Chapter 9

Breast Cancer Detection by Means of Artificial Neural Networks

Jose Manuel Ortiz-Rodriguez,
Carlos Guerrero-Mendez,
Maria del Rosario Martinez-Blanco,
Salvador Castro-Tapia, Mireya Moreno-Lucio,
Ramon Jaramillo-Martinez,
Luis Octavio Solis-Sanchez,
Margarita de la Luz Martinez-Fierro,
Idalia Garza-Veloz, Jose Cruz Moreira Galvan and
Jorge Alberto Barrios Garcia

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.71256

Abstract

Breast cancer is a fatal disease causing high mortality in women. Constant efforts are being made for creating more efficient techniques for early and accurate diagnosis. Classical methods require oncologists to examine the breast lesions for detection and classification of various stages of cancer. Such manual attempts are time consuming and inefficient in many cases. Hence, there is a need for efficient methods that diagnoses the cancerous cells without human involvement with high accuracies. In this research, image processing techniques were used to develop imaging biomarkers through mammography analysis and based on artificial intelligence technology aiming to detect breast cancer in early stages to support diagnosis and prioritization of high-risk patients. For automatic classification of breast cancer on mammograms, a generalized regression artificial neural network was trained and tested to separate malignant and benign tumors reaching an accuracy of 95.83%. With the biomarker and trained neural net, a computer-aided diagnosis system is being designed. The results obtained show that generalized regression artificial neural network is a promising and robust system for breast cancer detection. The Laboratorio de Innovacion y Desarrollo Tecnologico en Inteligencia Artificial is seeking collaboration with research groups interested in validating the technology being developed.

Keywords: breast cancer detection, digital image processing, artificial neural networks, biomarkers, computer-aided diagnosis

1. Introduction

1.1. Breast cancer and early detection

Nowadays, cancer is a massive public health problem around the world. According to the International Agency for Research on Cancer (IARC) [1], part of the World Health Organization (WHO), there were 8.2 million deaths caused by cancer in 2012 and 27 million of new cases of this disease are expected to occur until 2030 [2].

Cancer, medically defined as a malignant neoplasm, is a board group of disease involving unregulated cell growth [2]. In cancer, cell divides and grows uncontrollably forming malignant tumors, invading nearby parts of the body. The cancer can spread to all parts of the body through the lymphatic systems or blood streams [3].

Cancer can be diagnosed by classifying tumors in two different types such as malignant and benign. Benign tumors represent an unnatural outgrowth but rarely lead to a patient's death; yet, some types of benign tumors, too, can increase the possibility of developing cancer [4]. On the other hand, malignant tumors are more serious and their timely diagnosis contributes to a successful treatment. As a result, predication and diagnosis of cancer can boost the chances of treatment, decreasing the usually high costs of medical procedures for such patients [5].

Breast cancer (BC) is the most commonly diagnosed cancer and the most common cause of death in women all over the world. Among the cancer types, BC is the second most common cancer for women, excluding skin cancer [6]. Besides, the mortality of BC is very high when compared to other types of cancer [7]. BC, similar to other cancers, starts with a rapid and uncontrolled outgrowth and multiplication of a part of the breast tissue, which depending on its potential harm, is divided into benign and malignant types.

Generally, there are two types of BC that are in situ and invasive. In situ starts in the milk duct and does not spread to other organs even if it grows [8]. Invasive breast cancer on the contrary, is very aggressive and spreads to other nearby organs, and destroys them as well [9]. It is very important to detect the cancerous cell before it spreads to other organs; thus, the survival rate for patient will increase to more than 97% [10].

A major class of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. The evaluation of data taken from patients and decisions of experts are the most important factors in diagnosis. The correct diagnosis of BC is one of the major problems in the medical field. As BC can be very aggressive, only early detection can prevent mortality. Clinical diagnosis of BC helps in predicting the malignant cases and timely diagnosis can increase the chances of a patient's life expectancy from 56 to 86% [11].

BC has four early signs: microcalcification, mass, architectural distortion, and breast asymmetries [12]. The various common methods used for breast cancer diagnosis (BCD) are positron emission tomography (PET), magnetic resonance imaging (MRI), CT scan, X-ray, ultrasound, photo-acoustic imaging, tomography, diffuse optical tomography, elastography, electrical impedance tomography, opto-acoustic imaging, ophthalmology, mammogram, etc. [13]. The results obtained from these methods are used to recognize the patterns, which are aiming to help the doctors for classifying the malignant and benign cases.

Despite of recent advances in the comprehension of the molecular biology of BC progression and the discovery of new related molecular markers, the histopathological analysis remains the most widely used method for BC diagnosis [14]. Despite significant progress reached by diagnostic imaging technologies, the final BCD, including grading and staging, continues being done by pathologists applying visual inspection of histological samples under the microscope [15].

However, manual classification of images is a challenging and time-consuming task, being highly susceptible to interobserver variability and human errors, resulting in extremely poor critical outcomes, thus markedly increasing the workload of radiologists because of their significant shortage. In addition, medical care costs that are relevant to imaging rapidly increase. Therefore, new methods for diagnosis are required.

Currently, bioimaging quantification is an emerging technique in the field of radiology with a growing implantation in hospital centers. It provides relevant information that is not appreciable by the naked eye in conventional radiological reading. It consists of the generation of quantitative (numerical) data from images, mainly of high resolution, to provide information on which to support a clinical assessment [16]. Biomarkers can be said to be the transition from radiology to personalized medicine.

1.2. Bioimagen markers in breast cancer detection

Bioimagen markers allow to characterize and to study different diseases using some kind of information, such as genetic, histological, clinical imaging, etc. These biomarkers can be used to detect abnormalities in the data as genetic mutations that cause some diseases and can also be used in the clinic for the detection of patients with some types of disease [17].

The application of quantification of bioimaging markers to aid in the diagnosis, treatment, and follow-up of pathologies provides added value throughout the clinical practice process by providing additional information to conventional diagnostic tests [18]. From imaging tests processed in the right way, abnormalities in a tissue are evidenced before they are perceptible in the reading of the radiologist, fundamental objective of this type of biomarkers [19]. In addition, they allow the monitoring of the treatment effect from a quantitative point of view.

As mentioned before, BC is one of the leading causes of death in women around the world, accounting for nearly one-third of cancer-related deaths. Currently, the clinical screening by mammography is the most effective way for the early detection of this disease. Using analysis of mammograms obtained through X-rays allows radiologists to visualize early signs of cancer, such as calcifications, masses, and architectural distortions among other early signs of cancer. However, this analysis is a routine, monotonous, and exhausting task and it is estimated that only 0.3–0.4% of the cases are actually carcinogenic [20].

It has been shown that because of these problems and other factors intrinsic to cancer such as obscuration of abnormalities by fatty tissue, a radiologist can omit up to 30% of cancers. Moreover, because this type of analysis produces many false positives, the number of unnecessary biopsies is increased up to 35%, causing a high level of stress in the patient, and in turn saturating the health systems [21].

