Statistical analysis to assess automated Level of Suspicion scoring methods in breast ultrasound

Michael Galperin[*a]
[a]Almen Laboratories, Inc., 2105 Miller Ave., Escondido, CA 92025

Abstract

A well-defined rule-based system has been developed for scoring 0-5 the Level of Suspicion (LOS) based on qualitative lexicon describing the ultrasound appearance of breast lesion. The purposes of the research are to asses and select one of the automated LOS scoring quantitative methods developed during preliminary studies in benign biopsies reduction. The study has used Computer Aided Imaging System (CAIS) to improve the uniformity and accuracy of applying the LOS scheme by automatically detecting, analyzing and comparing breast masses. The overall goal is to reduce biopsies on the masses with lower levels of suspicion, rather that increasing the accuracy of diagnosis of cancers (will require biopsy anyway). On complex cysts and fibroadenoma cases experienced radiologists were up to 50% less certain in true negatives than CAIS. Full correlation analysis was applied to determine which of the proposed LOS quantification methods serves CAIS accuracy the best. This paper presents current results of applying statistical analysis for automated LOS scoring quantification for breast masses with known biopsy results. It was found that First Order Ranking method yielded most the accurate results. The CAIS system (*Image Companion*®, *Data Companion*® software) is developed by Almen Laboratories and was used to achieve the results.

Keywords: breast ultrasound, cancer, image analysis, segmentation, level of suspicion scoring, diagnostic imaging, ROC, computer-aided, classification, statistical similarity

1. Introduction

The National Cancer Institute estimates that approximately 700,000 women undergo breast biopsies (surgical or needle) in the U.S. each year. Approximately 80% of tumors biopsied are benign, 20% are malignant. Surgical biopsies--the most common--cost between $2,500 and $5,000 while needle biopsies cost from $750 to $1,000. Patients experience both physical and emotional effects when undergoing biopsy procedures and internal scarring may be problematic since it complicates interpretation of future mammograms. Until recently[1], ultrasound has only been used in distinguishing cystic from solid breast masses and to guide needle biopsies. A number of positive studies in Europe, Asia and the U.S. indicate that high-quality ultrasound can provide radiologists with a high degree of confidence in differentiating many benign from malignant or suspicious lesions detected by mammography.[1-3] Results suggest that ultrasound could help reduce the number of unnecessary biopsies by 40% with a cost savings of as much as $1 billion per year in the U.S.

As reported earlier this work has led[1-3] to the development of a well-defined system for scoring the Level of Suspicion (LOS) based on parameters describing the ultrasound appearance of the breast lesion[1,2]. Breast ultrasound protocol involves consideration of the following characteristics of the lesion: margins, shape, echogenicity, echo texture, orientation and posterior acoustic attenuation pattern. When a solid or solid-appearing mass is seen, the margins or degree of irregularity are evaluated. Benign masses usually have smooth margins while malignant tumors may have spiculations or finger-like extensions. The outline shape of the mass is examined to determine if it is ovoid, spherical, lobulated or irregular. Benign masses usually are spherical or egg shaped with three or less lobulations. Malignant masses tend to be irregular with less distinct boundaries.

The specific guidelines for differentiation of breast lesions are shown in Table 1, while the LOS score is assigned based on the number of benign and malignant criteria as shown in Table 2.

A number of promising efforts to improve the specificity of breast lesion classification using ultrasound may be grouped in two categories: 1) analysis of features in the display (image processing),[7-9] and 2) analysis of the ultrasound signal properties (tissue characterization).[10-15] Much of this work confirms that it is difficult to precisely classify masses because there is overlap in the acoustic properties of many solid benign and malignant lesions. Computer-aided diagnosis with artificial neural networks (ANN), a form of regression analysis, attempts to aid the radiologist in locating suspicious regions that might otherwise be missed[16,17]. We concentrated our efforts on the complementary problem of improving confidence in benign findings.

