Computational assessment of mammography accreditation phantom images and correlation with human observer analysis

Bruno Barufaldi[1], Kristen C. Lau[2], Homero Schiabel[1], Andrew D. A. Maidment[2]
[1] Department of Electrical Engineering, University of Sao Paulo
[2] Department of Radiology, University of Pennsylvania

ABSTRACT

Routine performance of basic test procedures and dose measurements are essential for assuring high quality of mammograms. International guidelines recommend that breast care providers ascertain that mammography systems produce a constant high quality image, using as low a radiation dose as is reasonably achievable. The main purpose of this research is to develop a framework to monitor radiation dose and image quality in a mixed breast screening and diagnostic imaging environment using an automated tracking system. This study presents a module of this framework, consisting of a computerized system to measure the image quality of the American College of Radiology mammography accreditation phantom. The methods developed combine correlation approaches, matched filters, and data mining techniques. These methods have been used to analyze radiological images of the accreditation phantom. The classification of structures of interest is based upon reports produced by four trained readers. As previously reported, human observers demonstrate great variation in their analysis due to the subjectivity of human visual inspection. The software tool was trained with three sets of 60 phantom images in order to generate decision trees using the software WEKA (Waikato Environment for Knowledge Analysis). When tested with 240 images during the classification step, the tool correctly classified 88%, 99%, and 98%, of fibers, speck groups and masses, respectively. The variation between the computer classification and human reading was comparable to the variation between human readers. This computerized system not only automates the quality control procedure in mammography, but also decreases the subjectivity in the expert evaluation of the phantom images.

Keywords: quality assurance, breast phantom, visual perception, correlation filters, mammography accreditation phantom.

1. INTRODUCTION

Mammography image quality is a constant concern in early breast cancer screening. Global guidelines recommend systematic actions to provide adequate confidence that the optimum quality of the entire diagnostic process in mammography [1–3] is achieved. These methods not only assure the fundamental principles of radiation protection, but also propose the standardization of image techniques and other procedures involved in mammography. The regulations require, among other aspects, the periodic evaluation of mammography images of breast phantoms. These phantoms are important tools for quality assurance, evaluation of image quality, and accurate characterization of dose.

In the Mammography Quality Standards Act (MQSA), enforced by the United States Food and Drug Administration (FDA) [4], facilities performing screening mammography are required to guarantee image quality by obtaining images using the American College of Radiology (ACR) accreditation phantom. Experts must meticulously examine the structures of interest in this phantom to evaluate the accuracy of the mammography equipment.

Unfortunately, human analysis of the simulated structures, such as low-contrast disks, tumor masses, low-contrast linear details and high-contrast details can be very subjective and error-prone, and can lead to disparities in reports. To detect certain structures in such images is a complex task due to the small sizes. In certain occasions, psycho-physiological factors

Medical Imaging 2015: Image Perception, Observer Performance, and Technology Assessment,
edited by Claudia R. Mello-Thoms, Matthew A. Kupinski, Proc. of SPIE Vol. 9416, 941606

such as eyestrain, as well as signal intensity and noise levels to determine a detection threshold, and lack of technical knowledge can also negatively affect the final diagnosis [5].

Computer Aided Detection techniques, which identify phantom structures, can help experts to overcome these drawbacks. Furthermore, it can provide a second evaluation, promote the integration between medicine and technology, and improve the detection of the structures of interest [6]. Previous studies have evaluated the use of computerized systems to assess the performance of mammography equipment, including studies that correlate computational systems with the human visual system [7–9].

This study aims to develop a computerized system and to automate the procedures of quality assurance in mammography. The system detects and classifies structures of interest in the *Mammography Accreditation Phantom* (Gammex-RMI Model 156) images, and correlates this classification with the human visual system. This system is integrated with a framework that extracts metadata from these phantom images, stores classification results in a database and presents phantom's dose information, using a reporting platform. Thus, the quality assurance procedures can be automated, hence reducing subjectivity in phantom image evaluations.

2. MATERIALS & METHODS

2.1 Image acquisition

In this study, three breast tomosynthesis systems (*Selenia Dimensions*, *Hologic Inc.* Bedford, MA) were used to produce the radiographic images. The Hologic system is equipped with an amorphous selenium (a-Se) detector layer with a thickness of 250 µm and pixel pitch of 70 µm. The radiographic images were taken using the automatic mode 2D with the same settings (X-ray tube voltage, filtration, automatic exposure control etc.), typical of those used during normal mammography practices.