Due to all the problems presented by mammography screening, great efforts have been made to support the radiologist in the search for these lesions through computer-aided detection (CAD) or biomarker systems, which try to help the radiologist by taking advantage of the latest advances in computer vision and their manipulation in digital form [22].

At present, computer-aided detection/diagnosis (CAD/CADx) systems are one of the numerous major research topics in diagnostic radiology and medical imaging. CAD systems allow the radiologist to manipulate mammography to highlight certain features that would otherwise be difficult to visualize. One of the most used techniques is the improvement of contrast, which allows to highlight objects in areas of low intensity. To date, CAD is a more suitable method for primary diagnosis of cancer in computed tomography, X-ray, MRI, or mammogram images. CAD system is an effective intermediate between input images and the radiologist. The output from CAD is not considered as an end result; nevertheless, the result is used as reference with regard to additional testing in the related field [23].

The CAD approach helps medical doctors to diagnose diseases with a higher degree of efficiency, while minimizing examination time and cost, as well as avoiding unnecessary biopsy procedures. However, CAD systems not only allow a better visualization of mammograms, but also using different digital image processing (DIP), knowledge discovery from data (KDD), artificial intelligence (AI) techniques such artificial neural networks (ANN) allow to preselect certain regions of interests (ROIs) for later analysis by the radiologist [24].

Classification of histopathology images into distinct histopathology patterns, corresponding to the noncancerous or cancerous condition of the analyzed tissue, is often the primordial goal in image analysis systems for cancer automatic-aided diagnosis applications. Recent advances in DIP, KDD, and AI techniques allow to build CAD/CADx that can assist pathologists to be more productive, objective, and consistent in diagnosis. The main challenge of such systems is dealing with the inherent complexity of histopathological images.

The aim of this research is to use advanced DIP to investigate and develop specific imaging biomarkers for Mexican patients through the quantitative mammography analysis and with this information to develop technology based on advanced KDD and AI techniques, aiming to detect breast cancer in the early stages in order to support the diagnosis and prioritization of high-risk patients.

2. Development of a CADx system to identify breast abnormalities in digital mammograms images using artificial neural networks

In this research, the study of BC disease using advanced techniques of DIP, KDD, and AI was carried out in order to develop imaging biomarkers that allow to carry out diverse studies for BCD. As is showed in **Figure 1**, the research was divided in three main stages.

Currently, there are no public databases of BC in Latin America or Mexico. Therefore, at first stage, different public mammography databases were used for developing and validating digital image processing algorithms capable to select ROIs from mammograms to extract

First stage
Mammographic feature extraction and artificial neural network training
Second stage
Retrospective study of Mexican patient's mammograms
Third stage
Prospective study of Mexican patient's mammograms

Figure 1. Main stages of research.

image features used to train a generalized regression artificial neural network (GRANN). The aim was to generate a methodology for the characterization of mammograms and their association with risk factors in BC patients as well as to integrate and to develop the technological tools for mammography analysis for BCD using AI technology. In this work, results obtained at first stage are presented.

Because there are no public databases of BC for Mexican patients, it was proposed to establish two protocols for the acquisition of mammograms. The first protocol, second stage, seeks to obtain data retrospectively, which will allow obtaining the mammograms necessary to validate in Mexican patients the methodology and technological tools developed at stage one. The aim of the second stage is the generation of an anonymous database of Mexican patients for free use by the scientific community for the study of BC in Mexican patients and to validate the methodologies developed in collaboration with the General Hospital of Zacatecas (GHZ) and the Molecular Medicine Laboratory (MML) from Autonomous University of Zacatecas, Mexico.

In the second protocol, third stage, it is sought to generate a long-term prospective protocol, which will allow the creation of a database with different risk factors associated with the development of BC. This protocol will allow to collect clinical data of patients with both high and low probabilities of developing cancer. These data will be able to validate the methodologies of cancer detection by the scientific community.

The generation of a prospective protocol will allow the expansion of the database for the study of breast cancer in Mexican patients. Unlike the retrospective protocol, prospective protocol aims to include clinical data, risk factors, and mammograms, among others. This database would present to the scientific community a reference for the development of new breast cancer detection techniques in Mexican patients.

2.1. Methodology: feature extraction and neural network training

Patient prioritization can play a very important role in the reach of health services in developing countries such as Mexico, where not all have access to these specialized oncology services. Therefore, this research seeks the study of BC by generating a methodology that allows the detection of patients with a high probability of BC. In this work, a new technology to generate mammographic biomarkers and a CADx system for breast cancer diagnosis was designed in order to analyze Digital Image Mammograms (DIM). With this knowledge, it is proposed to create a biomarker specifically designed for the Mexican population. As is showed in **Figure 2**, the first stage was divided into six main stages.

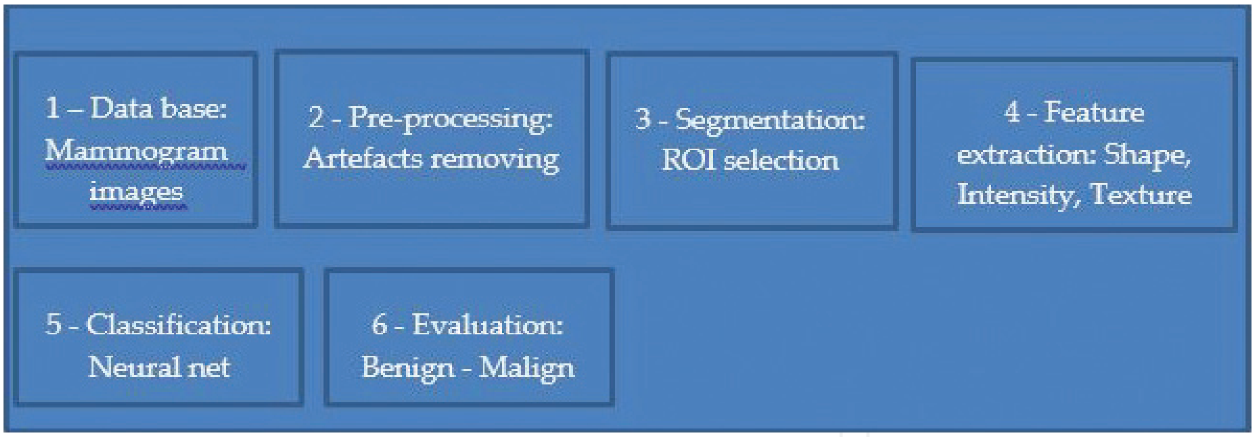
1 – Data base: Mammogram images
2 - Pre-processing: Artefacts removing
3 - Segmentation: ROI selection
4 - Feature extraction: Shape, Intensity, Texture
5 - Classification: Neural net
6 - Evaluation: Benign - Malign

Figure 2. First stage of research. Mammographic feature extraction and artificial neural network training.

