

Research Article

Automatic Quality Evaluation of Digital Mammographic Images Generated with Cdmam Phantom Correlated With the Human Vision

Maria AZ Sousa¹, Bruno Barufaldi², Regina B Medeiros³, and Homero Schiabel^{1*}

¹Department of Electrical and Computer Engineering, EESC, University of S. Paulo, Brasil

²Department of Radiology, University of Pennsylvania, USA

³Department of Imaging Diagnosis, EPM, Federal University of S. Paulo, Brasil

***Corresponding author**

Homero Schiabel, Associate Professor, Av. Trabalhador Saocarlense, 400, EESC/USP, 13566-590 Sao Carlos, SP, Brasil; Tel: +55 16 33739355; Fax: +55 16 33739372.

Submitted: 20 June 2023

Accepted: 10 July 2023

Published: 11 July 2023

ISSN: 2333-7095

Copyright

© 2023 Sousa MAZ, et al.

OPEN ACCESS

Keywords

- Digital mammography
- Quality control in mammography
- Phantom CDMAM
- Human visual perception

Abstract

The aim of this work was to develop a software tool that assists in the testing of image quality in mammography, addressing the challenges associated with the subjectivity and time-consuming nature of manual measurements. The software aims to correlate automated readings with the human visual system, eliminating the need for result correction, which is commonly required in many existing studies. To achieve this, a dataset of 46 images acquired from exposures of the phantom CDMAM to five computed radiography (CR) systems was used. The method employed for image quality assessment involved the use of circular correlator filters for detection. The correlation with human vision was based on Weber's parameters, which describe how the visual system discriminates contrast in digital images. The classification of image disks as visible or not visible was performed using the WEKA (Waikato Environment for Knowledge Analysis) data-mining tool in combination with the J48 algorithm, which facilitated the construction of decision tree models. The implementation of decision trees resulted in a software system that aids specialists in image quality assessment. The system provides stable and easily interpretable results, achieving accuracy rates of up to 95%. By automating the assessment process and reducing the dependence on observers, the software enhances the integrity of the evaluation and improves the accuracy of measurements.

ABBREVIATIONS

ACC: system accuracy; AUC: area under the ROC curve; CDC: Contrast-detail curve; COR : Correct observation ratio; CR : computed radiography; EFF: efficiency; FFDM: full-field digital mammography; FOM: Figure-of-merit; IQF: Image quality figure; κ -Kappa value; NPV: Negative Prediction Value; PPV: Positive Prediction Value; ROC-Receiver Operating Characteristic; TNR: true negative rate; TPR: true positive rate; WEKA: Waikato Environment for Knowledge Analysis.

INTRODUCTION

Statistics derived from the World Health Organization [1] indicate that the global count of breast cancer was estimated in 2.3 million new cases, as reported in 2020. This accounted for about 11.7% of all new cancer cases worldwide with an estimated surge in the number of new cases to 15.5 million by 2030. Among developed nations, breast cancer stands out as one of the most prevalent forms of cancer and a significant contributor to the elevated mortality rates among women worldwide. The majority of breast cancer deaths occur in low and middle-income countries, where late-stage diagnoses are more prevalent due to inadequate access to healthcare services and a lack of information.

To facilitate early detection, many healthcare institutions have implemented mammographic screening programs, which serve as the sole preventive measure against this disease. Early detection, coupled with appropriate treatment, substantially reduces the risk of fatality, enabling patients to lead a cancer-free life [2]. Consequently, screening mammography is considered the primary strategy for population-wide examinations.

The primary challenge in interpreting mammography images lies in the low contrast and size of malignant features, such as microcalcifications [3]. The ability to observe, extract, quantify, and interpret data for detecting and characterizing breast diseases is subjective and varies according to the expertise of the specialist. This capacity has implications for the number of unnecessary biopsies, leading to increased examination costs, and, in extreme cases, the possibility of failing to detect a lesion.

A significant advancement in mammography came with the introduction of full-field digital mammography (FFDM) and, alternatively, CR (Computed Radiography) systems. These technologies have largely replaced screen-film systems due to their extensive dynamic range, linear relationship between dose and signal intensity, the option of contrast enhancement through image processing algorithms, computer-aided diagnosis, and

immediate availability of images on the hospital information network [4].

Quality evaluation of these systems is crucial to ensure adequate visualization of lesions by radiologists. Various standards providing guidance for quality control have been developed worldwide [5-8]. These documents specify the parameters that must be measured regularly to ensure proper equipment functioning as per the manufacturer's guidelines. Many of these measures require obtaining parameters from the image of a breast phantom, and their accuracy is paramount as they pertain to elements sensitive to image quality.

The European Protocol recommendations [8], encompass the examination of parameters based on contrast-detail measurements obtained from images generated using the CDMAM phantom - *Artinis Contrast-Detail Phantom* [9] - specifically designed for this purpose. The method entails visual inspection of these images on a high-resolution display by one or more specialists to determine the disk thickness that falls within the visibility threshold among those present within the phantom. However, automatic measurement procedures using computational techniques to quantify mammography image quality can significantly streamline the routine tasks of skilled professionals and mitigate evaluation subjectivity, thereby enhancing the accuracy of results.

With this aim, this study was conducted to develop a computational tool for managing the interpretation of images generated by exposing the CDMAM phantom to digital mammography equipment with pre-certified quality parameters. The aim was to automate the image reading and calculation of four parameters central to image quality: contrast- detail curve, correct observation ratio (COR), image quality figure (IQF), and figure of merit (FOM).

This computational tool is expected to serve as an aid in quality assurance reporting, presenting easily analyzable results already correlated with human vision. It obviates the need for correction formulas used in other schemes or techniques previously described in the literature [10,11]. A comprehensive study was conducted on the human visual system and Artificial Intelligence techniques for feature extraction and object classification in digital images, specifically for this purpose. As a quantification strategy, the signal/detection ratio was determined using the statistical model of Receiver Operating Characteristic (ROC) curves [12].

