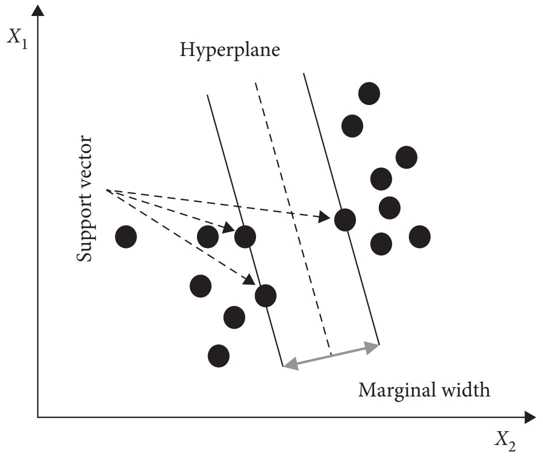

FIGURE 3: Support vectors, hyperplane, and marginal width with SVM [167].

The weight vector $\mathbf{w}$ and the bias b are found by minimizing the following model:

$$\begin{aligned}&L_d(\alpha) = 0.5\alpha^T H\alpha - f^T\alpha,\\ &\text{Subject to } y^T\alpha = 0,\\ &\alpha \geq 0,\end{aligned} \tag{47}$$

where H is the Hessian matrix given by

$$H = y_i y_j \left(x_i x_j\right), \tag{48}$$

and f is a unit vector. The values of α_{0i} can be determined by solving the dual optimization problem in equation (47). These values are used to find the values of $\mathbf{w}$ and b as follows:

$$\begin{aligned}\mathbf{w} &= \sum_{i=1}^{l} \alpha_{0i} y_i x_i,\\ b &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{y_i} - x_i^T\mathbf{w}\right),\end{aligned} \tag{49}$$

where N is the number of support vectors.

As mentioned before, for nonlinearly separable data, the data has to be mapped to a higher-dimensional feature space first using a suitable nonlinear mapping function $\varphi(x)$. A kernel function $K(x_i, x_j)$ that maps the data into a very high-dimensional feature space is defined as

$$K\left(x_i, x_j\right) = \varphi(x_i)^T \varphi\left(x_j\right), \tag{50}$$

and the hyperplane is defined as

$$d(\mathbf{X}) = \sum_{i=1}^{l} y_i \alpha_i K(x_i, \mathbf{X}). \tag{51}$$

The SVM produced in model (47) is called a hard margin classifier. Soft margin classifier can be produced with the same model (47) but with adding an additional constraint $0 \leq \alpha_i \leq C$, where C is defined by the user. Soft margin SVM is preferred over the hard SVM to preserve the smoothness of the hyperplane [161].

One other classifier used extensively in detecting cancer in mammograms is the artificial neural network (ANN). This classifier is imitating the biological neural network, such as the brain. The biological neural network consists of a tremendous amount of connected neurons through a junction called synapses. Each neuron is connected to thousands of other neurons and receives signals from them. If the sum of these signals exceeds a certain threshold, a response is sent through the axon. ANN imitates this setup. In an ANN, the neurons are called nodes and these nodes are connected to each other. The strength of the connections is represented by weights such that the weight between two nodes represents the strength of the connection between them. Figure 4 shows the generic structure for ANN in mammography where the network receives the features at the input nodes and provides the predicted class at the output node [169].

The inhabitation occurs when the weight is -1 and the excitation occurs when the weight is 1. Within each node's design, a transfer function is introduced [169]. The most used transfer functions are a unit step function, sigmoid function, Gaussian function, linear function, and a piecewise linear function. ANN usually has three layers of nodes: an input layer, a hidden layer, and an output layer.

ANNs have certain traits that make them suit breast cancer detection and diagnosis using mammograms. They are capable of learning complicated patterns [170, 171], they can handle missing data [172], and they are accurate classifiers [173–175]. In breast cancer detection and diagnosis using mammograms, the nodes of the input layer usually represent the features extracted from the region of interest (ROI) and the node in the output layer represents the class (either malignant or benign). The nodes of the input layer receive activation values as numeric information such that the higher the information, the greater the activation. The activation value is passed from node to node based on the weights and the transfer function such that each node sums the activation values that it receives and then modifies the sum based on its transfer function. The activation spread out in the network from the input layer nodes to the output layer node through the hidden layer where the output node represents the results in a meaningful way. The network learns through gradient descent algorithm where the error between the predicted value and the actual value is propagated backward by apportioning them to each node's weights according to the amount of this error the node is responsible for [176].

A deep learning (DL) or hierarchical learning classifier is a subset of machine learning that uses networks to simulate humanlike decision making based on the layers used in ANN. Unlike other machine learning techniques discussed until now, DL classifiers do not need features selection and extraction step as they adaptively learn the appropriate features extraction process from the input data with respect to the target output [177]. This is considered a big advantage for DL classifiers as the features selection and extraction step is challenging in most cases. For an image classification problem, the DL classifier needs three things to work properly: a large number of labeled images, neural network structure with many layers, and high computational power. It can reach high classification accuracy [178]. The most common type of DL architecture used to analyze images is Convolution Neural Network (CNN).

Different types of CNN were proposed recently to deal with breast cancer detection and diagnosis using mammograms problem [20, 166, 179–186]. For example, the VGG16 network is a deep CNN used to detect and diagnose lesions in mammograms. VGG16 consists of 16 layers with the final layer capable of detecting two kinds of lesions (benign and malignant) in the mammogram [186, 187]. VGG16 encloses each detected lesion with a box and attaches a confidence level in the predicted class for each detected lesion. Faster R-CNN is also a deep CNN used in breast cancer detection and diagnosis using mammograms. The basic Faster R-CNN is based on a convolutional neural network with an additional layer on the last convolutional layer called Region Proposal Network to detect, localize, and classify lesions. It uses various boxes with different sizes and aspect ratios to detect objects with different sizes and shapes [185]. A fast microcalcification detection and segmentation procedure utilizing two CNNs was developed in [186]. One of the CNNs was used for quick detection of candidate regions of interest and the other one was used to segment them. A context-sensitive deep neural network (DNN) is another CNN for detecting and diagnosing breast cancer using mammograms [188]. DNN takes into consideration both the local image features of a microcalcification and its surrounding tissues such that the DNN classifier automatically extracts the relevant features and the context of the mammogram. Handcraft descriptors and deep learning descriptors were used to characterize the microcalcification in mammograms [189]. The results showed that the deep learning descriptors outperformed the handcraft features. Pretrained ResNet-50 architecture and Class Activation Map technique along with Global Average Pooling for object localization were used in [190] to detect and diagnose breast cancer in mammograms. The results showed an area under the ROC of 0.96. A recent comprehensive technical review on the convolutional neural network applied to breast cancer detection and diagnosis using mammograms is found in [191].

To overcome the problem of overfitting in machine learning techniques such as DL and CNN, data augmentation techniques were used to generate artificial data by applying several transformations techniques to the actual data such as flipping, rotations, jittering, and random scaling to the actual data. Data augmentation is a very powerful method for overcoming overfitting. The augmented data represents a more complete set of data points. This will minimize the variance between the training and validation sets and any future testing sets. Data augmentation has been used in many studies along with DL and CNN such as [192–196].