2.1.1. Data base: mammogram images

The development of CAD/CADx systems involves their generation and validation by using mammograms obtained from clinical studies, or using public databases. However, at the global level, there are few public databases available to the scientific community to investigate. As mentioned before, currently, there are no public databases of BC in Latin America or Mexico for conducting this kind of studies. Therefore, at first stage, different public mammography databases were used for developing and validating digital image processing algorithms capable to select ROIs from mammograms and to extract image features used to train a GRANN capable to diagnose BC as an aid for radiologists.

Currently, there are only three public databases to conduct this type of research: the first one is the Digital Database for Screening Mammography (DDSM) that has a total of 2620 cases distributed in 625 normal, 1011 benign and 914 malignant and includes the two standardized views CC and MLO. Another available database is the Mammographic Image Analysis Society - Digital Mammogram Data Base (mini-MIAS) that has 322 cases; however, it only has the MLO view. DDSM and mini-MIAS databases are both form North America.

A newly created database is the Breast Cancer Digital Repository (BCDR) from Europe. The creation of BDR is supported by the IMED Project (Development of Algorithms for Medical Image Analysis). The IMED project was created by INEGI, FMUP-CHSJ University of Porto, Portugal and CIEMAT, Spain from 2009 to 2013. This database has 724 patients (723 women and 1 man), aged between 27 and 92 years.

In the first stage of this research, the mini-MIAS, DDSM, and BCDR databases were used to generate and validate the development of a biomarker, an artificial neural network approach with incremental learning and with both, the design of a CADx methodology, carried out in a general scope. However, it is important to highlight that these databases are formed by patients with ethnic characteristics typical of their region, which makes it difficult to transfer knowledge to other countries and their own features.

Moreover, as has been shown by the scientific community, BC varies widely between different etiologies and may prove that systems created for a population may not work for a different population in the way they were thought. This is further aggravated by different types of diets, customs, and lifestyle. Due to this, the development of biomarkers and CAD systems for the Mexican population needs an adaptation to the characteristics of our population. In second and third stages of this research, the designed methodology will be focused and refined for its operation in Mexican patients.

In this work, results obtained with BCDR database are presented. BCDR database contains useful information of each mammogram such as gender: masculine or feminine; segmentations of mammogram, marked in red pixels the ROI that contains the lesion found by the radiologist; patient ID; the age of the patient; breast density, i.e., the percentage of breast density according to Breast Imaging Reporting and Data System (BI-RADS) standard expressed as percentage of glandular and fibrous tissue; breast localization, depending on the location of breast of the RIO with the lesions; mammography, the type of lesion found by the mammographic image expert; biopsy result, anatomical pathology of the biopsy; categorization of the definitive diagnosis; the BI-RADS classification of the lesion; and finally, intensity and shape descriptors of ROI. However, it is important to mention that for this research, these descriptors were not used to train the neural network. Instead, a set of computer algorithms were designed in order to extract image descriptors of ROI of mammograms as described in later section.

2.1.2. Preprocessing: artifact removing and segmentation process

A mammogram image can be considered as a representation of the X-ray radiation density that reflects the tissue of the breast. A risk factor for BC in a patient is recognized when a white region appears on the mammogram image, which means a high tissue density, that may be considered abnormal. A breast abnormality is commonly called ROI. According to DIP techniques, segmentation of breast abnormalities on mammograms is a crucial step in CAD systems. It is a difficult task, since these types of medical images are in low-intensity contrast, making it difficult to identify the edges of a suspicious mass.

In the methodology used in this research, only lateral mammographic images taken from BCDR database were used. In all selected images, a lesion exists, which is considered as benign or malignant; digital mammographic images of BCDR database can be accessed in two forms: the first one from films (photographic films) and the second one from digital images taken from X-ray system (mammography images). Films' images require the design of digital image processing algorithms to eliminate artifacts such as red pixels and prenoise such as labels used by radiologists to identify left or right breast as well as patient identification information. Conversely, digital mammography images only require the design of algorithms for removing red pixels.

The films approach improves digital mammography images increasing the high frequency and eliminating the noise and unwanted artifacts in the ROI. As can be appreciated in **Figure 3**, at preprocessing stage, a computer tool was designed to automate the preprocessing of film digital mammographic images (FDMIs). All FDMIs are treated to eliminate image artifacts such as background, noise, and image labels. In the FDMIs, a common threshold was applied to create a region of the breast and other regions with the labels and artifacts on the mammography.

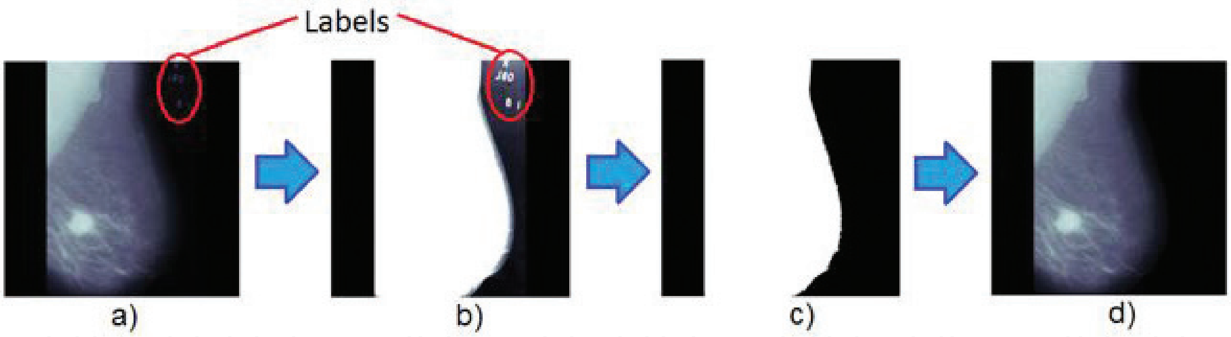

Figure 3. Preprocessing of a mammographic image. (a) Original image, (b) breast region, highlighting labels and background noise, (c) clean breast binary image, and (d) mammography image cleaned.

Using the designed automated computer tool, after creating the logical image, all small regions (less than 10,000 pixels) are eliminated to remove the regions considered as unnecessary in FDMIs. Then, the logical image (mask) is used instead of the original image to obtain an image of the breast without artifacts and labels. **Figure 3** shows the preprocessing method to remove noise and labels in a digital mammography image.

Converting a greyscale to a digital or logical image is a common task of digital image processing. There are many methods for calculating the threshold value for creating logical images. As can be appreciated in **Figure 4**, in this research, the threshold was calculated by converting the nonzero pixel's values to 1. To create a logical image that contains the ROI and the pectoral muscle, the gray tones were converted to white level.

For removing the pectoral region in the logical image, as is showed in **Figure 4**, the white region that is connected to the border of the binary image was eliminated. Therefore, the surplus white region represents the ROI detected in the mammography image. With a cleaned image, the next step is the segmentation process.