Table 1: Guidelines

Criteria Associated with Benign Lesions	Criteria Associated with Malignant Lesions
Spherical/ovoid/lobulated	Irregular shape
Linear margin	Poorly defined margin
Homogeneous texture	Central shadowing
Isoechoic/anechoic	Distorted architecture
Edge shadow	Calcifications
Parallel to the skin	Skin thickening
Distal enhancement	
Dilated duct/mobile	

Table 2: Scoring System

LOS	Diagnosis	Number of Criteria
5	Malignant	5 malignant criteria
4	Probably malignant	3-4 malignant criteria
3	Indeterminate criteria	1-2 malignant criteria
2	Probably benign	0 malignant criteria
1	Benign	0 malignant criteria & all benign criteria

2. Methods

Images of breast masses for 146 cases, with a wide range of known biopsy results, were selected retrospectively by an experienced radiologist. Images were stripped of all patient or diagnostic history information and then scored for LOS values by three independent radiologists. Segmentation of the suspicious mass was performed in both semi- automatic and automatic modes following pre-processing. These results were compared to manual segmentation performed by the radiologist (Fig. 1) and the accuracy of the segmentation was evaluated[1-3].

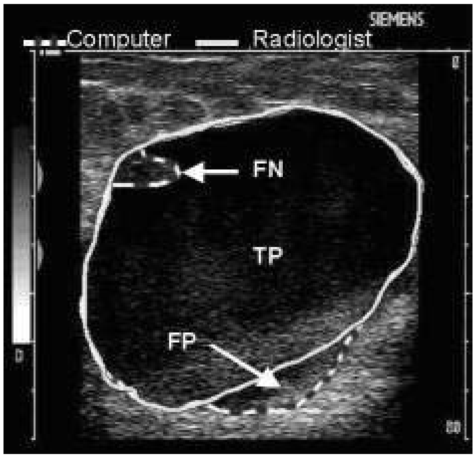


Figure 1: Large cyst with layered debris and a solid component. This illustrates a method to compute accuracy, sensitivity and specificity of computer segementation. The radiologists by consensus will manually outline the lesion ("truth," solid line).

The performance of the semi-automatic and automatic segmentation algorithms was assessed using a method established by a number of investigators to compute accuracy, sensitivity and specificity. As reported earlier[1-3,4-6], in this method, each pixel of an image is assigned to one of four categories: True Positive, False Positive, True Negative and False Negative. The pixels of the image that are assigned to the lesion by both the radiologist and the computer algorithm are considered True Positive (TP), pixels which are considered by both to be outside the lesion are labeled True Negative (TN), pixels which are considered inside the lesion by the computer algorithm but not be the radiologist are labeled False Positive (FP), and pixels assigned to a region outside the lesion by the algorithm but inside by the radiologist are labeled False Negative (FN). The number of pixels, N, in each category is summed and normalized to avoid any impact of relative size of the object. Accuracy, sensitivity and specificity will be measured for each image as defined by:

$$Accuracy\,(A_Z) = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \tag{1}$$

$$Sensitivity = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{2}$$

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

The averages of each of these terms may be calculated for all images and the results may be used as figures of merit for different segmentation algorithms. Uncertainties of these measurements were estimated from the standard deviations of the averages. Clearly, higher accuracy, sensitivity and specificity are preferred but the threshold of “acceptability” requires consideration of the classification performance as well. The examining radiologists by consensus manually outlined the lesion (the "truth"). If, for example, the computer algorithm produces the segmentation line, the number of pixels which deviate from this region will be scored as "False Positive", "False Negative", or "True Positive" (Fig. 1). Partial results of Sensitivity, Specifty and Predictability analyses are presented in Table 3.

Table 3: Partial Results of Accuracy Assessment

Observer	Sensitivity	Specificity	PPV[1]	NPV
Radiologist A	87	76	36	97
Radiologist B	93	66	30	99
Radiologist C	90	74	34	97
CAIS	**80**	**98**	**80**	**97**

The software calculated a large number of classes of parameters of segmented lesions including the classes of shape, density and texture of the mass. The selected set of used parameters was reported before[1-3] and here represented in Table 4. Standard correlation analysis was used to select a subset of parameters with less interdependency.