It is well established among the scholars who work on the problem of breast cancer detection and diagnosis using mammograms and on classification problems in general that there is no "one size fits all" classifier. The classifier who is trained on a certain dataset and certain features space may not work with the same efficacy on other datasets. This

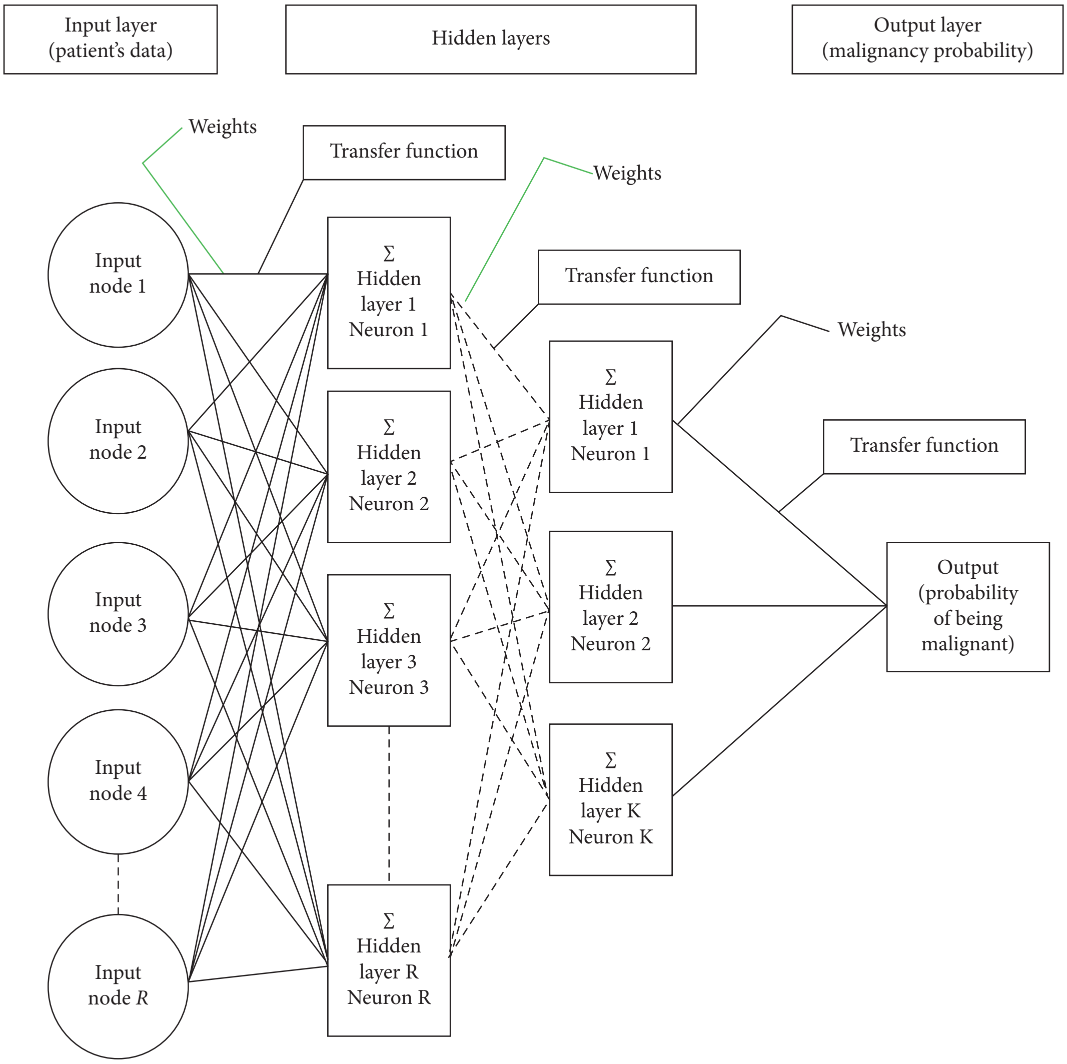
Input layer
(patient's data)
Hidden layers
Output layer
(malignancy probability)
Weights
Transfer function
Weights
Transfer function
Weights
Transfer function
Input node 1
Input node 2
Input node 3
Input node 4
Input node R
Σ Hidden layer 1 Neuron 1
Σ Hidden layer 2 Neuron 2
Σ Hidden layer 3 Neuron 3
Σ Hidden layer R Neuron R
Σ Hidden layer 1 Neuron 1
Σ Hidden layer 2 Neuron 2
Σ Hidden layer K Neuron K
Output (probability of being malignant)

Figure 4: Structure of ANN for typical breast cancer detection using mammogram [169].

problem is rooted in the "No Free Lunch Theorem" coined in [197]. "No Free Lunch Theorem" showed that there are no a priori differences between learning algorithms when it comes to an off-training-set error in a noise-free scenario where the loss function is the misclassification rate [197]. Therefore, each classifier has its own advantages and disadvantages based on the feature space and dataset used for learning [198]. For example, the features come in different representations like continuous, categorical, or binary variables and the features may have different physical meanings like energy, entropy, or size. Lump sums these diverse features into one features vector and then uses a single classifier that requires normalizing these features first, which is a tedious job. It may be easier to aggregate those features that share the same characteristics in terms of representation and physical meaning into several homogenous features vectors and then apply a different classifier to each vector separately. Moreover, even if the features are homogeneous in terms of representation and physical meaning but the number of features is large with a small number of training data points, then the estimation of the classifier parameters degrades as a result of the curse of dimensionality and peaking phenomena discussed earlier [118–120].

Table 1: The area under the ROC for some common classification techniques for mammograms.

Method	Area under ROC
Binary decision tree [210]	0.90
Linear classifier [210]	0.90
PCA–LS SVM [211]	0.94
ANN [212]	0.88
Multiple expert system [213]	0.79
Texture measure with ANN [214]	0.87
Multiresolution texture analysis [215]	0.86
Subregion Hotelling observers [216]	0.94
Logistic regression [217]	0.81
KNN [218]	0.82
NB [219]	0.56
DL [190]	0.96
Genetic algorithms with SVM [220]	0.97

Because of this, scholars can improve their classification accuracy by combining outputs from different classifiers

Table 2: List of common databases used in CAD-related techniques.

	MIAS [221]	DDSM [222]	UCSF/LLNL [223]	CALMa [224]	Banco Web [225]
Origin	UK	USA	USA	Italy	Brazil
Number of images	320	10480	198	3000	1400
File access	Free	Free	Paid	closed	Free, requires registration
Type of images	PGM	LJPEG	N/A	N/A	TIFF

through a combining schema. From an implementation point of view, combination topologies can be categorized into multiple, conditional, hierarchical, or hybrid topologies [199]. For a thorough discussion of combinational topologies, one can consult [200, 201]. A combining schema includes a rule to determine when a certain classifier should be invoked and how each classifier interacts with other classifiers in the combination [78, 202]. The majority votes and weighted majority votes are two widely used combining schemas in literature. In majority votes, all classifiers have the same vote weight and the test instance will have the class that has the highest number of votes from different classifiers, i.e., the class that is predicted by the majority of the classifiers used in the combination. In majority votes, the classifiers have to be independent [78] as it has been shown that using majority votes with a combination of dependent classifiers will not improve the overall classification performance [203]. In weighted majority votes, each classifier has its own weight that will be changed according to its efficacy such that the weight will be decreased every time the classifier has a wrong class. The test instance will be classified according to the highest weighted majority class [203, 204].