MATERIALS AND METHODS

CDMAM phantom characteristics and recommendations

The CDMAM phantom, also known as the *Artinis Contrast-Detail Phantom* [9], consists of an Aluminum (Al) base with gold disks - designed for contrast inspection, with thicknesses ranging from 0.03 to 2.0 μm , as illustrated in Figure 1(a). The positioning template of the disks is depicted in Figure 1(b). The gold disks are arranged in a 16x16 matrix, with two identical

disks placed in each cell. One disk is always positioned at the center of the cell, while the other is randomly placed at one of the vertices. Within each row of the matrix, the diameter of the disks remains constant, while the thickness increases logarithmically. Conversely, within each column, the thickness remains constant, while the diameter increases logarithmically.

To simulate the conditions faced in conventional mammography systems (with Mo anode target, Mo filter of 30 μm , and 28 kVp), the Al base is covered with acrylic (PMMA), resulting in a 10.0mm thickness. The CDMAM phantom includes 4 PMMA plates, each with a thickness of 10.0mm, which simulate various breast thicknesses. Typically, the phantom is placed on the Bucky with the disks of smallest diameters facing the thorax side, combined with the PMMA plates placed over the disks matrix. Following the recommendations outlined in the European Protocol [8], the test should be conducted annually, involving capturing six images with minimal processing while repositioning the phantom on the Bucky for each exposure. At least three experienced observers should independently interpret two different images, cross-referencing the actual positions of the disks using the provided template shown in Figure 1(b). The identified disk diameters should fall within the range of 0.1 to 2.0mm. Within this range, the minimum visible contrast for each diameter can be determined. Subsequent tests should display at least five distinct details [13].

Images acquisition for evaluation

Exposures were performed with the CDMAM phantom in a *Senographe Essential (GE Healthcare)* unit as well as in a *Selenia (Hologic Inc.)* one. In addition, five computed radiography (CR) systems (models *Fuji 50*, *Fuji 100*, *Kodak 975*, and *Agfa 85*) were used to obtain digital images from four other different sites where there were no FFDM systems [14]. A total of 75 images were acquired by using the mammography systems characterized as follows.

GE Senographe Essential and Hologic Selenia - typical full field digital mammography (FFDM) systems yielding breasts images with 12-bit contrast resolution;

Fuji 50, Fuji 100, Kodak 975 and Agfa 85 - CR systems with a minimum contrast resolution of 12 bits and a pixel size of 50 μm (except for Fuji 100, with pixel size of 100 μm).

Technique for disk detection

After testing the prototype with a large number of images, methods were refactored to process CDMAM images as input. Accordingly, the first version of the software was implemented and compared with the CDCOM [15], commonly used for reading CDMAM images. The CDCOM method for determining the threshold gold thickness is based on a fit of a psychometric curve for each detail diameter, after applying the nearest neighbor correction (NNC) in the reader analysis [8]. However, as previously described, the current study shows a different method for thresholding, using learning models in accordance

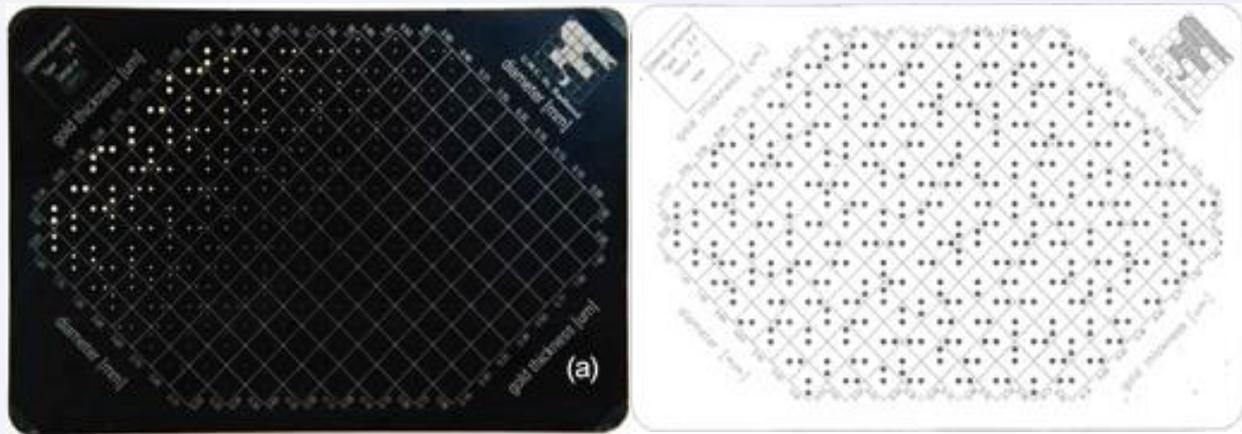


Figure 1 (a) Frontal view of phantom CDMAM 3.4 – Artinis Contrast-Detail Phantom. (b) Positions template of the gold disks.

with multiple visual inspections. Unlike the CDCOM program, the developed software was implemented using the Java programming language with *ImageJ* [16], an open source software aimed to the development of image processing applications and analysis.

The automation of CDMAM image reading was designed to detect the disks present in the images and classify them as either visible or not visible, based on human vision criteria. The detection process used circular correlation filters [17], which consisted of an inner region and an outer concentric region.

Figure 2 illustrates an example of the procedure specifically applied to typical phantom images. Figure 2(a) shows that the filters used in the correlation operations are composed of two parts, which correspond to inner and outer regions. Figure 2(b) demonstrates when the inner region matches the inner structure and the outer region matches the background. The filter's radii (r_1 and r_2) changes according to the size of each structure (inner or outer region).

Contrast Threshold

In our computational methodology, the determination of visibility classification for each target structure in the phantom images is accomplished using the J48 algorithm, which is a component of the WEKA data-mining tool [18]. The selection of the J48 algorithm was based on its proven simplicity, performance, accuracy, and effectiveness in previous research studies [18,19]. During the training stage, the algorithm generates a decision tree model for each structure type, utilizing pre-selected image features. The selected features encompass the average pixel intensity, standard deviation, mode, average intensity of structure pixels, average intensity of background pixels, the difference between structure and background average intensities, and the Weber Ratio.

A total of 24 CDMAM images were employed for training purposes, resulting in 13 learning models for different structure

diameters (ranging from 2.00mm to 0.13mm). The leave-one-out technique was applied to create the learning models, wherein the input features were iteratively excluded to enhance the accuracy and efficiency of the training process. The attribute selector offered by WEKA was used to improve the model's performance and reduce training effort.