Table 4: Default image parameters for lesion similarity calculation and further classification

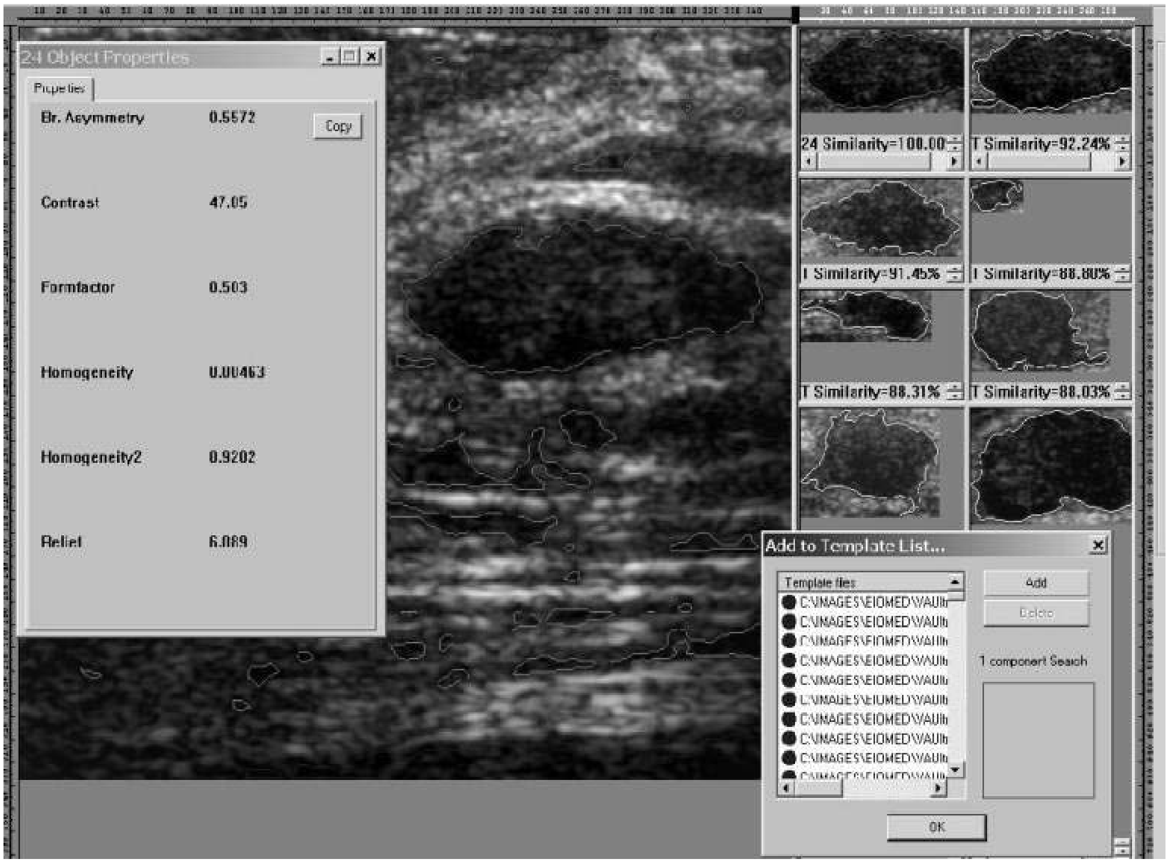
Image Criteria	**Sample of Associated Parameters**
Spherical/ovoid vs. irregular shape	Formfactor, Equivalent circular diameter/Form factor Perimeter/Area, Perimeter/Equivalent circular diameter, Aspect ratio
Linear margin vs. poorly defined margin	Edge gradient
Homogeneous texture vs. internal echoes Isoechoic/anechoic vs. echoic Calcifications	Homogeneity (multiple texture parameters) Relief, Contrast, Optical density, Integrated density Scatterer density, scatterer size, 2nd, 3rd, 4th moments of inertia
Edge shadowing vs. Central shadowing Distal enhancement	Density measures of a Distal ROI defined by X- and Y-Ferret coordinates
Parallel to skin vs. irregular	X-Ferret/Y-Ferret, Aspect ratio, Relative angle

[1] PPV – positive predictive value; NPV – negative predictive value.

The selected quantified lesion descriptions represent an N-dimensional (Tables 1-3) vector **P** that may be used to calculate the Relative Similarity of one lesion to another (*4*)

$$\left(\sum_{k=1}^{L} \left(p_{k}^{t} - P_{k}^{t} \right)^{s} * \omega_{k} \right)^{1/s} , \tag{4}$$

where ω_k is the statistical weight resultant from multi-factorial analysis; $\mathbf{P}_{\mathrm{it}}$ – is the index of this “template” object (compared to the other lesions); $\mathbf{P}_{k\,\_}$ is the feature vector of the current lesion and (k=1,…L) where L is the number of lesions.