On the other hand, the digital mammographic approach works as follows: to calculate the descriptors, the ROI of the lesion is manually segmented by the expert radiologist as can be appreciated in **Figure 5**. At preprocessing stage, the image is fitted for the segmentation. The pixels in red are turned into black. Using the black pixels, the ROI is separated from the rest of the breast image for making a segmentation of the ROI as is showed in **Figure 6**.

For the segmentation of the ROI, a binary or logical image with a very high binarization threshold is created where low gray levels become white. This approach considers most of

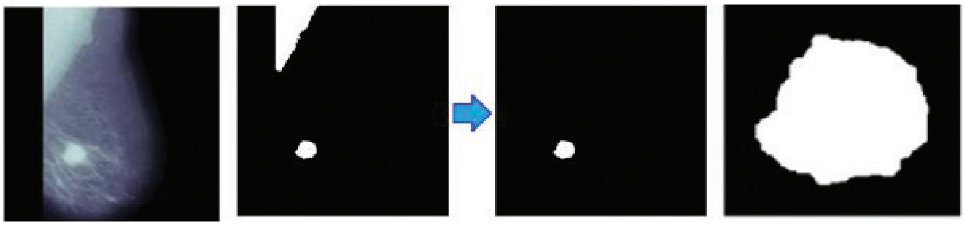

Figure 4. Pectoral region removing process.

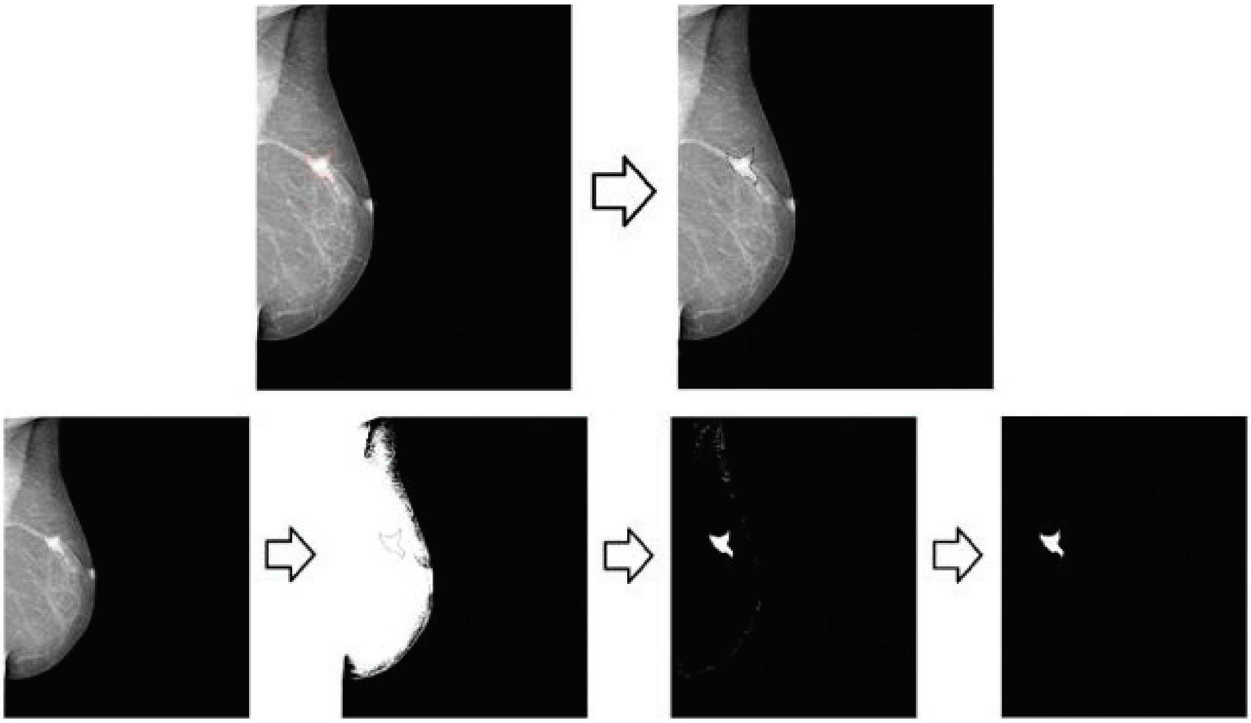

Figure 5. Preprocessing stage of digital mammogram approach.

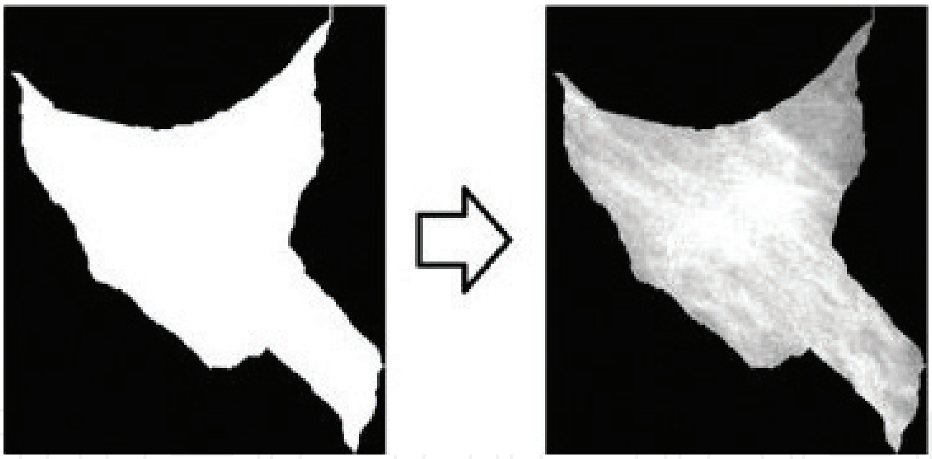

Figure 6. Binary mask and ROI in tones of grays.

gray pixels of the image looking not to lose many pixels from ROI. Afterward, the white logical region that is connected to the edge of the mammographic image is removed. Some white pixels pertaining to the contour of the breast are discarded when the pixels in the image with a small area are removed. Finally, the white region with a greater number of pixels is extracted, which would be considered as the neoplasia.

Next, a binary mask is created using the ROI obtained in the segmentation stage. With the mask and together with the complete image in shades of gray, we will get the ROI in shades of gray as is showed in **Figure 6**.

The next step in the operation of regular CAD systems is the feature extraction of the RIO. The feature extraction can be defined as the process to infer and quantify the parameters that characterize the object being studied. The feature extraction contributes to the analysis of the ROI. It is possible to quantify the shape, texture, size, border, and other tissue parameters that can contribute to the diagnosis and detection of a cancer risk factor. As is showed in **Figure 7**, in this work, shape, intensity, and texture features were extracted in order to create a biomarker for BCD using a CADx system that uses AI technology.