Once the learning models were obtained, they were integrated into the software for classification purposes. An independent set of 51 CDMAM images was processed for testing, thereby evaluating the performance of the developed prototype. The software implementation was carried out in the Java programming language using *ImageJ* [16], an open-source software specifically designed for image processing applications and analysis. A comparison was conducted between the developed software and the CDCOM method [15].

Statistical metrics such as system accuracy (ACC), true positive rate (TPR), true negative rate (TNR), efficiency (EFF), positive prediction value (PPV), negative prediction value (NPV), Kappa (κ) value, and area under the ROC curve (AUC) [13], were determined by comparing the results of the software with the technical expert reports obtained during the testing procedure. These statistical measures offer valuable insights into the performance and reliability of the developed software.

Features extraction and classification

During a visual inspection of an image to determine the objects visibility, it is common to classify a light-colored object on a dark background or a dark object on a light background as visible, in accordance with Weber's law [21]. Weber's law states that there is a contrast threshold between the object and the background, which determines whether the object is perceivable. However, this threshold is subjective and dependent on the lighting conditions during analysis as well as the human visual system. Thus, relying on a single feature to interpret image information may be insufficient to accurately describe object visibility.

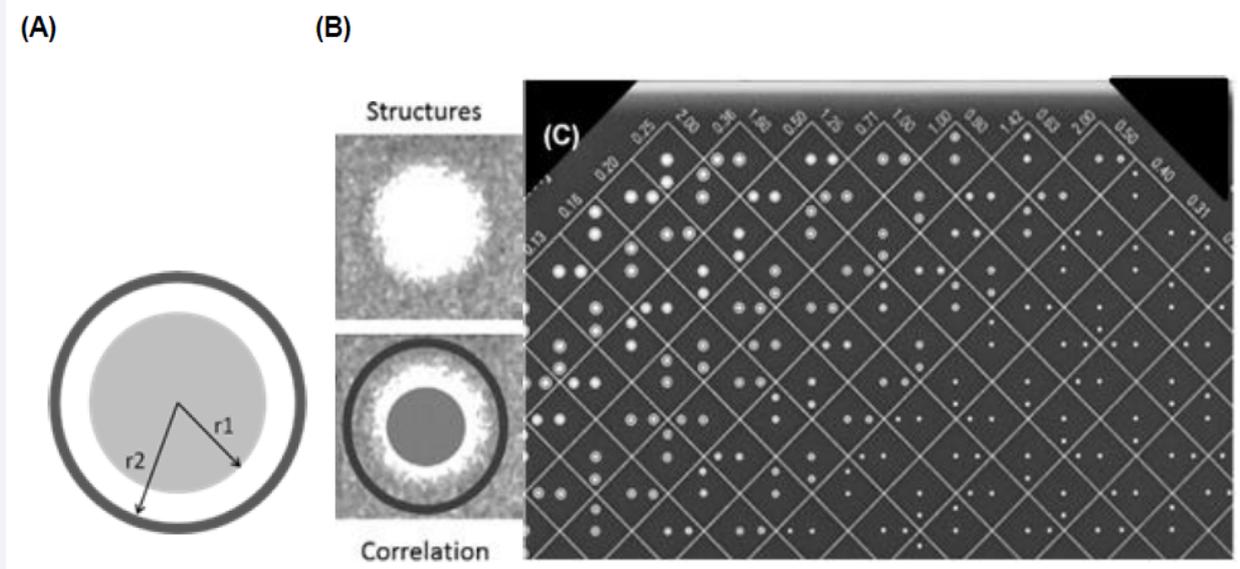


Figure 2 (A) Filter model used for disks detection. The r1 and r2 radii vary with the radius of the disk analyzed, according to a reference image. Filters were made by varying their diameters according to the diameters of the disks existent in the phantom image used as template; (B) example of a structure in the CDMAM phantom image. (C) Simulation of results, after locating the structures of interest.

When automated processes are used to identify objects in digital images, it becomes necessary to select and extract multiple specific characteristics. These characteristics are typically represented as a vector, which can be inputted into an automatic classifier [22]. In this study, predefined characteristics were used for the classification models: the gray levels of the image pixels, the pixel values of the disk obtained from the inner region of the circular filter (Figure 2), and the pixel values of the background corresponding to the outer region of the filter. Additionally, a region around each disk was selected, and other features were calculated from this region for training purposes [23]. Table 1 provides a description of each feature employed in the analysis, where (i, j) represents the position of the filter displacement on the image.

In this study, eight features were extracted from the images, and a statistical investigation was conducted to identify the features that best represented the natural patterns of each class (visible and not visible).

A total of 2,542 disks from 24 selected images were used exclusively for training purposes. Half of these disks were classified by experts as visible, while the other half were classified as not visible, ensuring the generation of reliable learning models.

For each diameter of the disks extracted from the images, a decision tree model [23], was generated during the training step. However, it was not possible to generate trees for diameters smaller than 0.13mm due to their poor visibility. As a result, a total of 13 decision trees were obtained, covering diameters ranging from 2.00mm to 0.13mm. Table 2 illustrates WEKA results for the training stage, considering some of the evaluation parameters mentioned at the end of section 2.4.

Table 1: Features extracted from the image after detection of structures (disks).

Feature	Equation
Average pixel values of the disk	$\mu = \frac{1}{e} \sum_{i=1}^w \sum_{j=1}^h p(i, j)$
Average pixel values of the background	$\mu_b = \frac{1}{b} \sum_{i=1}^w \sum_{j=1}^h p(i, j)$
Difference of average pixel values of the disk and background	$\Delta\mu = \mu_e - \mu_b$
Weber ratio	$W = \frac{\Delta\mu}{\mu_e}$
Average pixel values of the image	$\mu = \frac{1}{i} \sum_{i=1}^w \sum_{j=1}^h p(i, j)$
Average pixel values of the equalized image	$\mu = \frac{\sum_{i=1}^w \sum_{j=1}^h p_{eq}(i, j)}{eq}$
Image Variance	$v = \frac{1}{i} \sum_{i=1}^w \sum_{j=1}^h (p(i, j) - \mu)^2$
Image mode	Gray level more frequent in p(i,j)

Table 2: WEKA results for the training of contrast-details (n golden disks) in the CDMAM images. n+, n- represent the number of positive and negative structures.