Boosting and Bagging are also used to improve the accuracy of classification results. They were used successfully to improve the accuracy of the classifiers, like the DT classifier, by combining several classification results from the training data. Bagging combines classification results from different classifiers or from the same classifier using different subsets of training data, which is usually generated by bootstrapping. The main advantage of bootstrapping is to reduce the number of training datasets used. Bootstrapping resamples the same training dataset to create different training datasets that can be used with different classifiers or the same classifier. Bagging uses bootstrapping to create different datasets from one dataset. Bagging can be viewed as a voting combining technique and it has been implemented with majority voting or weighted majority voting such that the prediction is the class that has the majority votes or the weighted majority votes from different classifiers (or same classifier) with different training subsets. Bagging was used in the context of breast cancer detection using mammograms in several manuscripts like [205, 206]. Boosting technique attaches weights to different instances of training data such that lower weights are given to instances that were frequently classified correctly and higher weights for those who were frequently misclassified; therefore, these classes will be selected more frequently in the resampling to improve their performance. This is followed by another iteration of computing weights and this sequence is repeated until a termination condition is reached. The most popular version of boosting techniques is AdaBoost (stands for Adaptive Boosting) algorithm [207–209] which classifies the data instance as a weighted sum of the output of other weak classifiers. It is considered adaptive because the weights of the weak classifiers are changed adaptively based on their performance.

Table 1 shows a list of some common classifiers and their performance measures by the area under the receiver operating characteristic ROC curve registered in the respective papers where they were proposed/used.

For the sake of completeness, Table 2 shows a list of common databases used in CAD-related techniques. It should be noticed that the acquisition protocol of these databases normally must be rigorous and they are expensive.

5. Conclusions

In this study, we shed some light on CAD methods used in breast cancer detection and diagnosis using mammograms. We reviewed the different methods used in literature in the three major steps of the CAD system, which include preprocessing and enhancement, feature extraction, and selection and classification.

Studies reviewed in this article have shown that computer-aided detection and diagnosis of breast cancer from mammograms is limited by the low contrast between normal glandular breast tissues and malignant ones and between the cancerous lesions and the background, especially in dense breasts tissue. Moreover, quantum noise also reduces mammogram quality especially for small objects with low contrast such as a small tumor in a dense breast. The presence of noise in a mammogram gives it a grainy appearance, which reduces the visibility of some features within the image especially for small objects with low contrast. A wide range of histogram equalization techniques, among other techniques, were used in many articles for image enhancement to reduce the effect of low contrast by changing the intensity of the pixels for the input image. Different filters were proposed and used to reduce noises in mammograms such as Wiener filter and Bayesian estimator.

Morphological features were widely used by scholars to distinguish between benign and malignant masses. Benign masses are characterized by smooth, circumscribed, macrolobulated, and well-defined contours, while malignant masses are vague, irregular, microlobulated, and spiculated contours. Differential analysis was also used to compare the prior mammographic image with the most current one to find if the suspicious masses have changed in size or shape. The bilateral analysis is also used to compare the left and right mammograms to see any unusual differences between the left breast and right breast.

Table 1 gives a rough estimation of the average performance of the different CAD methods used and measured as the area under the ROC curve, which is about 0.86. This performance is encouraging but still not reliable enough to accept CAD systems as a standalone clinical procedure to detect and diagnose breast cancer using mammograms. Moreover, many results that were reported in the literature with excellent performance in cancer detection using CAD systems cannot be generalized as their analyses were conducted and tuned using a specific dataset. Therefore, unless a higher performance is reached with CAD systems by exploiting new promising methods like deep learning and higher computational power systems, CAD systems can only be used as a second opinion clinical procedure.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The author is grateful to the German Jordanian University, Mushaqar, Amman, Jordan, for the financial support granted to this research.

References

[1] M. Hejmadi, *Introduction to Cancer Biology*, Bookboon, London, UK, 2nd edition, 2010.

[2] American Cancer Society, *Breast Cancer Facts and Figures 2017-2018*, American Cancer Society, Atlanta, GA, USA, 2017.

[3] American Cancer Society, *Breast Cancer Facts and Figures 2019*, American Cancer Society, Atlanta, GA, USA, 2019.

[4] O. Ginsburg, F. Bray, M. P. Coleman et al., "The global burden of women's cancers: a grand challenge in global health," *The Lancet*, vol. 389, no. 10071, pp. 847–860, 2017.

[5] J. Eric, L. M. Wun, C. C. Boring, W. Flanders, J. Timmel, and T. Tong, "The lifetime risk of developing breast cancer," *Journal of the National Cancer Institute*, vol. 85, no. 11, pp. 892–897, 1993.

[6] B. E. Sirovich and H. C. Sox, "Breast cancer screening," *Surgical Clinics of North America*, vol. 79, no. 5, pp. 961–990, 1999.

[7] R. M. Rangayyan, F. J. Ayres, and J. E. Leo Desautels, "A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs," *Journal of the Franklin Institute*, vol. 344, no. 3-4, pp. 312–348, 2007.

[8] R. L. Helms, E. L. O'Hea, and M. Corso, "Body image issues in women with breast cancer," *Psychology, Health and Medicine*, vol. 13, no. 3, pp. 313–325, 2008.

[9] "Mayo clinic," 2019, https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475.

[10] S. Njor, L. Nyström, S. Moss et al., "Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies," *Journal of Medical Screening*, vol. 19, no. 1_suppl, pp. 33–41, 2012.

[11] S. Morrell, R. Taylor, D. Roder, and A. Dobson, "Mammography screening and breast cancer mortality in Australia: an aggregate cohort study," *Journal of Medical Screening*, vol. 19, no. 1, pp. 26–34, 2012.

[12] Independent UK Panel on Breast Cancer Screening, "The benefits and harms of breast cancer screening: an independent review," *The Lancet*, vol. 380, no. 9855, pp. 1778–1786, 2012.

[13] E. D. Pisano, C. Gatsonis, E. Hendrick et al., "Diagnostic performance of digital versus film mammography for breast-cancer screening," *New England Journal of Medicine*, vol. 353, no. 17, pp. 1773–1783, 2005.

[14] P. A. Carney, D. L. Miglioretti, B. C. Yankaskas et al., "Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography," *Annals of Internal Medicine*, vol. 138, no. 3, pp. 168–175, 2003.

[15] D. B. Woodard, A. E. Gelfand, W. E. Barlow, and J. G. Elmore, "Performance assessment for radiologists interpreting screening mammography," *Statistics in Medicine*, vol. 26, no. 7, pp. 1532–1551, 2007.

[16] E. B. Cole, E. D. Pisano, E. O. Kistner et al., "Diagnostic accuracy of digital mammography in patients with dense breasts who underwent problem-solving mammography: effects of image processing and lesion type," *Radiology*, vol. 226, pp. 153–160, 2003.