The image features of all Digital Image Mammography (DIM) of BCDR database were extracted and used to build a biomarker to train an ANN. The BCDR digital images are in RGB and gray-level digitalized in JEPG format with a depth of 24 and 8 bits per pixel, respectively, and a resolution of 3328 × 4084. The RGB mammograms are used to show the red remarked section by a radiologist to delimit the found anomaly.

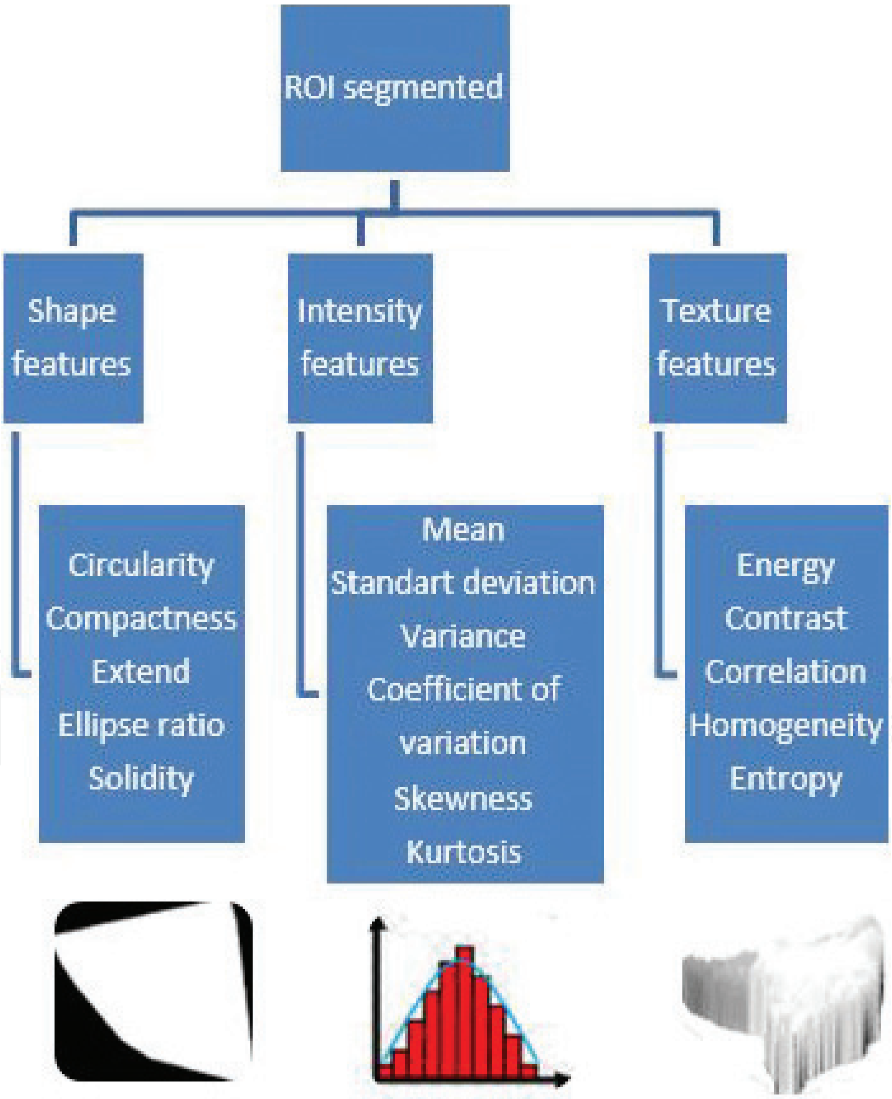

Figure 7. Image features extracted.

The segmentation process use the red section remarked in the RGB mediolateral oblique view mammograms to obtain the ROI. In the RGB mammograms, all the red pixels in the image and the pixels outside the original red region were eliminated. Finally, the remained pixels in the gray-level mammogram were used to get the ROI used for the features' calculation.

Using a custom-designed automated computer tool, a total of 361 images (36 malignant and 325 with benign abnormalities) from 239 patients were segmented in order to get the RIOs. This information was used as the entrance data for training and testing a GRANN. This computer tool saves a lot of time in the preprocessing stage of mammography analysis. This tool will be used at second and third stages of research to analyze the mammograms of Mexican patients in order to create a Mexican biomarker and a CADx system for BCD.

Feature extraction of digital images, such as obtained in digital mammograms, as is showed in **Figure 7**, is a manner to represent an element or an ROI in an image like a fingerprint, and these features are used in many research areas such as machine learning, patter recognition, image processing, or diagnosis of disease in medical science. Feature extraction is a crucial task before classifying an ROI or a pattern in an image.

2.1.3. Feature extraction: shape, intensity, and texture

Image processing is the most important area where the feature extraction is applied, in which mathematical algorithms are used to detect and isolate various desired portions or shapes, features, of digitalized images or video streams, it is particularly important in the area of pattern recognition or character recognition.

The feature extraction method is the measure of physical parameters visualized in a segmented region of an image. The aim of feature extraction is to find a mathematical way to represent the image information, which is important, in a compact form, for solving a computational task. In BCD, these features help to determine the kind of tumor detected in a mammogram image. The choice of features has a crucial influence on the accuracy of classification, the time needed for classification, the number of examples needed for learning, and the cost of performing classification. In breast abnormalities, classification of the differences in mass between benign and malignant on a mammography can be distinguished from their shape, textures, and the intensity in the image.

In this research, an automated computer tool was designed to calculate shape, intensity, and texture features from ROI extracted from BCDR mammograms. The shape features of MDI use the pixels inside and the border of the ROI. These descriptors, showed in **Table 1**, only have a valid meaning in binary or logical images, and some simple shape features are used to describe a ratio between some geometrical figures; for example: extend, ellipse_ratio, and solidity. Most common shape features are the area and perimeter of the region, but they are applied when the ROI size is invariant. However, the area and perimeter can be used to create a relation as the circularity and compactness.

The intensity features, showed in **Table 2**, use the shape intensity histogram to get information that describes the image; i.e., the intensity features use the probability and statistics from the values of the pixels in the image. The mean is the average intensity level. The standard is used

Shape features	
Circularity	Extend
$Circularity = \frac{4\pi ROI_{Area}}{ROI_{Perimeter}{}^{2}}$	$Extend = \frac{ROI_{Area}}{Box_{Area}}$
Compactness	Ellipse_ratio
$Compactness = \frac{ROI_{Perimeter}{}^{2}}{ROI_{Area}}$	$Ellipse_{ratio} = \frac{ROI_{Area}}{Ellipse_{Area}}$
	Solidity
	$Solidity = \frac{ROI_{Area}}{Convex_{Hull_{Area}}}$

Table 1. Shape features.

