

# Digital Mammogram Spiculated Mass Detection and Spicule Segmentation using Level Sets

John E. Ball, *Graduate Student Member, IEEE*, Lori Mann Bruce, *Senior Member, IEEE*

**Abstract**—This letter presents an automated mammographic computer aided diagnosis (CAD) system to detect and segment spicules in digital mammograms, termed spiculation segmentation with level sets (SSLS). SSLS begins with a segmentation of the suspicious mass periphery, which is created using a previously developed adaptive level set segmentation algorithm (ALSSM) by the authors. The mammogram is then analyzed using features derived from the Dixon and Taylor Line Operator (DTLO), which is a method of linear structure enhancement. Features are extracted, optimized, and then the suspicious mass is classified as benign or malignant. To assess the system efficacy, 60 difficult mammographic images from the Digital Database of Screening Mammography (DDSM), containing 30 benign non-spiculated cases, 17 malignant spiculated cases, and 13 malignant non-spiculated cases, are analyzed. The initial spiculation detection method found 100% of the spiculated lesions with no false positive detections, and has area under the receiver operating characteristics (ROC) curve  $A_z=1.0$ . The values using ALSSM (periphery segmentation only) are  $A_z=0.9687$  and  $0.9708$  for two investigated feature sets, and increases to  $A_z=0.9862$  using SSLS (spiculation segmentation). The best classification results are 93% overall accuracy (OA), with three false positives (FP) and one false negative (FN) using a 1-NN (Nearest Neighbor) or 2-NN classifier, and 92% OA with three FP and two FN using a maximum likelihood classifier.

**Keywords**—Breast Cancer, Cancer, Computer Aided Diagnosis, Digital Mammography, Image Segmentation, Level Sets, Spiculations, Stellate Lesions

## I. INTRODUCTION

Mammograms are the best method of detecting breast cancer [1]. According to the American Cancer Society (ACS) recommendations, women aged 40 or over should have annual mammographic screening [2]. Every thirteen minutes, four American women develop breast cancer, and one woman dies from breast cancer [3].

Since many malignant masses have a spiculated margin [4], mammographic computer aided diagnosis (CAD) systems that segment spiculations can potentially provide better results than a CAD system that does not segment spiculations. According to Winchester and Winchester,

mammographic mass analysis is based on the mass shape, margins, and density, and masses with irregular shapes or spiculations increase the probability of malignancy [5]. In digital mammograms, spiculations present in digital mammograms as nearly linear filaments extending from the mass periphery [5], and grow somewhat linearly within  $\pm 45^\circ$  normal to the mass boundary [6]. Jeske, Bernstein, and Stull state that the presence of irregular shaped masses and spiculations increase the probability of malignancy [7]. For many malignant masses, the presence of spicules is therefore very important in detecting breast cancer.

## II. PREVIOUS METHODS & BACKGROUND

In 2D image segmentation, the segmentation boundary is the zero level set of an implicit function  $\phi$ , where the segmentation boundary is controlled by the following partial differential equation  $\phi_t + F|\nabla\phi| = 0$ . The front (the locus of points with  $\phi=0$ ) advances or retreats in a direction normal to the front, and is controlled by the speed function,  $F(x,y)$ . If  $F > 0$  ( $F < 0$ ), then the level set boundary will expand outward (inward) normal to the boundary; if  $F = 0$ , then the boundary will not move [8].

Previous spiculation image enhancement methods include wavelet-based methods, statistical pixel orientation analysis, morphological processing, local gradient analysis, and custom filtering. Details on these methods may be found in [9]<sup>1</sup>. Spiculation segmentation is difficult because (1) spiculations are often ill-defined and have poor contrast [6], (2) the lack of adipose tissue (fat) in young breasts can make spiculation detection difficult [4], and (3) normal breast tissue, such as ligaments, ducts, or blood vessels, often appear as a criss-crossing set of linear structures [10].

Part of our proposed algorithm is based on extending the methods of spiculation image enhancement of Zwiggelaar *et al.* [11], which is in turn based on the Dixon and Taylor (DTLO) line enhancement algorithm [12], and these methods will be discussed in detail. Zwiggelaar *et al.* [11] analyzed several spiculation analysis methods, and found that the DTLO method provided the best performance. DTLO analysis produces a line strength map,  $S$ , indicating the potential presence of oriented lines. This method was originally developed to detect asbestos fibers. DTLO requires two parameters:  $K$ , which controls how many angular orientations are analyzed, and  $M$ , the line length

Manuscript received March 2, 2007.