Br. Asymmetry	0.5572
Contrast	47.05
Formfactor	0.503
Homogeneity	0.00463
Homogeneity2	0.9202
Relief	6.089

24 Similarity=100.00

T Similarity=92.24%

I Similarity=91.45%

I Similarity=88.80%

T Similarity=88.31%

T Similarity=88.03%

C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh
C:\IMAGES\EIOMED\WAUh

Figure 2: Complex cyst (benign) is compared to other images in the template database. The "unknown" mass in the upper left portion of the screen is a dark, relatively echo-free consistent with fluid-filled cyst but with irregular indistinct margins more consistent with a solid mass that might have a higher suspicion for cancer. Software automatically locates the mass contour. Measurements are made of the mass and its "relative similarity" is compared to a digital template database with known findings. Cases most “similar” to this suspicious mass are automatically retrieved and displayed in the thumbnail images on the right listed in rank order of this value (light contours in the right half of the screenshot). In this case, all of the “similar” masses were proven to be benign. The software allows the user to select additional digital templates with known findings and include them into calculations of the similarity.

To estimate statistical weights and obtain correlation results standard correlation analysis methods were used (the statistical methods were implemented in the proprietary statistical software system *Data Companion*® developed by Almen Laboratories, Inc.).

The weight assigned to a given parameter during this comparison process could be manually set by the user, or preferably set using a statistical method. This is especially useful when there is a structured set of rules for object characteristics. This data can be analyzed to determine how strongly different parameters of the parameter set values correlate with the presence or absence of the specific trait. The weight used for a given parameter in the comparison process may thus be derived from the values of the parameter vectors associated with the detected objects in the image database. In using this method a system is represented as a totality of factors. The statistical tools are correlation, regression, and multi-factor analyses, where the coefficients of pair-wise and multiple correlations are computed and both a linear and non-linear regression may be obtained. The data for a specific model experiment is represented as a matrix whose columns stand for factors describing the system and the rows for the experiments (values of these factors). This method of multifactor analysis was successfully developed, applied and verified in other biomedical fields[18].

The factor Y, for which the regression is obtained, is referred to as the system response. (Responses are integral indicators but theoretically, any factor can be a response. All the factors describing the system can be successively analyzed. In breast cancer, Y could be a biopsy result, lesion class or any other clinical indicator that is impacted by analyzed factors). The regression and covariance help to “redistribute” the multiple determination coefficient among the factors; in other words the “impact” of every factor to response variations is determined. The specific impact indicator of the factor is the fraction to which a response depending on a totality of factors in the model changes due to this factor. This specific impact indicator may then be used as the appropriate weight to assign to that factor (i.e., parameter set associated with the objects). The impact of a specific factor is described by a specific impact indicator, which is computed by the following algorithm: $\gamma_j = \alpha * [\, b_j * c_{0j} \,]$, j=1,2,...,k where γ is the specific impact indicator of the j-th factor; k is the number of factors studied simultaneously; b_j is the j-th multiple regression coefficient; c_{0j} – covariance coefficient and α - is the fraction of multiple determination related to the impact of the factor.

Three methods of **P** to LOS automated transformation were implemented in CAIS (*Methods 1-3*).

Algorithm of Method 1. First order ranking method. The software calculates and retrieves the 7 closest matches to the unknown lesion that was segmented and quantified. Then the rank is assigned according to a rule: if 4 out of 7 closest template lesions are benign then the rank of the score will be 1 (benign), otherwise the score will be 5 (malignancy). This is the method employed in the preliminary study of 146 cases.

Algorithm of Method 2. Simple Averaging Ranking Scoring. Continuum similarity values for the closest 7 templates with known findings are substituted by their binary ranks 1 (benign) or 5(malignant). Then the assigned score is an average of the 7 ranks.

Algorithm of Method 3. Scoring with the penalty function (this method was developed by Almen Labs and we are not aware of any similar methods published). The method uses only the 5 closest templates of the lesions with known findings. The values of calculated similarities between each template with known finding and the unknown lesion is substituted with the values that are calculated as follows:

For Templates of Malignant lesions:

Malignant Score – Penalty * Relative Similarity;

For Templates of Benign lesions:

Benign Score + Penalty * Relative Similarity.

A sample set of sequential cases was selected and correlation analyses (classic and rank) were performed to assess the most accurate method for CAIS LOS automated scoring. The original data is presented in Table 5 (partial).