Intensity features	**Texture features**
Mean	Energy
$\overline{\mu} = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}P_{RIO}(x,y),$	$Energy = \sum_{x=1}^{M}\sum_{y=1}^{N}P_{RIO}(x,y)^{2}$
where P_{RIO} is the intensity pixel value in the coordinates x and y.	
Standard deviation	Contrast
$\sigma = \sqrt{\frac{1}{MN-1}\sum_{x=1}^{M}\sum_{y=1}^{N}\left\vert P_{RIO}(x,y)-\overline{\mu}\right\vert^{2}}$	$Contrast = \sum_{x=1}^{M}\sum_{y=1}^{N}(x-y)^{2}P_{RIO}(x,y)$
Variance	Correlation
$variance = \sqrt{\sigma}$	$Correlation = \sum_{x=1}^{M}\sum_{y=1}^{N}\frac{\left(x-\overline{\mu_{x}}\right)\left(y-\overline{\mu_{y}}\right)P_{RIO}(x,y)}{\sigma_{x}\sigma_{y}},$ where $\overline{\mu_{x}}$, $\overline{\mu_{y}}$, σ_{x}, and σ_{y} are the mean values and the standard deviation P_{xRIO} and P_{yRIO}, respectively.
Coefficient of variation	Homogeneity
$Coefficient\ of\ variation = \frac{\sigma}{\overline{\mu}}$	$Homogeneity = \sum_{x=1}^{M}\sum_{y=1}^{N}\frac{P_{RIO}(x,y)}{1+\vert x-y\vert}$
Skewness	Entropy
$Skewness = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(\frac{P_{RIO}(x,y)-\overline{\mu}}{\sigma}\right)^{3}$	$Entropy = -\sum_{x=1}^{M}\sum_{y=1}^{N}P_{RIO}(x,y)\ \log\left[P_{RIO}(x,y)\right]$
Kurtosis	
$Kurtosis = \left\{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left[\frac{P_{RIO}(x,y)-\overline{\mu}}{\sigma}\right]^{4}\right\}-3$	

Table 2. Intensity and texture features.

to quantify the amount of variation of the set of intensities levels. The variance refers to the variation of the intensities around the mean value. The coefficient of variation is a standardized measure of dispersion in the values. Finally, the skewness and kurtosis measure the histogram symmetric.

The texture features, **Table 2**, describe the *roughness* of an image. Texture features attempt to capture features of the intensity fluctuations between groups of neighboring pixels. The texture is something to which the human eye is very sensitive. In this research, the energy, contrast, correlation, homogeneity, and entropy were used. The energy is a measure of textural uniformity of an image. The contrast refers to the difference in luminance in the ROI. Correlation texture measures the dependence of gray levels on those of neighboring pixels. Homogeneity measures the similarity of values in the ROI. Entropy measures the disorder of value pixels of an image.

As before mentioned, medical diagnosis is an important and complicated task that needs to be executed accurately and efficiently. At present, new techniques based on data mining, KDD, and AI in healthcare are being used mainly for predicting various diseases as well as assisting doctors in diagnosis in their clinical decision. One area where this effort has been most felt is the diagnosis of breast cancer in women. However, the absence of any fully effective, efficient method of BCD has led researchers to develop automated computational systems. In this research, automated CADx technology based on ANN as decision-making tool in the field of BCD is being developed.

2.1.4. Classification and evaluation

For automatic classification of BC on DIM, a GRANN was used to separate malignant and benign tumors [25]. GRANN falls into the category of probabilistic neural networks (PNN) [26–30]. GRANN is a neural network architecture of one-step-only learning that can solve any function approximation problem. The learning process is equivalent to finding a surface in a multidimensional space that provides a best fit to the training data. During the training process, it just stores training data and later uses it for predictions. This neural net is very useful to perform predictions and comparisons of system performance in practice. In GRANN architecture, there are no training parameters, just a smoothing factor (σ) that is applied after the network is trained. The choice of this factor is very important [26–30].

In this research, as is showed in **Figure 8**, a GRANN was trained and tested using a data set of 361 mammograms extracted from BCDR public database. For each mammogram, 35 image descriptors were calculated through an automated computer tool specifically designed for this purpose. These image features were used to train the neural net in order to classify benign and malignant BC for decision making in BDC.

As can be appreciated from **Figure 8**, the image features were used as entrance data, and the malignant (cancerous) and benign (noncancerous) instances were used as output data. In order to train the network, the dataset was randomly divided into two subsets, one with about 80% of the instances to training and another with around the remaining 20% of instances to testing.

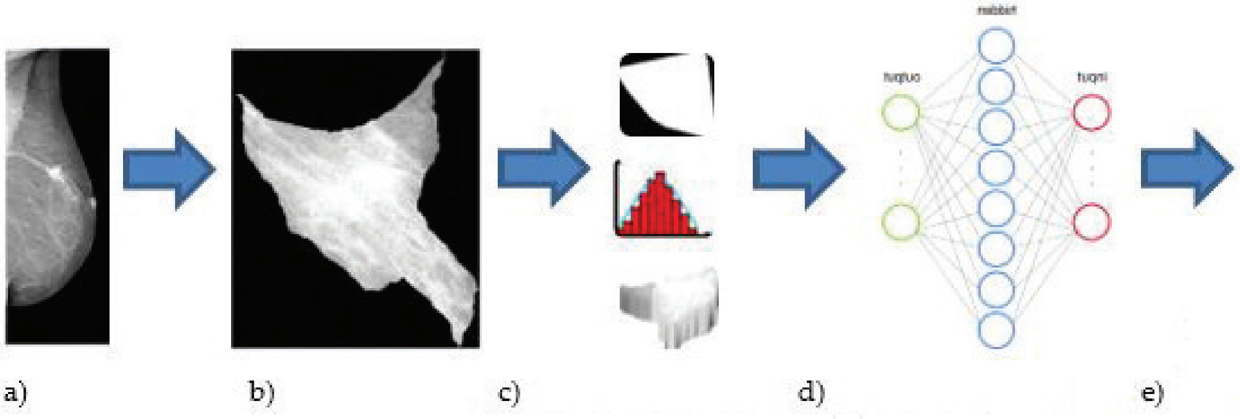

Figure 8. Training of GRANN for BCD. (a) Breast image, (b) segmentation, (c) image descriptors, (d) network training, and (e) BCD.

After 2000 network trainings, a smoothing factor equal to 1e−4 was calculated. This value was used for training the neural net reaching an accuracy of 95.83%. The results obtained in this work show that GRANN is a promising and robust system for BCD. The performance of a trained GRANN was evaluated using four performance measures: accuracy, sensitivity, specificity, and precision. These measures are defined by four decisions: true positive (TP), true negative (TN), false positive (FN), and false negative (FN). TP decision occurs when malignant instances are predicted rightly. TN decision benign instances are predicted rightly. FP decision occurs when benign instances are predicted as malignant. FN decision occurs when malignant instances are predicted as benign.