J.E. Ball is with the GeoResources Institute, Mississippi State University, Starkville, MS 39759, USA (e-mail: johnball@msubulldogs.org).

L.M. Bruce is with the Electrical and Computer Engineering Department and the GeoResources Institute, Mississippi State University, Starkville, MS 39759, USA (e-mail: bruce@hpc.msstate.edu).

<sup>1</sup> Dissertations from Mississippi State University can be searched at <http://www.library.msstate.edu>. Select "Search for Materials" then choose Electronic Thesis and Dissertations.

parameter, and the method is denoted by *DTLO*( $K,M$ ). For instance if  $K=12$  and  $M=5$ , then the angular displacement set is  $\Theta = \{0^\circ, 15^\circ, \dots, 165^\circ\}$ , and the line length is 5 pixels. For further details on the DTLO algorithm, refer to [11,12].

In spiculation segmentation, several methods have been employed, including region growing and specialized graylevel feature based segmentation. Since our proposed method will be loosely based on the work of Sahiner *et al.* and Chan *et al.* [6,13], these methods will be discussed in detail. References [6,13] used a search region outside of the mass periphery and used a custom spiculation measure to create a 2D spiculation likelihood map. From this map, an experimental threshold is used to segment one pixel of the suspected spiculated area. This process is repeated 30 times (a 3mm band) in order to segment spicules. If their system classified the mass as spiculated, the spicules are added to the periphery. For each point on the boundary of the mass periphery, it utilized a circular wedge-shaped search region with radius of 4 mm, with the apex of the wedge located at a given point on the mass periphery, and the wedge size defined by  $\pm 45^\circ$  from the normal to the mass boundary,  $\vec{n}_c$ , where  $c$  indicates the  $c$ -th point on the boundary. A spiculation measure,  $x(i_c, j_c)$ , is calculated for each point  $(i_c, j_c)$  on the mass periphery. The spiculation measure is the average of the acute angle between  $\vec{n}_c$  and the angle of the image gradient at each point in the search wedge. A final post processing step removes areas disjoint from the main mass. These methods used the image gradients, which are typically very noisy for mammographic images.

The authors previously developed a mammographic CAD system for mass periphery segmentation (segmentation of the main cancerous mass, but not the spiculated portions). The system used the narrowband level set methodology [8] with an adaptive segmentation threshold controlled by a border complexity term. This method is denoted the adaptive level set segmentation method (ALSSM) [9]. This paper uses the segmentations from this method as a basis for segmenting the spiculations in spiculated masses.

### III. METHODOLOGY

This analysis uses a difficult subset from the DDSM database [14]. The 60 cases in our data set consist of 30 benign and 30 malignant cases, where 17 of the malignant cases are spiculated [9]. This work is an extension of previous research [9] where the mass periphery was segmented using the ALSSM method. Our hypothesis is that overall results can be improved if the spicules of stellate lesions are also segmented, versus segmenting only the periphery, because extracting the spicule segmentations can give a more complete representation of the suspicious mass. The following sections discuss the proposed system in detail. To the best of our knowledge, this paper presents the first spiculation segmentation of digital mammograms using level sets.

The proposed CAD spiculation system uses the mass periphery segmentation via ALSSM as a starting point for

spiculation detection. Features extracted from DTLO analysis of the area around the seed point are analyzed and a decision of spiculation presence or absence is made. If spicules are present, then the spicules are segmented using the proposed level set methodology, and the new segmentation becomes the union of the previous mass periphery and the spiculation segmentations. Based on the new segmentation, features are extracted, and then classified. In our approach, both the  $k$  nearest neighbor ( $k$ -NN) and maximum likelihood (ML) [15] classifiers are used to make a benign or malignant decision. The proposed system will be evaluated for correct identification of spiculated masses, as well as overall performance compared to ALSSM [9].

SSLS uses features from the DTLO image to classify the suspicious mass as spiculated or non-spiculated. Features extracted from a DTLO image analysis are analyzed by a kernel classifier, and is an extension of Zwiggelaar *et al.* [11]. The DTLO algorithm is enhanced to provide a more robust detection of spiculations, which is an extension of Dixon and Taylor [12]. Finally, a search area and spiculation detection method similar to Sahiner *et al.* and Chan *et al.* [6,13] are used.