	Training												
	Diameter (mm)												
	2.00	1.60	1.25	1.00	0.80	0.63	0.50	0.40	0.31	0.25	0.20	0.16	0.13
ACC	85%	84%	90%	89%	92%	89%	97%	92%	94%	91%	81%	68%	95%
TPR	87%	80%	91%	90%	92%	91%	98%	95%	93%	92%	65%	63%	65%
TNR	83%	88%	89%	86%	92%	88%	97%	89%	95%	90%	96%	73%	98%
AUC	0.77	0.74	0.85	0.94	0.90	0.82	0.97	0.90	0.94	0.90	0.70	0.76	0.70
n+	48	58	68	70	78	78	88	76	60	52	52	40	10
n-	48	58	68	70	78	78	88	76	60	52	52	40	10
n total	96	116	136	140	156	156	176	152	120	104	104	80	20

A low variation of detection sensitivity and specificity rates for details with diameters between 1.25 mm and 0.25 mm can be noticed. The learning models could fit better for these diameters. On the other hand, for the smallest disks (<0.20 mm), the TNR (true negative rate) is always higher (even with low sensitivity).

To validate the models, the cross-validation technique was employed. This involved setting aside a portion of the dataset as “training data” to estimate the results, while using the remaining data as “validation data” to assess the model’s performance. Since the number of samples for each diameter varied, 10 samples were selected for training purposes to ensure a standardized process for generating decision trees.

Once the decision trees were obtained, the classification models could be implemented to automatically detect and classify the disks, aligning with human vision. Throughout the development process, adjustments were made to terminate the search whenever two consecutive regions along a line, encompassing a fixed diameter, were not found in any record. Examples of such cases are illustrated in Figure 3.

To ensure a closer resemblance to the human reading process, the marking of disks classified as “visible” was established based on specific criteria. In the classification process, a cell (region) was marked as “visible” if both disks within the cell or only a corner disk were detected. However, if only the central disk was found in a cell, that cell was disregarded in the procedure. This criterion aimed to align the program’s classification result with the approach typically followed by human readers.

For instance, in the example depicted in Figure 3, this criterion is illustrated for the thickness of $0.50\mu\text{m}$, which corresponds to the third region analyzed for the 0.80mm diameter line. The program dismisses the cell where only the central disk is located, as per the established limitation. These limitations imposed on the program ensure that the final classification outcome is in closer agreement with the human reading process, enhancing the reliability and accuracy of the classification results.

The results obtained from the disks classification in the previous stage were used to determine four image quality parameters [24,25] to be used as metrics to determine the software performance:

Contrast-detail curve (CDC) – the graphical correlation between the minimum diameter and thickness of each disk phantom in relation to a true reading;

Correct observation ratio (COR) – calculated as the ratio between the number of correctly identified disks (N_c) and the actual total number of disks (N_t), multiplied by 100;

Image quality figure (IQF) – determined by summing the product of the smallest scored object diameter ($D_{i,min}$) and its corresponding contrast (thickness) values;

Figure-of-merit (FOM) – used to assess the effects of dose and serves as a standard for comparing image quality.

Interface

To make easier the integration of all methods involved in the disks detection and classification process a user-friendly interface was designed, with a button to select the image to be processed. The software is compatible with TIFF and DICOM files, supporting images with 10, 12, 14, or 16-bit contrast resolution. It is important to position the images correctly to ensure the headers indicating the thickness and diameter are properly displayed, as illustrated in Figure 4.

Once the image is selected, it is displayed in the interface and positioned optimally for processing. If there are any alignment errors, the user can manually correct them by selecting two points belonging to a line on the image. The software will automatically adjust the alignment. During this step, the zoom tool is enabled, allowing the user to zoom in for better accuracy. After the alignment is complete, the processing can be run by selecting the starting point for the search, which should correspond to the center of a disk with a diameter of 2.00mm and a thickness of $0.25\mu\text{m}$. The starting point can be selected on the image using the zoom tool.

The software will search for the disks considered visible. The processed image will be marked with a black circle around each detected disk, and it will be displayed in the interface screen. At this stage, buttons will be enabled to generate the contrast-detail curve, which can be saved as an image. The parameters such as the correct observation ratio (COR), image quality figure (IQF),

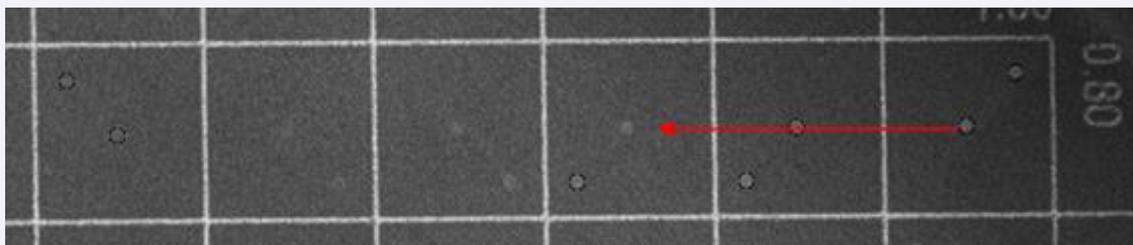


Figure 3 Image of an example where no disk has been located in two consecutive regions along the line of diameter = 0.80 mm. In this case, the subsequent disks would be marked as detected, but the restriction added to the system does not allow proceeding with the searching. On the other hand, the central disk of the third region (cell) would be marked as visible by the system.