Exposures were performed on the ACR MAP breast phantom (*RMI-156D, Gammex-RMI*. Madison, WI), which corresponds to a compressed breast of approximately 4.0 to 4.5 cm thick and includes objects simulating sets of 5 high-contrast speck groups (with individual diameters ranging from 0.16 mm up to 0.54 mm), 6 fibers (from 0.40 mm up to 1.56 mm wide) and 5 tumor masses (with diameters ranging from 0.25 mm up to 2.00 mm). Figure 1 illustrates this phantom and the structural representation.

In total, 20 images per month were produced in two sites, using three different serial numbers of the ACR MAP phantom. It should be stressed that these sites, Hospital of University of Pennsylvania (Philadelphia, PA) and Einstein Medical Center (Philadelphia, PA), consist of a full range of diagnostic, screening and radiological services, providing a large number of diagnostic studies in mammography.

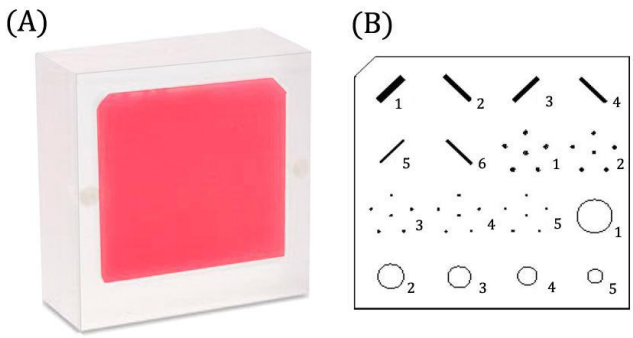

Figure 1. (A) ACR MAP phantom and (B) the location of the test objects in the red wax insert.

2.2 Human readings

As for the evaluation of the images, a team from the Hospital of University of Pennsylvania, composed of specialists in medical physics, was in charge of inspecting the phantom images and scoring each structure. They analyzed the visibility of the fibers, masses, and speck groups, according to the methods outlined in the Mammography Quality Control Manual [3]. This study used four experts for a more reliable analysis.

The image was rated in terms of "quality points" according to the requirements of the ACR. The specialist should count the number of visible objects from the largest structure of a given type (i.e., fiber, speck group, or mass) downward until a score of zero is reached, then stop counting for that object type. One fiber or mass is classified as visible if the full size of this structure is detected, and the location and orientation are accurate.

The speck group classification is slightly different. Experts should start from the largest to the smallest speck group and score one if four or more of the six specks in the cluster are visible in the proper locations. Magnification is allowed for viewing this structure.

Readings of incomplete structures (half-scores) or artifacts are not covered in this study.

2.3 Templates

The computerized system localizes the structures by defining different search regions of interest (ROIs). The ROIs depend on the image size and an initial point, which is previously defined by the user. In general, users define the center of the first mass as the initial point for searching, since this structure is easily detectable in the image.

Since the structures' locations slightly vary from one phantom to another, the user can also calibrate the system. Calibration is recommended only for exposures from different phantoms, i.e. for phantoms with distinct serial numbers. It is a simple procedure; the user defines the initial point for searching and selects the center of the largest fiber, speck group and mass.

After determining the initial point, matched filters were used to localize the structures of interest. The masses and specks were detected by using circular filters that consist of two parts, which match with these objects. In order to localize the fibers, filters with two external and one internal straight line are applied. Since the fibers are arranged within the phantom at various angles, to achieve better adjustment and greater correlation, the filters need to be applied using several rotations within the range of 37º and 53º with respect to the horizontal axis.

Figure 2 illustrates an example of the procedure specifically applied to typical images of the previously described phantoms. 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 (C) demonstrates the situation when the inner region matches the inner structure and the outer region matches the background. The filter's radii (r1 and r2) changes according to the size of each structure (inner or outer region).