Table 5: Example of original data used for correlation analysis.

Patient #	Biopsy Find	Radiol1	Radiol2	Radiol3	Method 3	Method 2	Method 1
1	1	1	2	1	1	2.04	1
2	1	2	4	2	1	2.85	2.3
3	1	2	2	1	1	2.76	1
4	1	2	1	1	1	2.14	1
5	1	4	5	4	1	1.93	1
6	1	4	3	4	1	2.79	1
7	1	2	1	1	1	1.85	1
8	1	4	5	4	1	2.62	1
9	1	2	1	1	1	1.64	1
10	1	1	1	1	1	2.44	1
11	1	5	5	5	1	2.50	1
12	1	1	1	1	1	2.13	1
13	1	1	1	1	1	2.33	1.8
14	1	5	5	5	2.5	2.93	3
15	5	5	5	5	5	3.26	3
16	1	1	1	1	1	1.82	1
17	1	1	3	1	1	2.61	1
18	1	1	1	1	1	1.71	1
19	1	1	1	1	1	1.96	1

3. Results

The rank order of the entire collection of masses was compared to the LOS score determined by the radiologist. The automated methods implemented in CAIS were applied and a data matrix was formed and processed with proprietary statistical software Data Companion®. Correlation analyses revealed that *Method 1* of those implemented was the most accurate and successful with correlation to biopsy findings above 0.80 (*Table 6, subset; Table 7, subset*). *Methods 2* and *3* yielded results closer to the radiologists scoring.

Table 6: Standard correlation analysis for three radiologists and three CAIS methods of LOS Scoring[2]

	Biopsy Find	Radiol1	Radiol2	Radiol3	Method 3	Method 2	Method 1
Biopsy Find	1.00	0.41	0.34	0.41	**0.48**	**0.60**	**0.82**
Radiol1	0.41	1.00	0.86	0.97	0.56	0.47	0.50
Radiol2	0.34	0.86	1.00	0.90	0.65	0.51	0.52
Radiol3	0.41	0.97	0.90	1.00	0.59	0.51	0.53
Method 3	**0.48**	0.56	0.65	0.59	1.00	0.66	0.71
Method 2	**0.60**	0.47	0.51	0.51	0.66	1.00	0.81
Method 1	**0.82**	0.50	0.52	0.53	0.71	0.81	1.00

[2] Total of 146 patients and 177 digital templates with biopsy verified findings were used in obtaining results of Table 6. The estimations were obtained with statistical credibility at 0.95 by Student.

It was interesting to note that Spearman's rank correlation yielded less conclusive results on the selected subset of data. While the trends were similar to standard correlations, *Method 2* revealed slightly higher correlation to biopsy findings. That fact can be explained by the method of deriving the score in *Method 2* that already includes partial ranking procedure.

Table 7: Spearman's rank correlation analysis for three radiologists and three CAIS methods of LOS scoring

	Biopsy Find	**Radiol1**	**Radiol2**	**Radiol3**	**Method 3**	**Method 2**	**Method 1**
Biopsy Find	1.00	0.36	0.32	0.40	**0.39**	**0.51**	**0.47**
Radiol1	0.36	1.00	0.77	0.88	0.52	0.34	0.48
Radiol2	0.32	0.77	1.00	0.89	0.67	0.38	0.52
Radiol3	0.40	0.88	0.89	1.00	0.62	0.48	0.63
Method 3	**0.39**	0.52	0.67	0.62	1.00	0.59	0.72
Method 2	**0.51**	0.34	0.38	0.48	0.59	1.00	0.81
Method 1	**0.47**	0.48	0.52	0.63	0.72	0.81	1.00

4. Breakthrough work presented

The assessed CAIS LOS scoring methods open a clear path to implementation of a BIRAD™ rule-based lexicon for breast ultrasound findings reporting. The application is based on a CAIS Relative Similarity and LOS score calculation for a digital database of breast lesions templates with known findings. The developed software and methods facilitated a high accuracy of LOS automated scoring and promised to yield results that will impact the existing clinical practice. For that additional prospective studies should be completed with an enlarged original data set.