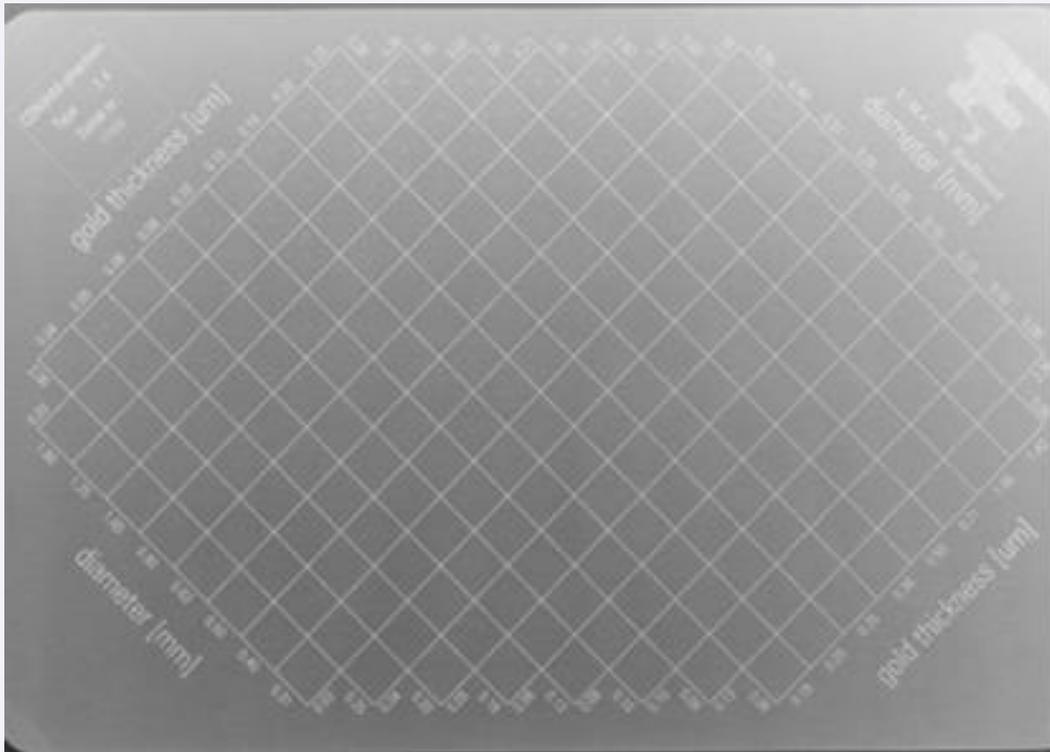


Figure 4 Example of an input image for the developed software.

and figure-of-merit (FOM) will also be calculated and displayed in the interface screen. The user can save all these results in a spreadsheet file.

To provide assistance and guidance, a “Help” button has been included in the interface, allowing users to access information on the proper usage of the computational scheme. The schematic example of the interface screen can be seen in Figure 5, providing users with a visual representation of the interface layout.

RESULTS AND DISCUSSION

The best selected characteristics for each disk diameter are presented in Table 3. For most diameters, the average pixel value of the disk (gray levels) and the average pixel value of the background were identified as the most relevant characteristics. Contrary to expectations, the contrast between the disk and the background did not appear to be as significant in this analysis. This tendency indicates that each feature was considered individually rather than focusing solely on their relative relationships. The selection of these key characteristics plays a crucial role in developing accurate and efficient classification models. By identifying the most informative features, the models can effectively differentiate between visible and non-visible disks, contributing to the overall performance and reliability of the automated classification process.

The average accuracy of 80.22% achieved during the selection of features can be reproduced when implementing

the decision tree in the final disks classification algorithm. This indicates that the chosen features, combined with the pruning process, contribute to the effectiveness of the classification models in accurately distinguishing between visible and non-visible disks. The decision tree algorithm, with its ability to create interpretable models, provides a reliable framework for automating the classification process based on the selected features.

The tests conducted on 51 images, excluding those used for training, provided valuable results for determining the quality parameters of the evaluated images. These parameters were expressed through contrast-detail curves and analyzed statistically using ROC curves [13], which measure the rate of true detections for each diameter under investigation. To establish a reference, contrast-detail curves were previously determined using measurements from five observers on a high-resolution display model BARCO E- 3620 [25]. A comparison between the contrast-detail curves generated by the developed scheme and their respective reference curves revealed a high degree of agreement between the data obtained from automatic classification and the technical report for the majority of the images.

The effectiveness of the classification process is demonstrated by the accuracy rates, which exceeded 70% (as shown in Table 4), and the ROC curves generated for each diameter. The areas under the ROC curves (AUC) were found to be greater than 0.72, as indicated in Table 5. These results further validate the

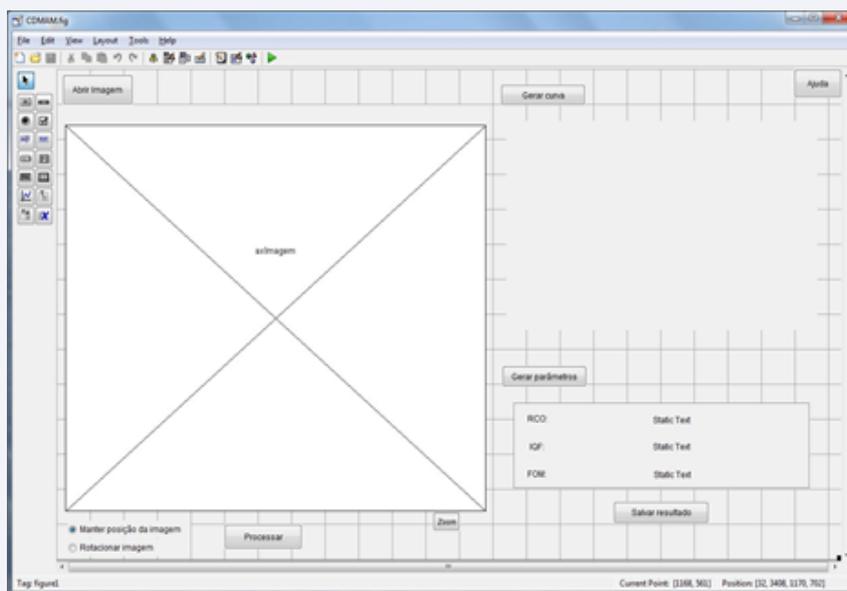


Figure 5 Schematic example of the interface screen.

Table 3: Accuracy rate for the disks detection according to the classification models generated and characteristics selected for each disk diameter.