[17] N. F. Boyd, H. Guo, L. J. Martin et al., "Mammographic density and the risk and detection of breast cancer," *New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[18] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology*, vol. 184, no. 3, pp. 613–617, 1992.

[19] K. Kerlikowske, P. A. Carney, B. Geller et al., "Performance of screening mammography among women with and without a first-degree relative with breast cancer," *Annals of Internal Medicine*, vol. 133, no. 11, pp. 855–863, 2000.

[20] M. G. Ertosun and D. L. Rubin, "Probabilistic visual search for masses within mammography images using deep learning," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Washington, DC, USA, November 2015.

[21] F. L. Nunes, H. Schiabel, and C. E. Goes, "Contrast enhancement in dense breast images to aid clustered microcalcifications detection," *Journal of Digital Imaging*, vol. 20, no. 1, pp. 53–66, 2007.

[22] G. Maskarinec, I. Pagano, Z. Chen, C. Nagata, and I. T. Gram, "Ethnic and geographic differences in mammographic density and their association with breast cancer incidence," *Breast Cancer Research and Treatment*, vol. 104, no. 1, pp. 47–56, 2007.

[23] H. D. Nelson, K. Tyne, A. Naik et al., "Screening for breast cancer: an update for the US preventive services task force," *Annals of Internal Medicine*, vol. 151, no. 10, pp. 727–737, 2009.

[24] M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-aided detection and diagnosis in mammography," *Handbook of Image and Video Processing*, Elsevier, London, UK, 2003.

[25] J. L. Jesneck, J. Y. Lo, and J. A. Baker, "Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors," *Radiology*, vol. 244, no. 2, pp. 390–398, 2007.

[26] J. Dinnes, S. Moss, J. Melia, R. Blanks, F. Song, and J. Kleijnen, "Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review," *The Breast*, vol. 10, no. 6, pp. 455–463, 2009.

[27] R. Warren and W. Duffy, "Comparison of single reading with double reading of mammograms, and change in effectiveness with experience," *The British Journal of Radiology*, vol. 68, no. 813, pp. 958–962, 1995.

[28] R. G. Blanks, M. G. Wallis, and S. M. Moss, "A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme," *Journal of Medical Screening*, vol. 5, no. 4, pp. 195–201, 1998.

[29] M. Posso, M. Carles, M. Rué, T. Puig, and X. Bonfill, "Cost-effectiveness of double reading versus single reading of mammograms in a breast cancer screening programme," *PLoS One*, vol. 11, no. 7, Article ID e0159806, 2016.

[30] D. Gur, J. H. Sumkin, H. E. Rockette et al., "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," *JNCI Journal of the National Cancer Institute*, vol. 96, no. 3, pp. 185–190, 2004.

[31] K. Doi, "Computer-aided diagnosis in medical imaging: achievements and challenges," in *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, Munich, Germany, 2009.

[32] C. Balleyguier, K. Kinkel, J. Fermanian et al., "Computer-aided detection (CAD) in mammography: does it help the junior or the senior radiologist?" *European Journal of Radiology*, vol. 54, no. 1, pp. 90–96, 2005.

[33] S. Sanchez Gómez, M. Torres Tabanera, A. Vega Bolivar et al., "Impact of a CAD system in a screen-film mammography screening program: a prospective study," *European Journal of Radiology*, vol. 80, no. 3, pp. e317–e321, 2011.

[34] A. Malich, T. Azhari, T. Böhm, M. Fleck, and W. Kaiser, "Reproducibility—an important factor determining the quality of computer aided detection (CAD) systems," *European Journal of Radiology*, vol. 36, no. 3, pp. 170–174, 2000.

[35] C. Marx, A. Malich, M. Facius et al., "Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of mammographic diagnosis with and without use of CAD," *European Journal of Radiology*, vol. 51, no. 1, pp. 66–72, 2004.

[36] M. L. Giger, N. Karssemeijer, and S. G. Armato, "Computer aided diagnosis in medical imaging," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1205–1208, 2001.

[37] M. L. Giger, "Computer-aided diagnosis of breast lesions in medical images," *Computer Science & Engineering*, vol. 2, no. 5, pp. 39–45, 2000.

[38] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, no. 3, pp. 781–786, 2001.

[39] M. P. Sampat, M. K. Markey, and A. C. Bovik, "Computer-aided detection and diagnosis in mammography," *Handbook of Image and Video Processing*, vol. 2, pp. 1195–1217, Academic Press, Cambridge, MA, USA, 2005.

[40] W. Zhang, K. Doi, M. L. Giger, Y. Wu, R. M. Nishikawa, and R. A. Schmidt, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Medical Physics*, vol. 21, no. 4, pp. 517–524, 1994.

[41] R. Mousa, Q. Munib, and A. Moussa, "Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural," *Expert Systems with Applications*, vol. 28, no. 4, pp. 713–723, 2005.

[42] M. Rizzi, M. D'Aloia, and B. Castagnolo, "Health care CAD systems for breast microcalcification cluster detection," *Journal of Medical and Biological Engineering*, vol. 32, pp. 147–156, 2012.

[43] M. J. Islam, M. Ahmadi, and M. A. Sid-Ahmed, "Computer-aided detection and classification of masses in digitized mammograms using artificial neural network," in *Proceedings of the Advances in Swarm Intelligence: First International Conference, ICSI 2010*, pp. 327–334, Beijing, China, 2010.

[44] E. Kozegar, M. Soryani, B. Minaei, and I. Domingues, "Assessment of a novel mass detection algorithm in mammograms," *Journal of Cancer Research and Therapeutics*, vol. 9, no. 4, pp. 592–600, 2013.

[45] A. Jalalian, S. Mashohor, R. Mahmud, B. Karasfi, M. I. B. Saripan, and A. R. B. Ramli, "Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection," *EXCLI Journal*, vol. 16, pp. 113–137, 2017.

[46] A. Oliver, J. Freixenet, J. Martí et al., "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.

[47] E. D. Pisano, M. J. Yaffe, and C. M. Kuzmiak, *Digital Mammography*, Lippincott Williams and Wilkins, Philadelphia, PA, USA, 2004.

[48] H. O. Peitgen, "Digital mammography," in *Proceedings of the IWDM 2002-6th International Workshop on Digital Mammography*, Springer-Verlag, Bremen, Germany, 2003.

[49] K. Doi, H. MacMahon, S. Katsuragawa, R. M. Nishikawa, and Y. Jiang, "Computer-aided diagnosis in radiology: potential and pitfalls," *European Journal of Radiology*, vol. 31, no. 2, pp. 97–109, 1999.

[50] C. J. Vyborny, M. L. Giger, and R. M. Nishikawa, "Computer aided detection and diagnosis of breast cancer," *Radiologic Clinics of North America*, vol. 38, no. 4, pp. 725–740, 2000.

[51] Z. Suhail, E. R. E. Denton, and R. Zwiggelaar, "Classification of micro-calcification in mammograms using scalable linear fisher discriminant analysis," *Medical & Biological Engineering & Computing*, vol. 56, no. 8, pp. 1475–1485, 2018.

[52] J. Scharcanski and C. R. Jung, "Denoising and enhancing digital mammographic images for visual screening," *Computerized Medical Imag. Graphics*, vol. 30, no. 4, pp. 243–254, 2006.