Based on good results with contrast limited adaptive histogram equalization (CLAHE) in the ALSSM method [9], the original mammographic images are first processed using CLAHE to remove noise, and then are normalized. This enhancement image is denoted  $E_1$ . Next, linear structures are enhanced using a generalized version of DTLO and employing image enhancement techniques on the DTLO line strength image,  $S$ . According to Hagay, spicules may be very numerous, or spread in an irregular manner around the mass, and may have thicknesses from a tenth of a mm up to three to four mm, and lengths from several up to eight cm [16]. In order to better capture linear structures that are more than one pixel thick, a modified version of DTLO, called generalized DTLO (GDTLO), is proposed. The main difference in GDTLO and DTLO is that the former has a foreground kernel with a line of width  $N$ , while the latter has a foreground kernel with a line of width one. The GDTLO analysis will be denoted with the following nomenclature: **GDTLO**( $K,M,N$ ) denotes GDTLO where  $K$  is the number of angular orientations, and  $M$  and  $N$  are the line length and width parameters, respectively.

The GDTLO parameters selected for creation of the strength map in this analysis are  $M=19$ ,  $N=7$ , and  $K=12$ , which was experimentally determined. Three additional steps are performed once the strength map is created. First, the strength map is normalized. This image is denoted  $S_N$ . Second,  $S_N$  is processed by histogram equalization. The image resulting from both these preprocessing steps is denoted  $S_{NEQ}$ . Third, a final enhancement image is created by element-by-element multiplication of  $S_{NEQ}$  with the CLAHE enhancement image. This final step further enhances bright linear areas in the image.

In order to determine which mammograms are spiculated, statistical features are extracted from  $S_{NEQ}$ . The image graylevel histogram is analyzed using a logarithmic scale, since there can be large differences in bin values. The convention that  $\log(0)=0$  is used and the number of bins is 256. The feature vector is the log histogram. This feature vector is analyzed using generalized discriminant analysis (GDA) to provide a non-linear classification. In GDA, the data are mapped by a kernel function to a high dimensional feature space, and then linear discriminant analysis is applied. The GDA algorithm [17] is utilized with a radial basis function (RBF) kernel with  $\sigma = 10^{-2}, 10^{-1}, \dots, 10^2$ . GDA is the non-linear extension of linear discriminant analysis, and is described in detail in [18], and the Statistical Pattern Recognition toolbox for Matlab implementation was used [17].

The GDA analysis is performed using a leave one out (LOO) methodology. In a round robin fashion, one case is sequestered for testing, while the others are used for training. The distances to the nearest malignant and nearest benign neighbor are used to classify the test case, using the reduced features. Spiculation segmentation is only performed if the mass is classified by GDA as spiculated.

If a mass was classified as spiculated by the above GDA analysis, then seed points are extracted from the mass periphery in order to facilitate spiculation segmentation. The proposed seed point selection algorithm is closely based on the concepts proposed in [6,13]. For each point  $(i_c, j_c)$  on the mass periphery, the outward normal,  $\vec{n}_c$ , is calculated using the method outlined on pp. 518-519 of [19] with a parameter  $K=20$ . Then, a rectangular search area,  $R_c$ , is defined such that the rectangle extended from the periphery at  $(i_c, j_c)$  and  $R_c$  is oriented with the long axis along the direction of  $\vec{n}_c$ . Next, the pixels in  $R_c$  are extracted such that the pixel at the analysis point,  $(i_c, j_c)$ , is located at the top center pixel in  $R_c$ , and pixels radially outward from the mass boundary are filled in from top to bottom (i.e. larger radii are closer to the bottom of  $R_c$ ). The size of  $R_c$  is set to 150 pixels long (i.e., along the normal direction) by 80 pixels wide (6.53 mm long by 3.48 mm wide).