Diameter (mm)	Sample	Accuracy rate (%)	Characteristics
2.00	160	80.63	Contrast
1.60	194	85.57	Contrast, Average pixel values of the disk
1.25	223	87.00	Contrast, Average pixel values of the disk
1.00	251	87.65	Contrast, Average pixel values of the disk
0.80	266	79.32	Average pixel values of the background, variance, average pixel values of the image, average pixel values of the disk
0.63	280	78.21	Contrast, average pixel values of the disk, weber ratio, average pixel values of the background, mode
0.50	292	73.63	Average pixel values of the disk, average pixel values of the background, variance
0.40	256	83.98	Average pixel values of the disk, mode
0.31	206	83.98	Average pixel values of the disk, average pixel values of the background, Mode
0.25	170	72.94	Structure Average
0.20	130	79.23	Average pixel values of the disk, average pixel values of the background, Mode
0.16	90	80.00	average pixel values of the image, average pixel values of the background
0.13	24	70.83	Average pixel values of the disk
TOTAL	2542	80.22	

Table 4: Contingency table and accuracy rates for the classification obtained by the developed scheme.

Diameter (mm)	Sample	TP	TN	FP	FN	Accuracy rate (%)
2.00	153	73	54	5	21	83.01
1.60	187	90	82	4	11	91.98
1.25	204	104	85	4	11	92.65
1.00	221	110	94	9	8	92.31
0.80	238	64	106	3	65	71.43
0.63	255	78	125	0	52	79.61
0.50	272	132	111	28	1	89.34
0.40	272	97	147	11	17	89.71
0.31	272	81	164	15	12	90.07
0.25	255	82	156	16	1	93.33
0.20	238	60	165	9	4	94.54
0.16	221	36	157	22	6	87.33
0.13	204	6	170	20	8	86.27
TOTAL	2992	1008	1614	144	217	87.63

(TP: true positive; TN: true negative; FP: false positive; FN: false negative cases).

performance of the classification models and the overall success of the developed scheme in accurately detecting and classifying disks. Overall, the scheme exhibits promising accuracy rates and AUC values, indicating its potential for practical application in automatic disk classification and image quality assessment.

The variation observed in the values of the areas under the ROC curves (AUC) is worth noting, as they can differ by up to 25% between the best and worst results. Some diameters may exhibit easier detection than others, leading to varying AUC values. However, when considering the average AUC value across all diameters, it remains consistent with the AUC value obtained for the overall scheme (AUC = 0.86). This can be observed by comparing the data from Table 5 and Figure 6(a).

Additionally, Figure 6(b) displays the ROC curve generated by the CDCOM software using the nearest neighbor correction (NNC) for the same set of images tested with the developed computational tool. The AUC value obtained using the developed tool is slightly higher compared to the CDCOM software, indicating a better performance in terms of classification accuracy.

These results highlight the effectiveness of the developed computational tool in achieving accurate classification and image quality assessment. The average AUC value of 0.86 indicates a reliable and robust performance across different disk diameters, demonstrating its potential for practical application in the automated analysis of CDMAM images.

The evaluation of the scheme's performance can be further enhanced by analyzing the visibility potential of each disk phantom and calculating the true classification rates of the software compared to technical reports. Table 6 provides an overview of this evaluation, highlighting the diameters where the software achieved 100% accuracy in detecting disks across all analyzed images. Additionally, it identifies diameters where

misclassifications were more likely, leading to a decrease in the true response rate for those specific diameters. For example, the diameter of 0.63mm with a thickness of 0.20 μ m shows a lower accuracy rate. It is important to attempt that the mean values obtained for each diameter in Table 6 are very similar to those shown in Table 5, further confirming the accuracy and consistency of the scheme's results.

Similar data were collected for tests performed with the CDCOM software, as presented in Table 7, allowing for a comparison of performance between the developed computational tool and the CDCOM software.

The comparison between the computational tool developed in this work and the CDCOM software revealed a significant improvement in the accuracy rates, with an increase of up to 30% in the mean values calculated for each diameter. For diameters of 0.08mm and 0.63mm, the difference in accuracy rates between the two techniques was less than 5%, indicating a correlation in the difficulty of detecting these specific disk diameters. Similarly, for the 0.20mm and 0.16mm diameters, the difference was less than 2%, but in these cases, the correlation is due to the inability to detect such disks, resulting in a high number of true negative cases.

The most noteworthy results from the comparison between the two techniques are highlighted in Tables 6 and 7, where the minimum accuracy rates are even 50% higher for the reading obtained with the computational tool proposed in this work.

Regarding the image quality parameters determined by the software, the best results indicate lower values for the correct observation ratio (COR) and higher values for the image quality figure (IQF). However, accurate conclusions about the figure-of-merit (FOM) cannot be drawn. Table 8 provides an example of the numbers achieved for five different images, illustrating the

Table 5: Results obtained from ROC curves generated for each disk diameter under study.

Diameter (mm)	2.00	1.60	1.25	1.00	0.80	0.63	0.50	0.40	0.31	0.25	0.20	0.16	0.13
Area under curve (AUC)	0.84	0.89	0.85	0.90	0.73	0.78	0.89	0.89	0.89	0.96	0.94	0.89	0.72

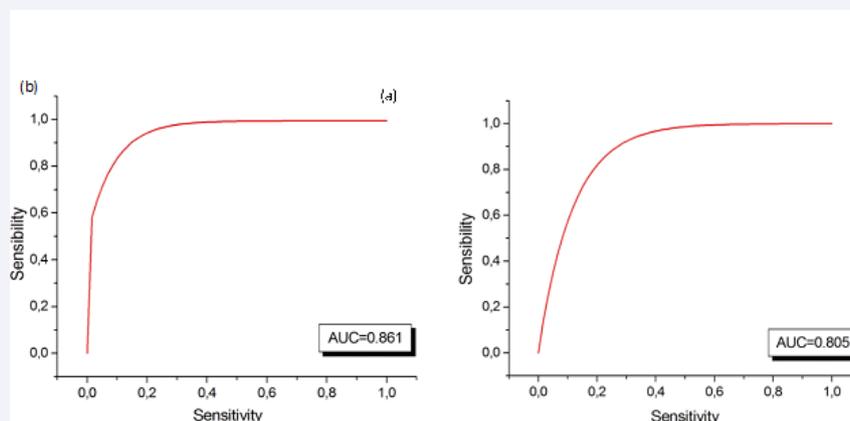


Figure 6 ROC curves generated for the classification results from (a) the computational tool developed and (b) CDCOM software.

Table 6: Accuracy rate of the software classification in relation to technical reports, considering each disk individually.