[53] The web-based edition of the physical principles of medical imaging," 2019, http://www.sprawls.org/ppmi2/.

[54] E. D. Pisano, E. B. Cole, B. M. Hemminger et al., "Image processing algorithms for digital mammography: a pictorial essay," *Radiographics*, vol. 20, no. 5, pp. 1479–1491, 2000.

[55] N. Jamal, K. H. Ng, and D. McLean, "A study of mean glandular dose during diagnostic mammography in Malaysia and some of the factors affecting it," *The British Journal of Radiology*, vol. 76, no. 904, pp. 238–245, 2003.

[56] N. Jamal, K.-H. Ng, D. McLean, L.-M. Looi, and F. Moosa, "Mammographic breast glandularity in Malaysian women derived from radiographic data," *American Journal of Roentgenology*, vol. 182, pp. 713–717, 2004.

[57] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 2nd edition, 2002.

[58] Y.-T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Transactions on Consumer Electronics*, vol. 43, pp. 1–8, 1997.

[59] C. Zuo, Q. Chen, and X. Sui, “Range limited bi-histogram equalization for image contrast enhancement,” *Optik*, vol. 124, no. 5, pp. 425–431, 2013.

[60] Y. Wang, Q. Chen, and B. Zhang, “Image enhancement based on equal area dualistic subimage histogram equalization method,” *IEEE Transactions on Consumer Electronics*, vol. 45, pp. 68–75, 1999.

[61] C. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[62] C. H. Ooi, N. Kong, and H. Ibrahim, “Bi-histogram equalization with a plateau limit for digital image enhancement,” *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2072–2080, 2009.

[63] C. H. Ooi and N. Isa, “Adaptive contrast enhancement methods with brightness preserving,” *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2543–2551, 2010.

[64] K. Singh and R. Kapoor, “Image enhancement via median-mean based sub-image-clipped histogram equalization,” *Optik—International Journal for Light and Electron Optics*, vol. 125, no. 17, pp. 4646–4651, 2014.

[65] S.-D. Chen and A. R. Ramli, “Minimum mean brightness error bi-histogram equalization in contrast enhancement,” *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 1310–1319, 2003.

[66] R. A. Lerski, K. Straughan, L. R. Schad, D. Boyce, S. Bluml, and I. Zuna, “MR image texture analysis—an approach to tissue characterization,” *Magnetic Resonance Imaging*, vol. 11, no. 6, pp. 873–887, 1993.

[67] R. M. Haralick, K. Shanmugan, and I. Dinstein, “Textural features for image classification,” *IEEE transactions on systems: man, and cybernetics SMC*, vol. 3, no. 6, pp. 610–621, 1973.

[68] E. P. Simoncelli and E. H. Adelson, “Noise removal via bayesian wavelet coring,” in *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 1, pp. 379–382, Lausanne, Switzerland, September 1996.

[69] S. G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.

[70] M. Mozammel Hoque Chowdhury and A. Khatun, “Image compression using discrete wavelet transform,” *IJCSI International Journal of Computer Science*, vol. 9, no. 1, 2012.

[71] A. P. Dhawan, Y. Chitre, and C. Kaiser-Bonasso, “Analysis of mammographic microcalcifications using gray-level image structure features,” *IEEE Transactions on Medical Imaging*, vol. 15, no. 3, pp. 246–259, 1996.

[72] A. E. Minarno, Y. Munarko, A. Kurniawardhani, F. Bimantoro, and N. Suciati, “Texture feature extraction using co-occurrence matrices of sub-band image for batik image classification,” in *Proceedings of the 2014 2nd International Conference on Information and Communication Technology (ICoICT)*, pp. 249–254, Bandung, Indonesia, 2014.

[73] H. Yoshida, Z. Wei, W. Cai, K. Doi, R. M. Nishikawa, and M. L. Giger, “Optimizing wavelet transform based on supervised learning for detection of microcalcifications in digital mammograms,” in *Proceedings of the International Conference Image Processing*, vol. 3, pp. 152–155, Washington, DC, USA, October 1995.

[74] P. Heinlein, J. Drexl, and W. Schneider, “Integrated wavelets for enhancement of microcalcifications in digital mammography,” *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 402–413, 2003.

[75] H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, “Computerized detection of malignant tumors on digital mammograms,” *IEEE Transactions on Medical Imaging*, vol. 18, no. 5, pp. 369–378, 1999.

[76] S. Watanabe, *Pattern Recognition: Human and Mechanical*, Wiley, New York, NY, USA, 1985.

[77] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Berlin, Germany, 2006.

[78] A. K. Jain, R. P. W. Duin, and J. C. Mao, “Statistical pattern recognition: a review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[79] M. Sameti, R. K. Ward, J. Morgan-Parkes, and B. Palcic, “Image feature extraction in the last screening mammography prior to detection of breast cancer,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 46–52, 2011.

[80] R. Hupes and N. Karssemeijer, “Use of normal tissue context in computer-aided detection of masses in mammograms,” *IEEE Transactions on Medical Imaging*, vol. 28, no. 12, pp. 2033–2041, 2010.

[81] A. Lambrou, H. Papadopoulos, and A. Gammerman, “Evolutionary conformal prediction for breast cancer diagnosis,” in *Proceedings of the IEEE 9th International Conference on Information Technology and Applications in Biomedicine*, pp. 1–4, Los Alamitos, CA, USA, 2011.

[82] X. Gao, Y. Wang, X. Li, and D. Tao, “On combining morphological component analysis and concentric morphology model for mammographic mass detection,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 266–273, 2011.

[83] A. R. Webb, *Statistical Pattern Recognition*, Wiley, New York, NY, USA, 2nd edition, 2002.

[84] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Elsevier, New Providence, NJ, USA, 3rd edition, 2006.

[85] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, Berlin, Germany, 1996.

[86] R. O. Duda and P. E. Hart, *Pattren Classification and Scene Analysis*, John Wiley & Sons, New York, NY, USA, 1973.

[87] N. Karssemeijer and G. Te Brake, “Detection of stellate distortions in mammograms,” *IEEE Transactions on Medical Imaging*, vol. 15, no. 10, pp. 611–619, 1996.

[88] G. Fan and X.-G. Xia, “Wavelet-based texture analysis and synthesis using hidden markov models,” *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 50, no. 1, pp. 106–120, 2003.

[89] S. L. Kok, J. M. Brady, and L. Tarassenko, “The detection of abnormalities in mammograms,” in *Proceedings of the 2nd International Workshop on Digital Mammography*, pp. 261–270, York, UK, 1994.

[90] S. Timp, C. Varela, and N. Karssemeijer, “Temporal change analysis for characterization of mass lesions in mammography,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 7, pp. 945–953, 2007.

[91] S. Sidhu and K. Raahemifar, “Texture classification using wavelet transform and support vector machines,” in *Proceedings of the 2005 Canadian Conference on Electrical and Computer Engineering*, IEEE, Saskatoon, Canada, pp. 941–944, May 2005.

[92] S. H. Amroabadi, M. R. Ahmadzadeh, and A. Hekmatnia, “Mass detection in mammograms using GA based PCA and Haralick features selection,” in *Proceedings of the 19th*

Iranian Conference on Electrical Engineering (ICEE), p. 1, Tehran, Iran, 2011.