Accuracy can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensitivity can be calculated as:

$$Sensitivity(recall) = \frac{TP}{TP + FN} \tag{2}$$

Specificity can be calculated as:

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

Precision can be calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

The confusion matrix for the data set, showed in **Table 3**, was computed using these values into above equations to find accuracy, sensitivity, specificity, and precision. **Table 3** shows the

		Obtained		
		0	1	
Goal	0	66	2	
		91.70%	2.80%	
	1	1	3	
		1.40%	**4.20%**	
				95.80%
				4.20%

Table 3. Confusion matrix.

classification results of BC. The confusion matrix is a table that allows to visualize the execution of an algorithm, usually a supervised learning. In this case, it is a classifier of two classes, the current or expected and those obtained by the system (predictions). Each column of the matrix represents the cases that the system, in this case, the neural network, predicted, while the rows represent the expected values.

The diagonal indicates the successes achieved by the system; that is, the trained neural network obtained 91.7% + 4.2% = 95.8% of accuracy, successfully predicting 66 + 3 = 69 lesions of a population of 72, representing an error of 4.2% with three errors from a population of 72 lesions.

With both biomarkers obtained from mammograms and the trained GRANN, a CADx technology system, as showed in **Figure 9**, is being created to be used at second stage in collaboration with GHZ and MML.

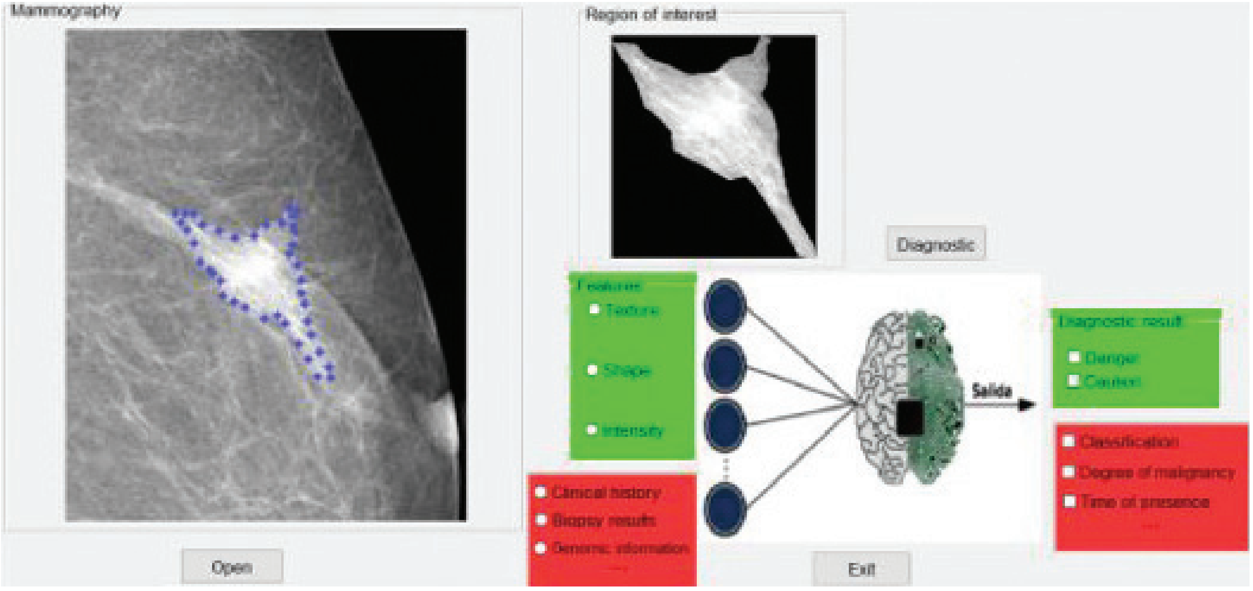

Figure 9. CADx system being developed based on DIP, KDD, and AI methodologies.

The Computer Aided Diagnosis system consists of two main stages: the first one detects the suspicious regions with high sensitivity presenting the results to the radiologist aiming to reduce false positives. This process is given in the first instance by a preprocessing algorithm based on advanced DIP techniques designed to reduce the noise acquired in the image and in an improvement of the same, and then it executed a segmentation process of different ROIs designed to detect high suspicion of some signs of cancer. By using the information obtained in the segmentation process, the classification of positives or negatives prediction of BC is obtained through a GRANN. After concluding the development of this CADx technological computer tool, it is planned to be used at real workplaces such as GHZ, making a validation of the prediction obtained by the neural network compared with predictions made by specialized oncologists aiming that it can be used as an aid in the early breast cancer diagnosis.

The aim of developing this CADx system for Mexican patients is to expand the knowledge of the database of BCD including image mammograms information, clinical data, risk factors, biopsy results, genomic information, etc., as is showed in **Figure 9**, at the bottom of the main window, with the aim to obtain more information on the resulting diagnostics such as degree of malignancy, time of presence, etc.

3. Conclusions

As mentioned before, conventionally, BCD and classification is performed by a clinician or a pathologist by observing stained biopsy images under the microscope. However, this is time consuming and can lead to erroneous results. Therefore, there is a rise in the need for developing intelligent and automated technology. As this problem can be computationally modeled as a problem of retrieving relevant information from a plethora of reports or tests, the development of a computationally efficient and detection-wise effective system for BCD can be seen as an issue from the field of KDD. Being cross-dimensional, KDD uses algorithms and techniques from a vast array of fields like soft computing, pattern recognition, machine learning statistics, artificial intelligence (AI), natural language processing (NLP), etc.

In this research, a method based on advanced DIP, KDD, and AI techniques was used for the extraction of fundamental features in DIM in order to detect breast lesions by using a GRANN. The used methodology was divided in two main stages. The first one used advanced DIP techniques for extracting image features of DMI in order to create a biomarker for BCD. With this information, in the second stage, a GRANN was trained and tested in order to classify BC.

After 2000 network trainings, a smoothing factor equal to 1e−4 was calculated. This value was used for training the neural net reaching an accuracy of 95.83%. The performance of trained GRANN was evaluated using four performance measures: accuracy, sensitivity, specificity, and precision. The results obtained in this work show that GRANN is a promising and robust system for BCD.

In a third stage, a CADx system based on AI technology is being designed in order to be applied in Mexican patients in collaboration with GHZ and MML. The proposed system aims to eliminate the unnecessary waiting time as well as reducing human and technical errors in diagnosing BC. The Laboratorio de Innovacion y Desarrollo Tecnologico en Inteligencia Artificial is seeking collaboration with research groups interested in validating the technology being developed.

Acknowledgements

This work was partially supported by Fondo Sectorial de Investigación para la Eduación under contract no. 241771 and PRODEP under contract no. 511-6/17-7763. The second and sixth authors want to thank the postdoctorate and doctorate scholarships, with scholarship holder numbers 25467 and 25976, respectively, received by Fondo Sectorial de Investigación para la Eduación under contract no. 241771. The fourth and fifth authors want to thank CONACYT PNPC scholarships, with scholarship holder numbers 285593 and 577554, respectively, received through Posgrado en Ingenieria y Tecnologia Aplicada, member of Programa Nacional de Posgrados de Calidad (PNPC) of CONACYT, México.