The correlation method was implemented along with the matched filters. This method is described by Gonzales and Woods [10]:

$$c(x,y) = \sum_s \sum_t f(s,t) * w(x+s, y+t), x = 0,1,2,\ldots,M\text{-}1 \text{ and } y = 0,1,2,\ldots,N\text{-}1$$

Consider an image *f* of size *M x N* and a mask *w* (filter) of size *J x K*, with $J \leq M$ and $K \leq N$. The correlation is calculated where the mask *w* overlaps the image *f* during the convolution procedure. *w* moves over the image *f*, resulting in the function *c(x, y)*. The maximum value of the function *c(x,y)* indicates the position of the best match between *w* and *f*.

2.4 Learning Models

The computerized system was developed using the Java programming language with *ImageJ* [11], an open source program that is focused on the development of image processing applications and analysis. The software interface and the algorithms

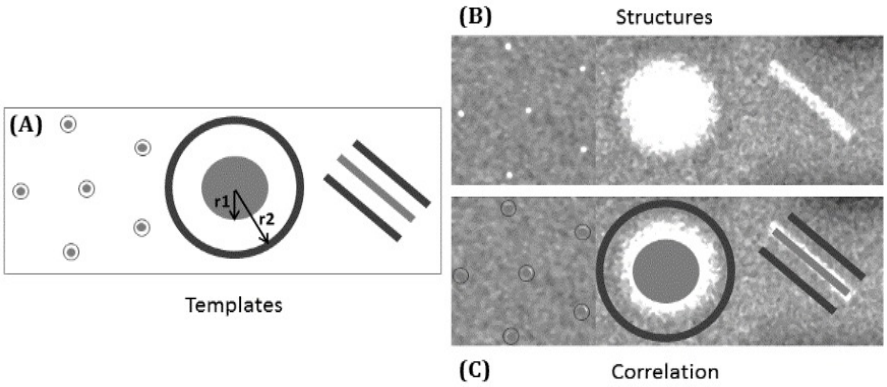

Figure 2. (A) Filters used for the detection and the classification of (B) structures of the phantom image. (C) Simulation of results, after locating the structure of interest.

developed use the classes and methods from the *ImageJ* library (ij.jar). In addition, a data-mining tool known as WEKA [12] was also used along with the software developed to generate decision trees through algorithms of predictive analytics.

The visibility classification of fibers, masses and speck groups in the ACR MAP images in our computational methodology is based on the J48 algorithm from the WEKA tool. The J48 algorithm was chosen not only for its simplicity and performance but also for its accuracy and effectiveness in previous studies [13]. This classifier generates a decision tree model for each type of structure, where pre-selected image features are used for the training stage. The average of pixel images, standard deviation, mode, average of structure pixels, average of background pixels, difference of structure and background averages, and Weber Ratio [10] were pre-selected for each learning model.

Three sets of 60 images each were used for training purposes. Each set included five images that were produced over each of 12 months (January 4 to December 20, 2011). Three different learning models for a given object type were generated in each image group. The procedure of leave-one-out was executed from the input features to generate the learning models in this step. Not all features used in the training are employed in decision trees since some image characteristics are not considered relevant by the classifier and hence are automatically discarded. Using the attribute selector provided by WEKA, the model's performance will be improved and substantially reduce the training effort.

Each model obtained in the training stage was implemented and incorporated into the software developed. The classification procedure utilizes decision trees based on the models of each structure. A set of 240 images were processed for classification purposes. Statistical features such as system accuracy and efficiency can be determined from this classification process by the comparison between the technical expert reports and the software results.

2.5 Results analysis

By comparing the technical expert reports and the software results, a discrepancy was noticed in the reports provided by the four observers due to human visual subjectivity. To minimize the variation of the results obtained by the readers, the mode was used to discard outliers. The mode selects the most frequent score among the readers. Noteworthy, there is always at least two identical scores in the structure classifications.

The software and the readers' results are stored in a relational database by a tracking system, which selects the information provided by the DICOM header and external sources by date, room, technologist, procedure, view, etc. [14] This tracking system is composed of a Custom DICOM Service Class Provider (SCP), data-cleaning software, and a client-side application. The intermediate software cleans the information extracted from DICOM headers and from external sources. The DICOM SCP then maps the metadata into a MSSQL database using an object relational mapper.

A SharePoint application graphically exhibits the information stored in this database by date using the software and the report results. Thus, it is possible to evaluate mammography system performance to produce part of the annual report from the equipment.