[93] L. Ke, N. Mu, and Y. Kang, "Mass computer-aided diagnosis method in mammogram based on texture features," in *Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 1, pp. 354–357, Yantai, China, 2010.

[94] D. Guliato, R. M. Rangayyan, W. A. Carnielli, J. A. Zuffo, and J. E. L. Desautels, "Segmentation of breast tumors in mammograms by fuzzy region growing," in *Proceedings of the 20th Annual International Conference on IEEE Engineering Medicine Biology Society*, Hong Kong, China, 1998.

[95] R. M. Rangayyan, N. M. El-Faramawy, J. E. L. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumours," *IEEE Transactions on Medical Imaging*, vol. 16, no. 12, pp. 799–810, 1997.

[96] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[97] M. H. Bharati, J. Liu, and J. F. MacGregor, "Image texture analysis: methods and comparisons," *Chemometrics and intelligent laboratory Systems*, vol. 72, no. 1, pp. 57–71, 2004.

[98] J. Zhang and T. Tan, "Brief review of invariant texture analysis methods," *Pattern Recognition*, vol. 35, no. 3, pp. 735–747, 2002.

[99] J. Y. Tou, Y. H. Tay, and P. Y. Lau, "Recent trends in texture classification: a review," in *Proceedings Symposium on Progress on Information and Communication Technology 2009 (SPICT 2009)*, pp. 63–68, Kuala Lumpur, Malaysia, 2009.

[100] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels, "Gradient and texture analysis for the classification of mammographic masses," *IEEE Transactions on Medical Imaging*, vol. 19, no. 10, pp. 1032–1043, 2000.

[101] S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 269–285, 1976.

[102] L. H. Siew, R. H. Hodgson, and E. J. Wood, "Texture measures for carpet wear asessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 1, pp. 92–105, 1988.

[103] H.-P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology*, vol. 212, no. 3, pp. 817–827, 1999.

[104] B. Sahiner, C. Heang-Ping, N. Petrick et al., "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 598–610, 1996.

[105] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[106] J. Suckling, J. Parker, D. Dance et al., "Mammographic image analysis society (MIAS) database v1.21 [dataset]," 2015, https://www.repository.cam.ac.uk/handle/1810/250394.

[107] S. Wang, "A review of gradient-based and edge-based feature extraction methods for object detection," in *Proceedings of the 11th IEEE International Conference on Computer and Information Technology, CIT 2011*, pp. 277–282, Pafos, Cyprus, August 2011.

[108] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny, "Structure adaptive approach for dimension reduction," *The Annals of Statistics*, vol. 29, no. 6, pp. 1537–1566, 2001.

[109] A. M. Samarov, "Exploring regression structure using nonparametric functional estimation," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 836–847, 1993.

[110] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. Ng, "Computer-aided breast cancer detection using mammograms: a review," *IEEE Reviews in Biomedical Engineering*, vol. 6, pp. 77–98, 2012.

[111] D. Guliato, R. M. Rangayyan, J. D. Carvalho, and S. A. Santiago, "Spiculation-preserving polygonal modeling of contours of breast tumors," in *Proceedings of the 28th International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2791–2794, New York, NY, USA, August 2006.

[112] I. Cheikhrouhou, K. Djemal, D. Sellami, H. Maaref, and N. Derbel, "New mass description in mammographies," in *Proceedings of the Image Processing Theory, Tools and Applications*, 2008.

[113] S. A. Feig and M. J. Yaffe, *Digital Mammography, Computer-Aided Diagnosis and Telemammography, The Radiologic Clinics of North America*, Breast Imaging Press, vol. 33, 1995.

[114] A. R. Domínguez and A. K. Nandi, "Toward breast cancer diagnosis based on automated segmentation of masses in mammograms," *Pattern Recognition*, vol. 42, no. 6, pp. 1138–1148, 2009.

[115] R. M. Rangayyan and T. M. Nguyen, "Pattern classification of breast masses via fractal analysis of their contours," *International Congress Series*, vol. 1281, pp. 1041–1046, 2005.

[116] M. Rangayyan, Rangaraj, R. Ferrari, and A. Frère, "Analysis of bilateral asymmetry in mammograms using directional, morphological, and density features," *Journal of Electronic Imaging*, vol. 16, no. 1, Article ID 013003, 2007.

[117] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford Calrendon Press, Oxford, UK, 1995.

[118] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics*, Elsevier, vol. 2, pp. 835–855, Amsterdam, Netherlands, 1982.

[119] S. J. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 3, pp. 242–252, 1980.

[120] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[121] S. Kulback and R. A. Liebler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[122] L. Mthembu and J. Greene, "A comparison of three class separability measures," in *Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 63–67, Grabouw, South Africa, November 2004.

[123] S. Watanabe, *Pattern Recognition: Human and Mechanical*, Wiley, New York, NY, USA, 1995.

[124] "Mutual information," 2019, http://en.wikipedia.org/wiki/Mutual_information.

[125] "Pearson product," 2019, http://en.wikipedia.org/wiki/Pearson_product_moment_correlation_coefficient.

[126] K. Kira and L. A. Rendell, “A practical approach to feature selection,” in *ML92 Proceedings of the 9th International Workshop on Machine Learning*, 1992.

[127] Y. Saeys, T. Abeel, and Y. V. de Peer, “Robust feature selection using ensemble feature selection techniques,” in *Machine Learning and Knowledge Discovery in Databases*, pp. 313–325, Berlin, Germany, 2008.

[128] I. Guyon and A. Elissee, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1180, 2003.

[129] J. Luo, B. Hu, X.-T. Ling, and R.-W. Liu, “Principal independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 10, no. 4, 1999.

[130] P. Comon, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[131] E. Oja, “The nonlinear PCA learning rule and signal separation mathematical analysis,” Technical report A26, Helsinki University of Technology, Espoo, Finland, 1995.

[132] E. Oja, J. Karhunen, L. Wang, and R. Vigario, “Principal and independent components in neural networks—recent developments,” in *Proceedings of the 7th Italian Workshop on Neural Nets WIRN’95*, Vietri sulMare, Italy, May 1995.

[133] I. Christoyianni, A. Koutras, E. Dermatas, and G. Kokkinakis, “Computer aided diagnosis of breast cancer in digitized mammograms,” *Computerized Medical Imaging and Graphics*, vol. 26, no. 5, pp. 309–319, 2002.

[134] http://chem-eng.utoronto.ca/%7Edatamining/dmc/zeror.htm, 2019.

[135] C. Nasa and S. Suman, “Evaluation of different classification techniques for web data,” *International Journal of Computer Applications*, vol. 52, no. 9, pp. 34–40, 2012.

[136] G. Buddhinath and D. Derry, *A Simple Enhancement to One Rule Classification*, Department of Computer Science & Software Engineering University of Melbourne, Melbourne, Australia, 2006.

[137] V. Parsania, N. N. Jani, and N. Bhalodiya, “Applying naïve bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis,” *IJDI-ERET*, vol. 3, 2014.

[138] M. E. Maron, “Automatic indexing: an experimental inquiry,” *Journal of the ACM*, vol. 8, no. 3, pp. 404–417, 1961.