5. Conclusions

A large number of features were measured that correspond to lesion images used to assess CAIS LOS implementation. A considerably larger patient database and additional expert observers are needed in future work to identify image features that maximize the accuracy of categorization of masses based on CAIS LOS scoring methods assessed in this study. The five-category of both – lexicon based and CAIS based - LOS scores are suited to further Receiver-Operator Characteristic (ROC) estimation for which the calculated area under the ROC curve, A_z, can be used to compare the performance of the selected methods (based on correlations) to the radiologist. The ROC application results are being reported separately. It was statistically proven that the developed CAIS and its LOS automated scoring methods are accurate enough to advance the research in the direction of complete automation of the scoring process as a part of a computer-aided diagnostic system. The software used was a reliable and effective tool aiding the research. The methods can be utilized for education and training procedures of practitioners and technologists working in the area of diagnostic and treatment of breast cancer. The results were obtained using *Image Companion*® and *Data Companion*® proprietary software packages developed by Almen Laboratories, Inc.

References

1. M.P. André, M. Galperin, et. al., "Preliminary investigation of a method to assess breast ultrasound level of suspicion," SPIE *Medical Imaging* 4322:507-512, 2001
2. M.P. André, M. Galperin, et. al., "Improving the accuracy of diagnostic breast ultrasound", *Medical Design*, November, 2000
3. M.P. André, M. Galperin, et. al., "Internet-based medical imaging productivity system (in its application to breast ultrasound)", *MEDNET2001*, 2001, no.28

4. A.T. Stavros, D. Thickman, C.L. Rapp, M.A. Dennis, S.H. Parker, and G. A. Sisney, “Solid Breast Nodules: Use of Sonography to Distinguish between Benign and Malignant Lesions,” *Radiology* **196**, pp. 123-134, 1995.
5. P.E. Undrill, R. Gupta, S. Henry and M. Downing, “The use of texture-analysis and boundary refinement to delineate suspicious masses in mammograms,” *Proc. SPIE* **2701**, pp. 301-310, 1996.
6. M. Kallergi, G.M. Carney and J. Gaviria, “Evaluating the performance of detection algorithms in digital mammography,” *Med. Phys.* **26**, pp. 267-275, 1999.
7. V. Goldberg, A. Manduca, et. al., “Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence,” *Med. Phys.* **19**, pp. 1475-1481, 1992.
8. B.S. Garra, B.H. Krasner, S.C. Horii, et al., “Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis,” *Ultrasonic Imag.* **15**, pp. 267-285, 1993.
9. Y. Zheng, J.F. Greenleaf, J.J. Grisvold, “Reduction of breast biopsies with a modified self-organizing map,” *IEEE Trans. Neural Net,* **8**, pp. 1386-1396, 1997.
10. F. Lefebvre, M. Meunier, F. Thibault, et al., "Computerized ultrasound B-scan characterization of breast nodules," *Ultrasound in Med & Biol*, **26**(9):1421-1428, 2000.
11. N.F. Maklad, J. Ophir, I. Cespedes, H.N.F. Ponnekanti, S. Pinero, "Elastography of the breast: Preliminary results," *Radiology* **189**(P), 154 , 1993.
12. J.F. Greenleaf, S. A. Johnson, W. F. Samayoa, F. A. Duck, "Algebraic reconstruction of spatial distributions of acoustic velocities in tissue from their time-of-flight profiles," Acoustic Holography, 1975.
13. J. Bamber, "Ultrasound propagation properties of the breast," In: Ultrasonic Examination of the Breast, J. Wiley and Sons, Chichester, 1983.
14. CD Lehman, MP André, et. al., “Optical sonography applied to breast imaging.” *Academic Radiology* **7**(2):100-107, 2000.
15. M.P. André, H.S. Janée, et. al. : “High-speed data acquisition in a diffraction tomography system employing large-scale toroidal arrays." *Intl J Imaging Systems Technol* **8**(1), pp. 137-147, 1997.
16. M. Kallergi, G.M. Carney and J. Gaviria, “Evaluating the performance of detection algorithms in digital mammography,” *Med. Phys 26*, pp. 267-275, 1999.
17. A. Huo, M.L. Giger, D.E. Wolverton, et al., "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection," *Med. Phys*. **27**, pp. 4-12, 2000.
18. M. Galperin, Applied Simulation of Marine Communities, - *Akademie-Verlag*, Berlin, 1989. 240 Pages

[*] mgalperin@almenlabs.com; phone: 1-760-489-0100; http://www.almenlabs.com