This graphical application also allowed a phantom dose analysis to verify the consistency of information reported by the manufacturer, and to analyze the correlation between dose and number of structures detected by the software. Since the software results and the image metadata were stored in a MSSQL database (Figure 3), users might define which view they want in order to exhibit these data [14]. Thus, the entrance skin kerma (ESAK), the DICOM dose and the number of structures detected were selected from the period that the images were acquired.

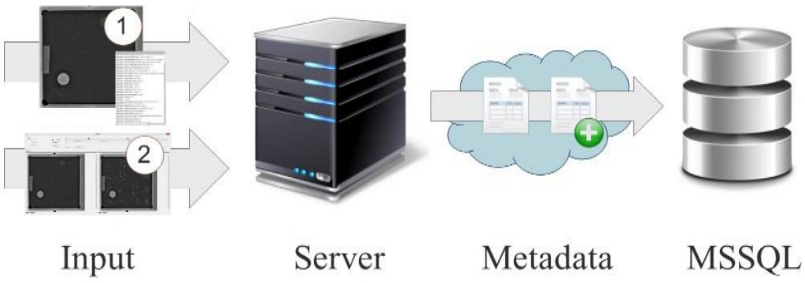

Figure 3. Flow chart that shows the DICOM SCP process from the tracking system. (1) represents the input image and its DICOM header, and (2) input from external sources.

3. RESULTS & DISCUSSION

3.1 Overview

The results refer to the classification of all structures of interest in the phantom images. These results include the training procedure and the correlation between the expert's reports and the computational analysis of the structures. Additionally, the results contain the effect of the correlation filters and results of both evaluations, software and human inspection. Finally, it shows the relationship between the dose reported by the manufacturer and the classification of phantom structures after correction, using the mode for scoring.

3.2 Software training

The software classification was based on learning models generated from the features selected by the tool WEKA (*Attribute Selection*). Table 1 displays the selected features present in the learning models, in which p, p_s and p_b indicate the pixel value regarding the position (i,j) in the image, structure of interest and background around the structure of interest respectively. Table 2 shows the best models generated for each structure type, after the training stage.

Table 1. Selected features present in the learning models for each structure of interest.

Training	
Selected Features	**Equation**
Average of the structure pixels	$\mu_s = \sum_{i=1}^{w}\sum_{j=1}^{h}\frac{p_s(i,j)}{(w*h)}$
Average of the background pixels	$\mu_b = \sum_{i=1}^{w}\sum_{j=1}^{h}\frac{p_b(i,j)}{(w*h)}$
Weber ratio [17]	$W = \frac{\mu_s - \mu_b}{\mu_e}$

Table 2. Best results for the training of fibers, speck groups and masses in the ACR MAP images.

Training			
	Fibers	**Specks**	**Masses**
Correctly classified	92%	96%	97%
Sensitivity	0.94	0.96	0.99
Specificity	0.83	0.95	0.89
ROC (A_z)	0.91	0.96	0.99
Total of structures	360	1,800	300

After training the software against the series of datasets, WEKA tool provided three different decision trees and statistical results for each structure of interest. The correctly classified rates, sensitivity, specificity and A_z of each decision tree were analyzed to select which model should be implemented for the software testing.

It should be stressed that the software was trained based on reports provided by a team of experienced observers to inspect this type of image. These observers' reports are used not only for training the system but also for limiting the structures detected by software reading. Otherwise, the system would detect virtually all structures in the phantom images, even those imperceptible by human observers.

3.3 Software testing

Figure 4 illustrates the automatic delimitation of structures in the ACR MAP images after being processed by the system. Note that the structures of interest are accurately localized by the correlation filters. After this detection, each structure of interest was classified as visible or not visible by our computational methods.