[139] J. Rennie, L. Shih, J. Teevan, and D. Karger, “Tackling the poor assumptions of Naive Bayes classifiers,” in *Proceedings of the 20th International Conference on Machine Learning*, Washington, DC, USA, 2003.

[140] I. Rish, “An empirical study of the naive Bayes classifier,” in *Proceedings of the 2001 IJCAI Workshop on Empirical Methods in AI*, New York, NY, USA, 2001.

[141] L. Yuan, “An improved naive Bayes text classification algorithm in Chinese information processing,” in *Proceedings of the 3rd International Symposium on Computer Science and Computational Technology*, pp. 14-15, Jiaozhou, China, 2010.

[142] L. Vibha, G. M. Harshavardhan, K. Pranaw, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, “Classification of mammograms using decision trees,” in *Proceedings of the 2006 10th International Database Engineering and Applications Symposium (IDEAS’06)*, pp. 263–266, Delhi, India, December 2006.

[143] A. K. Mohanty, M. R. Senapati, S. Beberta, and S. K. Lenka, “Texture-based features for classification of mammograms using decision tree,” *Neural Computing & Applications*, vol. 23, no. 3-4, pp. 1011–1017, 2013.

[144] K. Polat and S. Güneş, “A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 1587–1592, 2009.

[145] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[146] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[147] H.-P. Chan, D. Wei, M. A. Helvie et al., “Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space,” *Physics in Medicine and Biology*, vol. 40, no. 5, pp. 857–876, 1995.

[148] J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn, K. A. Shaffer, and E. S. Burnside, “A logistic regression model based on the national mammography database format to aid breast cancer diagnosis,” *American Journal of Roentgenology*, vol. 192, no. 4, pp. 1117–1127, 2009.

[149] W. E. Barlow, E. White, R. Ballard-Barbash et al., “Prospective breast cancer risk prediction model for women undergoing screening mammography,” *JNCI: Journal of the National Cancer Institute*, vol. 98, no. 17, pp. 1204–1214, 2006.

[150] X. Zhou, K. Y. Liu, and S. T. C. Wong, “Cancer classification and prediction using logistic regression with Bayesian gene selection,” *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 249–259, 2004.

[151] B. Samanta, G. L. Bird, M. Kuijpers et al., “Prediction of periventricular leukomalacia. Part I: selection of hemodynamic features using logistic regression and decision tree algorithms,” *Artificial Intelligence in Medicine*, vol. 46, no. 3, pp. 201–215, 2009.

[152] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley, New York, NY, USA, 2000.

[153] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley, New York, NY, USA, 1989.

[154] A. C. Nusantara, E. Purwanti, and S. Soelistiono, “Classification of digital mammogram based on nearest-neighbor method for breast cancer detection,” *International Journal of Technology*, vol. 7, no. 1, p. 71, 2016.

[155] B. N. Prathibha and V. Sadasivam, “Multi-resolution texture analysis of mammograms using nearest neighbor classification techniques,” *International Journal of Information Acquisition*, vol. 7, no. 2, pp. 109–118, 2010.

[156] N. Alpaslan, A. Kara, B. Zenċir, and D. Hanbay, “Classification of breast masses in mammogram images using KNN,” in *Proceedings of the 2015 23rd Signal Processing and Communications Applications Conference (SIU)*, pp. 1469–1472, Malatya, Turkey, 2015.

[157] V. Rodriguez, K. Sharma, and D. Walker, “Breast cancer prediction with K-nearest neighbor algorithm using different distance measurements,” 2018.

[158] V. B. Prasath, H. Arafat Abu Alfeilat, O. Lasassmeh, and A. B. A. Hassanat, “Distance and similarity measures effect on the performance of k-nearest neighbor classifier—a review,” 2017, https://arxiv.org/pdf/1708.04321.pdf.

[159] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA, 1995.

[160] V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*, MIT Press, Cambridge, MA, USA, 2001.

[161] G. Sakr, I. Elhajj, and H. A.-S. Huijer, “Support vector machines to define and detect agitation transition,” *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 98–108, 2010.

[162] J. M. Lesniak, R. Hupse, R. Blanc, N. Karssemeijer, and G. Székely, "Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography," *Physics in Medicine and Biology*, vol. 57, no. 16, pp. 5295–5307, 2012.

[163] I. El-Naqa, Y. Yang, M. Wernick, N. Galatsanos, and R. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.

[164] J. Wei, H. P. Chan, C. Zhou et al., "Computer-aided detection of breast masses: four-view strategy for screening mammography," *Medical Physics*, vol. 38, no. 4, pp. 1867–1876, 2011.

[165] W. H. Land, D. Mckee, R. Velazquez, L. Wong, J. Y. Lo, and F. R. Anderson, "Application of support vector machines to breast cancer screening using mammogram and clinical history data," in *Proceedings of SPIE—Medical Imaging 2003: Image Processing*, vol. 5032, San Diego, CA, USA, 2003.

[166] K. Prabhpreet, S. Gurvinder, and K. Parminder, "Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification," *Informatics in Medicine Unlocked*, vol. 16, Article ID 100151, 2019.

[167] https://www.saedsayad.com/support_vector_machine.htm, 2019.

[168] https://www.saedsayad.com/images/SVM_2.png, 2019.

[169] T. Ayer, Q. Chen, and E. Burnside, "Artificial neural networks in mammography interpretation and diagnostic decision making," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 832509, 10 pages, 2013.

[170] T. Ayer, J. Chhatwal, O. Alagoz, C. E. Kahn, R. W. Woods, and E. S. Burnside, "Comparison of logistic regression and artificial neural network models in breast cancer risk estimation," *Radiographics*, vol. 30, no. 1, pp. 13–22, 2010.

[171] J. E. Dayhoff and J. M. DeLeo, "Artificial neural networks: opening the black box," *Cancer*, vol. 91, no. 8, pp. 1615–1635, 2001.

[172] M. K. Markey, G. D. Tourassi, M. Margolis, and D. M. DeLong, "Impact of missing data in evaluating artificial neural networks trained on complete data," *Computers in Biology and Medicine*, vol. 36, no. 5, pp. 516–525, 2006.

[173] R. G. Stafford, J. Beutel, D. J. Mickewich, and S. L. Albers, "Application of neural networks to computer-aided pathology detection in mammography," *Proceedings of SPIE*, vol. 1896, pp. 341–352, 1993.

[174] J. Lawrence, *Introduction to Neural Networks*, California Scientific Software, Nevada City, CA, USA, 1993.

[175] A. J. Maren, C. T. Harston, and R. M. Pap, *Handbook of Neural Computing Applications*, Academic Press, San Diego, CA, USA, 1990.

[176] https://www.saedsayad.com/artificial_neural_network.htm, 2019.

[177] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[178] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, "Deep learning in mammography and breast histology, an overview and future trends," *Medical Image Analysis*, vol. 47, pp. 45–67, 2018.

[179] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 652–660, Munich, Germany, October 2015.

[180] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, no. 3, Article ID 034501, 2016.

[181] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Im-agenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, Springer, Berlin, Germany, 2012.

[182] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," *Neurocomputing*, vol. 197, pp. 221–231, 2016.