		Thickness (µm)																
		0.03	0.04	0.05	0.06	0.08	0.10	0.13	0.16	0.20	0.25	0.36	0.50	0.71	1.00	1.42	2.00	Average
Diameter (mm)	2.00		100	100	94	47	29	94	94	100	100							84
	1.60	100	100	100	94	82	59	82	94	100	100	100						92
	1.25	100	100	100	100	76	59	94	94	94	94	100	100					93
	1.00	100	100	100	100	59	59	71	94	94	100	100	100	100				90
	0.80	100	100	100	100	94	82	35	35	47	47	53	53	59	100			72
	0.63	100	100	100	100	100	100	82	41	18	29	71	76	88	88	100		80
	0.50	100	100	100	100	100	100	24	41	82	88	100	100	100	100	100	100	90
	0.40	100	100	100	100	100	100	100	100	100	53	29	76	94	94	100	100	90
	0.31	100	100	100	100	100	100	100	100	100	94	53	71	82	88	100	100	93
	0.25		100	100	100	100	100	100	100	100	100	100	35	94	100	100	100	95
	0.20			100	100	100	100	100	100	100	100	100	94	65	71	94	100	95
	0.16				100	100	100	100	100	100	100	100	100	94	29	29	88	87
	0.13					100	100	100	100	100	100	100	100	100	93	100	36	57

Table 7: Accuracy rate of the CDCOM software automatic reading in relation to technical reports, considering each disk individually.

		Thickness (µm)																
		0.03	0.04	0.05	0.06	0.08	0.10	0.13	0.16	0.20	0.25	0.36	0.50	0.71	1.00	1.42	2.00	Average
Diameter (mm)	2.00		42	42	27	27	46	100	100	100	100							65
	1.60	92	69	50	27	23	46	100	100	100	100	100						73
	1.25	100	73	54	23	27	35	100	100	100	100	100	100					76
	1.00	100	88	62	38	35	38	38	100	100	100	100	100	100				77
	0.80	100	88	73	54	38	27	27	46	81	85	85	85	85	100			70
	0.63	100	100	96	73	54	27	77	27	42	77	92	92	100	100	100		77
	0.50	100	100	96	81	77	62	8	15	54	77	100	100	100	100	100	100	79
	0.40	100	100	100	92	100	77	50	42	27	77	73	96	96	100	100	100	83
	0.31	100	100	100	100	100	96	77	58	46	12	92	92	92	100	100	100	85
	0.25		100	100	100	100	100	96	88	81	50	15	50	85	100	100	100	84
	0.20			100	100	100	100	100	100	100	81	69	96	62	96	100	100	93
	0.16				100	100	100	100	100	100	92	81	54	8	96	92	96	86
	0.13					94	94	94	100	94	94	88	100	47	24	41	88	80

Table 8: Image quality parameters and accuracy calculated by the program developed for five images tested.

Image	COR	IQF	FOM	Accuracy (%)
1	282.86	0.16	14.69	0.91
2	260.53	0.09	66.35	0.83
3	319.35	0.16	1.12	0.86
4	257.14	0.26	-68.27	0.95
5	282.86	0.11	-135.78	0.91

performance of the software in terms of these image quality parameters.

In a more detailed analysis, Table 9 shows the accuracy rates, sensitivity, specificity, efficiency, positive prediction and negative prediction, Kappa values, and area under the ROC curve (AUC). Following the trend of models obtained by the training step, the software efficiency is higher (>90%) for details between 1.25mm and 0.25mm.

The results obtained in this study confirm the nonlinear relationship between the image quality parameters and the accuracy ratio, as previously observed in Thomas' work [24]. These findings emphasize the importance of exercising caution when analyzing data from different CDMAM images, as the relationship between the parameters may vary. The perfect

correlation observed between the correct observation ratio (COR) and the accuracy ratio for each image should be highlighted. Identical values were registered for these two parameters, indicating their strong association and reliability in assessing the performance of the classification system.

CONCLUSIONS

The methodology presented in this study has successfully developed a tool that aids specialists in preparing the final report on the quality of CDMAM images. The tool provides stable and easily interpretable results, strengthening the integrity of the evaluation process. Furthermore, it has the potential to be refined and integrated into a comprehensive system for image quality evaluation in mammography, enhancing the availability of quality certification processes in line with national and international standards.

A key focus during the development of this software was to achieve automatic readings that were as close as possible to the human visual assessment, and without the need to correct the results – as it happens in many systems found in literature. This aim was achieved by only using automatic classification techniques, which improved upon the current state-of-the-art represented by the CDCOM tool.

Table 9. Developed software efficacy for classification of all gold disks in the CDMAM 3.4 images. Results are displayed in descending order according to the diameter of the disks.

	Diameter (mm)												
	2.00	1.60	1.25	1.00	0.80	0.63	0.50	0.40	0.31	0.25	0.20	0.16	0.13
ACC	78%	89%	91%	90%	89%	94%	93%	94%	95%	95%	90%	89%	96%
TPR	61%	83%	91%	88%	81%	91%	87%	91%	92%	94%	57%	69%	66%
TNR	97%	95%	90%	91%	96%	96%	97%	96%	97%	96%	100%	93%	99%
EFF	79%	89%	91%	90%	90%	94%	92%	93%	95%	95%	78%	81%	82%
PPV	96%	93%	90%	91%	96%	95%	96%	93%	93%	90%	100%	70%	88%
NPV	70%	86%	91%	89%	84%	93%	90%	95%	96%	98%	88%	93%	97%
κ	0.65	0.83	0.86	0.85	0.83	0.91	0.88	0.90	0.91	0.90	0.68	0.60	0.72
AUC	0.82	0.91	0.94	0.95	0.95	0.97	0.95	0.98	0.98	0.97	0.72	0.75	0.72

(ACC: accuracy; TPR: true positive rate; TNR: true negative rate; EFF: efficiency; PPV: positive prediction value; NPV: negative prediction value; κ : Kappa value; AUC: area under the ROC curve).

A critical aspect of automatic classification is the proper selection of the most relevant features for training, ensuring they adequately represent the criteria used by human vision in recognizing objects in digital imaging. Classification models were obtained with accuracies of up to 87%, which were successfully reproduced during the implementation of decision trees in the final algorithm. The results achieved even surpassed the initial accuracy rates for most disk diameters, as evident from the comparison between Tables 3 and 4.