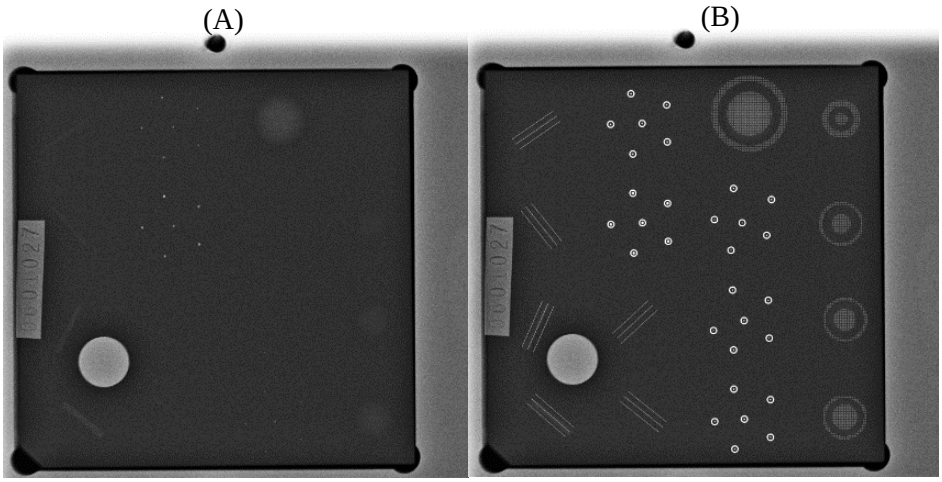

Figure 4. ACR MAP phantom image (A) before and (B) after processing, with all structures marked.

3.3.1 All scores

The best models obtained in the training stage were implemented and incorporated into the software developed. Table 3 presents the correctly and incorrectly classified rates, sensitivity, specificity, efficiency, positive prediction and negative prediction, Kappa values, Matthew's coefficients and ROC (A_z), for the classification of each structure of interest from the entire dataset (240 images). The accuracy rate of each structure of interest reached to at least 81%, a reliable behavior for

Table 3. Average results for the classification of fibers, speck groups and masses in the ACR MAP images.

	Testing						
	Fibers	**Specks**	**Masses**		**Fibers**	**Specks**	**Masses**
Correctly Classified	81%	90%	89%	Negative Prediction	0.94	0.97	0.98
Incorrectly Classified	19%	10%	11%	Sensitivity	0.87	0.78	0.79
Kappa statistic	0.75	0.80	0.78	Specificity	0.37	0.88	0.86
Matthews Coefficient	0.51	0.75	0.74	Efficiency	0.62	0.83	0.83
Positive Prediction	0.80	0.88	0.87	ROC (A_z)	0.83	0.94	0.92

classification of ACR MAP images. It is noteworthy that these results were validated with four different reports produced by human observers.

Despite of the size of the speck groups, the classification of these structures represented a high success rate of the system (90%). These results are due to the specks having the highest contrast of all objects.

Although the system achieved high classification rates for the structures of interest, the results were worse than those presented in the training step. Thus, a further investigation was conducted to verify the reasons why the system behaved differently from the learning models.

Most of the low rates of sensitivity and specificity are due to variation in readers' analysis for the classification of structures, given that the learning models are generated based on observer's reports. Figure 5 shows the observer's subjectivity in the analysis of the structures of interest in the ACR MAP phantom.

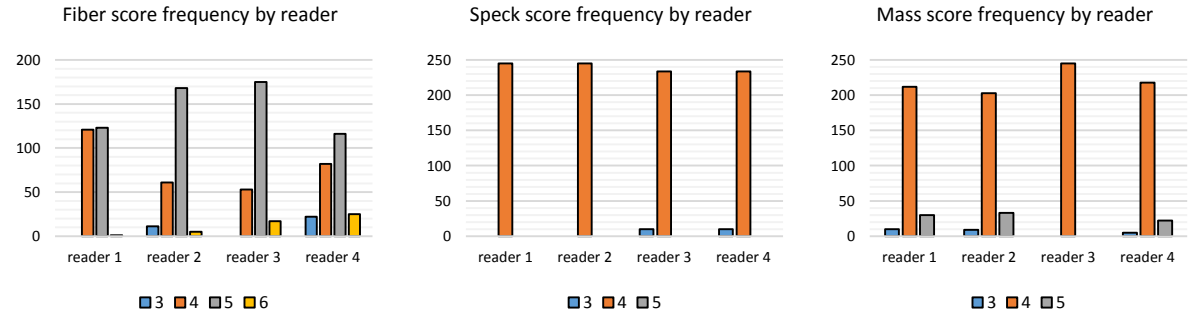

Figure 5. Score frequency of each structure by four different human observers in the breast phantom images.

After inspecting the phantom images by four readers, the Kappa values achieved were 0.56, 0.99 and 0.79 for the structures that simulate fibers, speck groups and masses, respectively. Note that the most stable results were in the structures that represents speck groups. This fact influenced the best results of the training and, consequently, the software testing.