[183] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016.

[184] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific Reports*, vol. 8, no. 1, p. 4165, 2018.

[185] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, Springer, Berlin, Germany, 2015.

[186] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[187] G. Valvano, G. Santini, N. Martini et al., "Convolutional neural networks for the segmentation of microcalcification in mammography imaging," *Journal of Healthcare Engineering*, vol. 2019, Article ID 9360941, 9 pages, 2019.

[188] J. Wang and Y. Yang, "A context-sensitive deep learning approach for microcalcification detection in mammograms," *Pattern Recognition*, vol. 78, pp. 12–22, 2018.

[189] H. Cai, Q. Huang, W. Rong et al., "Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms," *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 2717454, 10 pages, 2019.

[190] W. Fathy and A. Ghoneim, "A deep learning approach for breast cancer mass detection," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019.

[191] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, "A technical review of convolutional neural network-based mammographic breast cancer diagnosis," *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 6509357, 16 pages, 2019.

[192] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, "Deep convolutional neural networks for mammography: advances, challenges and applications," *BMC Bioinformatics*, vol. 20, p. 281, 2019.

[193] N. Dhungel, G. Carneiro, and A. P. Bradley, "The automated learning of deep features for breast mass classification from mammograms," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Athens, Greece, pp. 106–114, 2016.

[194] H. Chougrad, H. Zouaki, and O. Alheyane, "Convolutional neural networks for breast cancer screening: transfer learning with exponential decay," 2017, https://arxiv.org/abs/1711.10752.

[195] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon, "Three-class mammogram classification based on descriptive CNN features," *BioMed Research International*, vol. 2017, Article ID 3640901, 11 pages, 2017.

[196] T. Kooi, A. Gubern-Merida, J. J. Mordang, R. Mann, R. Pijnappel, and K. Schuur, "A comparison between a deep convolutional neuralnetwork and radiologists for classifying regions of interest in mammography," in *Proceedings of the 2016 International Workshop on Digital Mammography*, Springer, Malmö, Sweden, pp. 51–56, June 2016.

[197] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[198] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 24, no. 3, pp. 371–380, 2005.

[199] L. Lam, "Classifier combinations: implementations and theoretical issues," in *Multiple Classifier Systems*, Vol. 1857, Springer, Berlin, Germany, 2000.

[200] R. K. Powalka, N. Sherkat, and R. J. Whitrow, "Multiple recognizer combination topologies," in *Handwriting and Drawing Research: Basic and Applied Issues*, M. L. Simner, C. G. Leedham, and A. J. W. M. Thomasson, Eds., pp. 329–342, IOS Press, Amsterdam, Netherlands, 1996.

[201] A. F. R. Rahman and M. C. Fairhurst, "An evaluation of multi-expert configurations for the recognition of handwritten numerals," *Pattern Recognition*, vol. 31, no. 9, pp. 1255–1273, 1998.

[202] L. Lam, "Classifier combinations: implementations and theoretical issues," in *Multiple Classifier Systems*, vol. 1857, pp. 78–86, Springer, Cagliari, Italy, 2000.

[203] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin, "Limits on the majority vote accuracy in classifier fusion," *Pattern Analysis & Applications*, vol. 6, no. 1, pp. 22–31, 2003.

[204] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Information Fusion*, vol. 6, no. 1, pp. 63–81, 2005.

[205] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[206] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[207] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on Machine Learning*, vol. 39, pp. 148–156, Lanzhou, China, 1996.

[208] X. Zhang, "A new ensemble learning approach for microcalcification clusters detection," *Journal of Software*, vol. 4, no. 9, pp. 1014–1021, 2009.

[209] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[210] K. S. Woods, C. C. Doss, K. W. Bowyer, J. L. Solka, C. E. Priebe, and W. P. Kegelmeyer, "Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography," *International Journal on Pattern Recognition and Artificial Intelligence*, vol. 7, no. 6, pp. 1417–1436, 1993.

[211] A. Devos, A. W. V. Simonetti, M. Graaf et al., "The use of multivariate MR imaging intensities versus metabolic data from MR classification," *Journal of Magnetic Resonance*, vol. 173, no. 2, pp. 218–228, 2005.

[212] J. K. Kim, J. M. Park, K. S. Song, and W. Park, "Detection of clustered microcalcifications on mammograms using surrounding region dependence method and artificial neural network," *Journal on VLSI Signal Process*, vol. 18, pp. 251–262, 1998.

[213] M. A. Alolfe, W. A. Mohamed, A.-B. M. Youssef, A. S. Mohamed, and Y. M. Kadah, "Computer aided diagnosis in digital mammography using support vector machine and linear discriminant analysis classification," in *Proceedings of the 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2609–2612, Cairo, Egypt, November 2009.

[214] Y. Chitre, A. P. Dhawan, and M. Moskowitz, "Artificial neural network based classification of mammographic microcalcifications using image structure features," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 12, pp. 1377–1402, 1993.

[215] D. Wei, H. P. Chan, M. A. Helvie et al., "Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis," *Medical Physics*, vol. 22, no. 9, pp. 1501–1513, 1995.

[216] A. H. Baydush, D. M. Catarious, C. K. Abbey, and C. E. Floyd, "Computer aided detection of masses in mammography using subregion hotelling observers," *Medical Physics*, vol. 30, no. 7, pp. 1781–1787, 2003.

[217] Y. Wang, F. Aghaei, A. Zarafshani, Y. Qiu, W. Qian, and B. Zheng, "Computer-aided classification of mammographic masses using visually sensitive image features," *Journal of X-Ray Science and Technology*, vol. 25, no. 1, pp. 171–186, 2016.

[218] A. M. Khuzi, R. Besar, W. M. D. Wan Zaki, and N. N. Ahmad, "Identification of masses in digital mammogram using gray level co-occurrence matrices," *Biomedical Imaging and Intervention Journal*, vol. 5, no. 3, p. e17, 2009.

[219] M. Marina, J. Dragan, and P. Aleksandar, "Comparative analysis of breast cancer detection in mammograms and thermograms," *Biomedical Engineering/Biomedizinische Technik*, vol. 60, no. 1, pp. 49–56, 2015.

[220] S. Ramadan and M. El-Banna, "Breast cancer diagnosis in digital mammography images using automatic detection for region of interest," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 15, 2019.

[221] J. Suckling, "The mammographic image analysis society digital mammogram database exerpta medica," *International Congress Series*, vol. 1069, pp. 375–378, 1994.

[222] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. Philip Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th International Workshop on Digital Mammography*, M. J. Yaffe, Ed., Medical Physics Publishing, Toronto, Canada, pp. 212–218, June 2000.

[223] Lawrence Livermore National Library/UCSF Digital Mammogram Database, *Center for Health Care Technologies Livermore*, Lawrence Livermore National Library, Livermore, CA, USA, 2010, http://gdo-biomed.ucllnl.org/pub/mammo-db/.

[224] S. R. Amendolia, M. G. Bisogni, and U. Bottigli, "The CALMA project," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 461, no. 1–3, pp. 428-429, 2001.

[225] PROENG, "Image processing and image analysis applied to mastology," 2012, http://visual.ic.uff.br/en/proeng/.