A comprehensive system analysis was conducted by calculating the software's classification accuracy relative to technical reports, considering each disk individually. This analysis highlighted the visibility potential of each phantom disk, demonstrating that the system's behavior is consistent with human readings. The accuracy rates were lower primarily for disks near the visibility threshold, further validating the system's performance.

Regarding image quality parameters, the findings confirmed the premise previously mentioned by other authors, such as Thomas et al. [24], that the use of different CDMAM image quality parameters can potentially lead to different conclusions about image quality itself. However, the notable correlation observed between the correct observation ratio (COR) and the accuracy rate for each image suggests a convergence that can be explored in future studies.

In conclusion, this computational scheme has achieved its initial objectives of providing a more specific and less sensitive system for analyzing image quality. Unlike diagnostic schemes, the system's behavior is desirable for image quality analysis, as it reduces the chances of certifying an inadequate image as suitable for medical analysis. By serving as a second opinion and directing analysis to specialists when necessary, the system mitigates the risks associated with both false positives and false negatives.

ACKNOWLEDGEMENTS

The authors thank to Fatima F. Rodrigues Alves, PhD, for her technical assistance.

REFERENCES

1. Health Organization. Programmes and projects. Cancer Screening and early detection of cancer. Breast cancer: prevention and control.
2. Stewart BW, Wild CP. World cancer report 2014. Lyon: IARC Press; 2014. 630.
3. SheN L, Rangayyan RM, Desautels LJE. Detection and Classification of Mammographic Calcifications. Intern J Pattern Recognit Artif Intell. 1993; 7: 1403-1416.
4. Jing H, Yang Y, Wernick MN, Yarusso LM, Nishikawa RM. A comparison study of image features between FFDM and film mammogram images. Med Phys. 2012; 39: 4386-4394.
5. American Association of Physicists in Medicine. Equipment Requirements and quality control for mammography. New York, 1990. AAPM Rep. 29.
6. American College of Radiology. Mammography quality control manual. Reston, VA: ACR, 1999. ISBN: 1559031425.
7. IAEA Library Cataloguing in Publication Data. Quality assurance programme for digital mammography. Vienna: International Atomic Energy Agency, 2011.
8. Perry N, Broeders M, Wolf C DE, Törnberg S, Holland R, Von Karsa L. European Guidelines for Quality Assurance in Mammography Screening and Diagnosis. 4 ed. Luxembourg: Office for Official Publications of the European Communities, 2006.
9. Bijker KR, Thijssen MAO, Arnoldussen Th JM. Manual of CDMAM-phantom type 3.4, University Medical Centre Nijmegen, St. Radboud Department of Diagnostic Radiology, Section of Physics and Computer Science, 2000.
10. Monnin P, Marshall N W, Bosmans H, Bochud FO, Verdun FR. Image quality assessment in digital mammography: part II. NPWE as a validated alternative for contrast detail analysis. Phys Med Biol. 2011; 56: 4221-4238.
11. Figl M, Hoffmann R, Kaar M, Semturs F, Brasik N, Birkfellner W, et al. Factors for conversion between human and automatic read-outs of CDMAM images. Med Phys. 2011; 38: 5090-5093.
12. Evans AL. The evaluation of medical images. Filadélfia: Heyden & Son, 1981; 130.
13. Young KC, Johnson B, Bosmans H, Van Engen RE. Development of minimum standards for image quality and dose in digital mammography. In: 7th INTERNATIONAL WORKSHOP ON DIGITAL MAMMOGRAPHY. 2005. Proceedings of the 7th International Workshop on Digital Mammography. 2005; 149-154.
14. Sousa MAZ, Medeiros RB, Schiabel H. Investigating the visual inspection subjectivity on the contrast-detail evaluation in digital mammography images - SPIE Medical Imaging 2014: Image Perception, Observer Performance and Technology Assessment. 2014; 9037: 9037161- 9037167.

15. Karssemeijer N, Zhijsen MAO. Determination of contrast-detail curves of mammography systems by automated image analysis. In: DIGITAL MAMMOGRAPHY'96, 1996, Chicago. Proceedings of the 3rd International Workshop on Digital Mammography. Chicago: 1996; 155-160.
16. National Institute of Mental Health, "ImageJ: Image Processing and Analysis in Java,". 2023.
17. Gonzalez RC, Woods RE. Digital Image Processing, 2 ed. New Jersey: Prentice Hall, 2002.
18. Frank E, Hall M, Trigg L, Holmes G, Witten I H. Data mining in bioinformatics using WEKA. Bioinformatics. 2004; 20: 2479-2481.
19. Cheng HD, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition. 2010; 43: 299-317.
20. Barufaldi B, Lau KC, Schiabel H, Maidment ADA. "Computational assessment of mammography accreditation phantom images and correlation with human observer analysis," in SPIE Medical Imaging. 2015; 9416: 9416061-94160610.
21. Tseng DC, Huang MY. Automatic thresholding based on human visual perception. Image and Vision Computing. 1993; 11: 539-548.
22. Duda RO, Hart PE, Stork DG. Pattern classification. 2 ed. Nova Iorque: John Wiley & Sons, 2000. 654p.
23. Barufaldi B, Oliveira SS, Batista LV, Schiabel H, Almeida MSC. A computer aided detection system for microcalcifications in breast phantom images. In: INTERNATIONAL CONFERENCE ON BIO-INSPIRED SYSTEMS AND SIGNAL PROCESSING, Vilamoura. 5th International Joint Conference on Biomedical Engineering Systems and Technologies. 2012; 414-418.
24. Thomas JA, Chakraborti K, Kaczmarek R, Romanyukha A. Contrast-detail phantom scoring methodology. Med Phys. 2005; 32: 807-814. American Association of Physicists in Medicine.
25. Sousa MAZ, Medeiros RB, Schiabel H. Investigating the visual inspection subjectivity on the contrast-detail evaluation in digital mammography images. In: SPIE MEDICAL IMAGING, 2014, San Diego, CA. Proceedings of SPIE: Medical Imaging 2014.
26. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8: 283-298.