3.3.2 Mode for scoring

The most frequent value (mode) eliminates outliers from the results obtained by the readers' reports. Figure 6 shows the average rating of each structure of interest as a function of month, according to both methods. The average variation

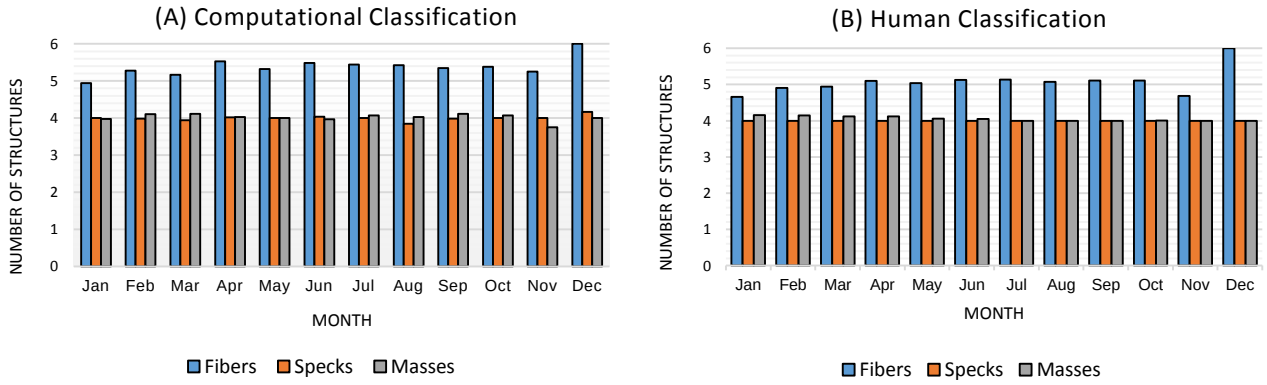

Figure 6. Mean of classification in function of the months of 2011. (A) Computational assessment of the structures of interest in the phantom images, and (B) same classification by visual inspection of the radiologists, using the mode method.

between both computer and human analysis for the classification of structures that simulate fibers, speck groups and masses was 6%, 1% and 1%, respectively.

Table 4 presents the same statistical analysis previously shown, after the use of the mode for scoring in the computational assessment. Note that there was an improvement of at least 7% in the correctly classified structures.

Table 4. Results for the classification of fibers, speck groups and masses in the ACR MAP images.

Testing							
	Fibers	**Specks**	**Masses**		**Fibers**	**Specks**	**Masses**
Correctly Classified	88%	99%	98%	Negative Prediction	0.87	0.98	0.96
Incorrectly Classified	12%	1%	2%	Sensitivity	0.98	1.00	0.99
Kappa statistic	0.87	0.99	0.98	Specificity	0.54	0.97	0.95
Matthews Coefficient	0.63	0.97	0.94	Efficiency	0.76	0.98	0.97
Positive Prediction	0.89	0.99	0.99	ROC (A_z)	0.91	0.99	0.98

After discarding the outliers, there was a reduction in the number of false positives and negatives, resulting in a substantial improvement in the sensitivity and specificity rates for all structures of interest. Moreover, all Kappa values of the system were equal to or greater than that the values between the readers.

3.4 Phantom dose analysis

The classification method was performed in different phantoms, making it compatible for processing images from phantoms with structures with various shapes and sizes [13]. Because all information acquired by the DICOM header is stored in a database, different features, such as the dose reported by the manufacturer, can be assessed in the phantom image.

Figure 7 shows the entrance skin kerma (ESAK) and the dose reported by the DICOM header for the information stored in the database from the tracking system. The average of the ESAK and DICOM dose were 3.14±0.12 mGy and 1.02±0.10 mGy for the mammography system used.