Authors want to thank the participation of Arturo Serrano Muñoz, an undergraduate student of Information Technology and Communication Bachelor from Universidad Tecnologica del Estado de Zacatecas, México, who is receiving a scholarship from PRODEP under contract no. 511-6/17-7763. This student performed the installation and configuration of the software and several services on the server where calculations were performed for extracting the features of the mammographic images and the neural network trainings.

Authors want to thank the participation of Edgar Viveros Llamas, an undergraduate student of Robotics and Mechatronics Bachelor from Universidad Autonoma de Zacatecas, Mexico. This student is making a social service in Laboratorio de Innovacion y Desarrollo Tecnologico en Inteligencia Artificial, helping with general activities related to this research.

Author details

Jose Manuel Ortiz-Rodriguez[1]*, Carlos Guerrero-Mendez[1], Maria del Rosario Martinez-Blanco[1], Salvador Castro-Tapia[1], Mireya Moreno-Lucio[1], Ramon Jaramillo-Martinez[1], Luis Octavio Solis-Sanchez[1], Margarita de la Luz Martinez-Fierro[2], Idalia Garza-Veloz[2], Jose Cruz Moreira Galvan[3] and Jorge Alberto Barrios Garcia[3]

*Address all correspondence to: morvymm@yahoo.com.mx

1 Laboratorio de Innovación y Desarrollo Tecnológico en Inteligencia Artificial, Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Zacatecas, México

2 Laboratorio de Medicina Molecular, Unidad Académica de Medicina Humana y Ciencias de la Salud, Universidad Autónoma de Zacatecas, Zacatecas, México

3 Tecnologías de Información y Comunicación, Universidad Tecnológica del Estado de Zacatecas, Guadalupe, Zacatecas, México

References

[1] Bernard WS, Christopher PW. World Cancer Report 2014. International Agency for Research on Cancer. WHO Press. World Health Organization; Switzerland. 2014

[2] Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the human development index (2008-2030): A population base study. The Lancet Oncology. 2012;**13**(8):790-801

[3] Weinberg RA, editor. One Renegade Cell: How Cancer Begins. Basic Books; New York, USA. 1999

[4] Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell. 1996;**87**(2):159-170

[5] Steyeberg E. Clinical Prediction Models: A Practical Approach to Development, Validation and Updating. Springer Science & Business Media; New York, USA. 2008

[6] Siegel R, Ward E, Brawley O, Jemal A. Cancer statistics. CA: A Cancer Journal for Clinicians. 2011;**61**(4):212-236

[7] Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, JA. Cancer treatment and survivorship statistics. CA: A Cancer Journal for Clinicians. 2016;**66**(4):271-289

[8] Love SM, Barsky SH. Breast-duct endoscopy to study stages of cancerous breast disease. The Lancet. 1996;**348**(9033):997-999

[9] Hortobagyi GN. Treatment of breast cancer. New England Journal of Medicine. 1998;**339**(14): 974-984

[10] Osborne C, Ostir GV, Du X, Peek MK, Goodwin JS. The influence of marital status on the stage at diagnosis, treatment and survival of alder women with breast cancer. Breast Cancer Research and Treatment. 2005;**93**(1):41-47

[11] Frank SA. Genetic predisposition to cancer — Insights from population genetics. Nature Reviews. 2004;**5**(10):764

[12] Sickles EA. Mammographic features of “early” breast cancer. American Journal of Roentgenology. 1984;**143**(3):461-464

[13] Nover AB, Jagtap S, Anjum W, Yegingil H, Shih WY, Shih WH, Brooks AD. Modern breast cancer detection: A technological review. Journal of Biomedical Imaging. 2009;**26**

[14] Kollias J, Elston CW, Ellis IO, Robertson JF, Blamey RW. Early-onset breast cancer — Histopathological and prognostic considerations. British Journal of Cancer. 1997;**75**(9):1318

[15] Krishnamurthy S, Mathews K, McClure S, Murray M, Gilcrease M, Albarracin C, Cohen A. Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin–stained breast tissue sections. Archives of Pathology and Laboratory Medicine. 2013;**137**(2):1733-1739

[16] Ganesan K, Acharya UR, Chua CK, Min LC, Abraham KT, Ng KH. Computer-aided breast cancer detection using mammograms: A review. IEEE Reviews in Biomedical Engineering. 2013;**6**:77-98

[17] Gelasca ED, Obara B, Fedorov D, Kvilekval K, Manjunath BS. A biosegmentation benchmark for evaluation of bioimage analysis methods. BMC Bioinformatics. 2009;**10**(1): 368

[18] Shamir L, Delaney JD, Orlov N, Eckley DM, Goldberg IG. Pattern recognition software and techniques for biological image analysis. PLoS Computational Biology. 2010;**6**(11)

[19] Kaplan SS. Clinical utility of bilateral whole-breast US in the evaluation of women with dense breast tissue. Radiology. 2001;**221**(3):641-649

[20] Rosenblatt P, Suzuki I, DeRidder A, Patel N. Breast cancer survivorship. Handbook of Breast Cancer and Related Breast Disease; New York, USA. Demos Medical Publishing. 2016

[21] Preventive Services US. Task force. Screening for breast cancer: Recommendations and rationale. Annals of Internal Medicine. 2002;**137**(5):344

[22] Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Jong R. Diagnostic performance of digital versus film mammography for breast-cancer screening. New England Journal of Medicine. 2005;**353**:1773-1783

[23] Júnior G, de Oliveira Martins L, Silva A, de Paiva A. Computer-Aided Detection and Diagnosis of Breast Cancer Using Machine Learning Texture and Shape Features. IGI Global; Hersey, PA, UZA. 2010

[24] Velayutham C, Thangavel K. Unsupervised feature selection in digital mammogram image using rough set based entropy measure. Information and Communication Technologies (WICT), 2011 World Congress; 2011

[25] Gupta M, Jin L, Homma N. Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory. John Wiley Sons: New Jersey, USA; 2003

[26] Huang DS. Radial basis probabilistic neural networks: Model and applications. International Journal of Pattern Recognition and Artificial Intelligence. 1999;**13**(7):1083-1101

[27] Mao K, Tan K, Ser W. Probabilistic neural network structure determination for pattern classification. IEEE Transactions on Neural Networks. 2000;**11**(4):1009-1016

[28] Spetch DF. Probabilistic neural networks for classification, mapping or associative memory. IEEE International Conference on Neural Networks. 1998;**1**:525-532

[29] Spetch DF. Probabilistic neural networks. Neural Networks. 1990;**3**(1):109-118

[30] Spetch DF, Romsdhal H. Experience with adaptive probabilistic neural networks and adaptive general regression neural networks. IEEE International Conference on Neural networks. 1994;**2**:1203-1208