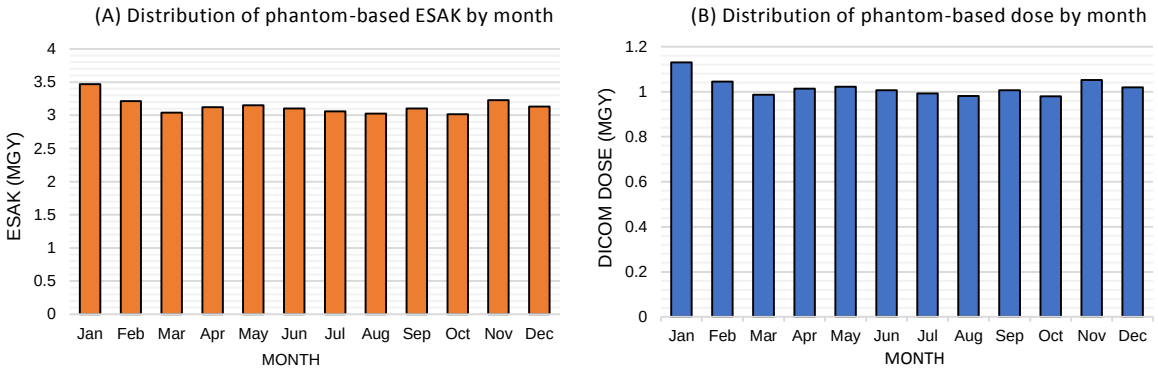

Figure 7. Histograms of (A) ESAK (mGy) and (B) DICOM DOSE (mGy) in function of the months of 2011.

There were no major variations in the doses reported by the manufacturer. Figure 8 shows the number of structures detected as a function of dose in both methods. Using the automatic mode and the exposure factors previously mentioned are not sufficient to find a correlation between the DICOM dose and the number of detectable structures.

It should be stressed that the manufacturer dose is based on phantoms measurements that simulate breast glandularity. The dose for patients should be more accurately calculated to evaluate the actual dose from mammography systems. While the ACR mammography accreditation provides a standard for the dose of a phantom, there is no such standard for human exposures [14]. This research is tied to an ongoing study being performed at the University of Pennsylvania to assess dose measurements, using breast density factors from patients.

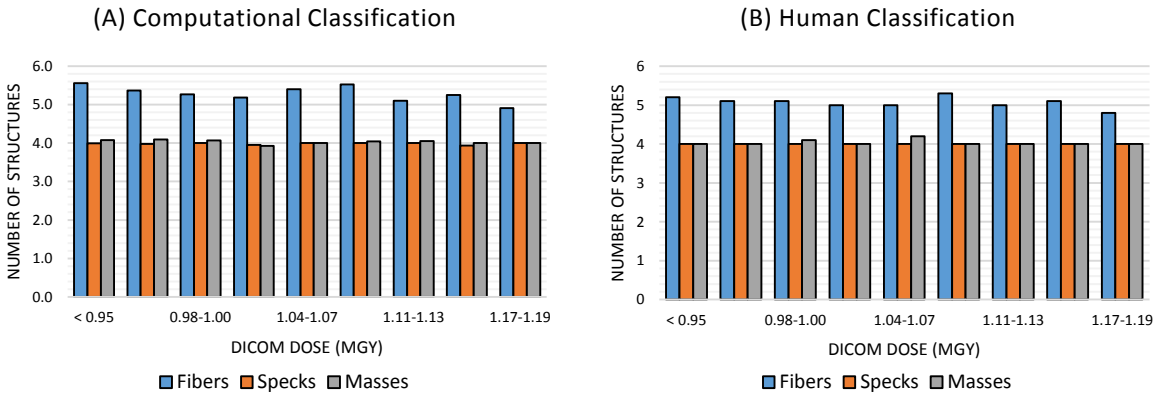

Figure 8. Mean of classification in function of DICOM dose (mGy). (A) Computational assessment of the structures of interest in the phantom images, and (B) same classification by visual inspection of the human observers, using the mode for scoring.

5. CONCLUSION

This computerized system is used as a tool for automating the human visual inspection of phantom ACR MAP images. The system can be used to reduce subjectivity and the number of persons-hours for obtaining a second opinion. However, even with the accuracy rates presented, the reports generated by the proposed system should always be analyzed by at least one human expert.

Despite the small sizes of simulated speck groups, the system classification performance was the highest, with an accuracy rate of 99%. However, software was trained based on the analysis provided by four readers who reported the images with similar scores. Both, training and testing step are inclined to classify these structures with the same score, although there are a few software misclassifications (1%).

The structures that simulate tumor masses also obtained high rates of correct classifications, with 98% of accuracy. Moreover, all structures obtained κ and A_z values greater than 0.87 and 0.91, which are better than the human results and indicate a high degree of reliability for the system.

Therefore, by making use of this computerized system, the mammography quality control process can be automated, hence reducing the subjectivity in the phantom image evaluations. Thus, establishments that offer mammography services can perform their own quality control efficiently and with appropriate reliability.

The developed tracking system is a tool that allows for widespread utilization in mammography systems of different modalities, regardless of manufacturer or model. Phantom dose analysis is just an example of additional information that this system can provide. Various data views support different user roles (e.g., technical or medical supervisor physicist).

Once we have a better understanding of how the quality control is being performed in breast imaging, we will integrate this tracking system with different frameworks to assess image quality.

ACKNOWLEDGMENTS

The project described is supported by CAPES and NIH, process number 99999.014175/2013-04 (Methods for Quality Control in Digital Mammography), U54-CA163313-04 (Penn Center for Innovation in Personalized Breast Screening), ACRIN PA 4006 (Comparison of Full-Field Digital Mammography with Digital Breast Tomosynthesis Image Acquisition).

The authors also wish to thank the Hospital of University of Pennsylvania (Philadelphia, PA) and the Albert Einstein Healthcare Center (Philadelphia, PA) for providing the mammography systems and the images used in this study.

REFERENCES

[1] N. Perry., Broeders, M., Wolf, C. de., Törnberg, S., Holland, R.., Karsa, L. von., European guidelines for quality assurance in breast cancer screening and diagnosis, 4th ed., 1–160, Luxembourg, Belgium (2013).

[2] IAEA., "Quality Assurance Programme for Digital Mammography," Vienna, Austria, 105–108 (2011).

[3] ACR., Mammography Quality Control: Radiologist's manual, radiologist technologist's manual, medical physicist's manual., 4th ed., Reston, VA (1999).

[4] FDA., "Mammography Quality Standards Act Regulations," 1–37 (1998).

[5] Byng, J. W., Yaffe, M. J., Lockwood, G. A., Little, L. E., Tritchler, D. L.., Boyd, N. F., "Automated Analysis of Mammographic Densities and Breast Carcinoma Risk," Cancer **80**(1), 66–74 (1997).

[6] Freer, T. W.., Ulissey, M. J., "Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center," Radiology(220), 781–786 (2001).

[7] Lanconelli, N., Rivetti, S., Golinelli, P., Serafini, M., Bertolini, M.., Borasi, G., "Comparison of Human observers and CDCOM software reading for CDMAM images," Proc. SPIE **6515**, Y. Jiang and B. Sahiner, Eds., 65150E – 1 (2007).

[8] Prieto, G., Chevalier, M.., Guibelalde, E., "Automatic Scoring of CDMAM Using a Model of the Recognition Threshold of the Human Visual System: R*," Image Process. (ICIP), 2009 16th IEEE Int. Conf. Image Process. **5**, 2489–2492 (2009).

[9] Perez-Ponce, H., Daul, C., Wolf, D.., Noel, A., "Computation of realistic virtual phantom images for an objective lesion detectability assessment in digital mammography.," Med. Eng. Phys. **33**(10), 1276–1286, Institute of Physics and Engineering in Medicine (2011).

[10] Gonzalez, R. C., Woods, R. E.., Hall, P., Digital Image Processing, 2nd ed., A. Dworkin, Ed., 1–190, Prentice-Hall, Inc., New Jersey (2002).

[11] National Institute of Mental Health., "ImageJ: Image Processing and Analysis in Java," <http://imagej.nih.gov/ij/index.html> (20 May 2001).

[12] Waikato Environment for Knowledge Analysis., "Weka 3: Data Mining Software in Java," <http://www.cs.waikato.ac.nz/ml/weka/> (20 May 2001).

[13] Barufaldi, B., Oliveira, S. S. De., Batista, L. V.., Schiabel, H., "A Computer Aided Detection System For Microcalcifications in Breast Phantom Images," Int. Conf. Bio-inspired Syst. Signal Process., 1–5, Valemoura, Portugal. (2011).

[14] Barufaldi, B., Maidment, A. D. A., Cook, T., Synnestvedt, M., Conant, E.., Schnall, M., "A Radiation Dose Reporting System for Mammography and Digital Breast Tomosynthesis," Radiol. Soc. North Am., Chicago, USA. (2014).