Next, a DTLO-based analysis is performed. The angles analyzed,  $\Theta_S$ , is a set of equally spaced angles in the range of  $\pm 45^\circ$  with a step size of  $2^\circ$ . For each angle  $\theta_m$ ,  $S_F(\theta_m, c)$  and  $S_B(\theta_m, c)$ , the sums of the enhanced image pixels times the corresponding foreground and background mask pixels, respectively, are calculated, where the DTLO angle,  $\theta_m$ , is the angle and the periphery pixel index is  $c$ . The spiculation feature,  $f_c$ , is

$$f_c = \max_{\theta \in \Theta_S} \{S_F(\theta_m, c) - S_B(\theta_m, c)\}. \quad (1)$$

Then, two experimentally determined thresholds,  $T_1 = 0.4$  and  $T_2 = 0.1$ , are applied in order to remove areas with low contrast or low background mask values from  $f_c$ , as follows:

$$f'_c = \begin{cases} f_c & f_c > T_1 \text{ and } S_B(\theta_m^{OPT}, c) > T_2 \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

where  $\theta_m^{OPT}$  is the optimum angle selected in eq. (1).

For each set of consecutive nonzero values in  $f'_c$ , the index with the maximum  $f'_c$  value is chosen as a seed point. Since the search areas of nearby pixels could potentially point to the same spicule, redundant seed points could be detected. Therefore, any seed point which had a neighboring seed point within ten border pixels with a lower value of  $f'_c$  is removed from the seed point list. Finally, any seed points which have an optimal direction such that a straight line in that direction would leave and then re-enter the periphery are discarded, which could happen for areas that are very convoluted.

Let  $ns$  be the final number of seed points detected. The results of the spiculation seed point detection are sets of: (1) spiculation seed points,  $SP = \{\vec{P}_1, \dots, \vec{P}_{ns}\}$ , (2) optimal angles for each seed point,  $\Theta_{OPT} = \{\theta_1^{OPT}, \dots, \theta_{ns}^{OPT}\}$ , and (3) optimal unit magnitude vectors pointing in the optimal spiculation direction,  $D_{OPT} = \{\vec{D}_1^{OPT}, \dots, \vec{D}_{ns}^{OPT}\}$ . These seed points correspond to the set of optimal locations on the periphery where the line detector response is maximized.

In order to segment the spicules, a simple, yet effective method is to examine the magnitude of the GDTLO spiculation enhancement image,  $E_{SPIC}$ , in combination with the direction difference between the optimal spiculation vector,  $\vec{D}_c^{OPT}$ , and the GDTLO angle,  $\vec{D}_{DT}$ , for each point in a rectangle oriented with the long part in the direction of  $\vec{D}_c^{OPT}$ , as follows:

$$F_c(x, y) = E_{SPIC}(x, y) \left| \left\langle \vec{D}_c^{OPT}, \vec{D}_{DT}(x, y) \right\rangle \right|, \quad (3)$$

where  $|\cdot|$  denotes absolute value and  $\langle \cdot, \cdot \rangle$  denotes the inner (dot) product. The absolute values are required since the Dixon Taylor optimal angle is limited to the range  $[0, \pi)$  and the optimal angle has range  $[0, 2\pi)$ .

A threshold is applied to the result from eq. (3). In order to remove small spurs from the image, for each pixel in  $F_c(x, y)$  which are  $\geq T_3$ , the orthogonal projection to the optimal angle vector  $\vec{D}_c^{OPT}$ , is calculated. The threshold value is experimentally determined as  $T_3 = 0.4$ . Values greater than 12 pixels away are clipped at 12 pixels. Then,

for each set of pixels in the direction of the optimal tangent  $\bar{T}_c^{OPT}$ , the values are filtered with a low pass filter in order to smooth the signal. The filter impulse response  $h[n]=1/15$  for  $n \in \{0, \dots, 14\}$ . The filter length is experimentally determined. For each pixel along the optimal angle vector  $\bar{D}_c^{OPT}$ , the mean pixel distance in the direction of  $\bar{T}_c^{OPT}$  and  $-\bar{T}_c^{OPT}$  is calculated. Let  $D_{CL}(x,y)$  be the smallest distance from the point  $(x,y)$  to the mean pixel distance center line. The final speed functions is

$$F'_c(x,y) = F_c(x,y) \exp\left(-\frac{1}{8}D_{CL}(x,y)\right) - T_3, \quad (4)$$

where the constant  $1/8$  in the exponential term is chosen to allow the spiculation to grow to a certain size tangentially, and the constant  $T_3$  ensures that the areas segmented have a minimal graylevel value. Once all of the spicules have been segmented, the spiculation segmentation becomes the union of the original mass periphery and each individual spiculation segmentation.

The features extracted include patient age, morphological features, statistical graylevel features, and features based on the segmentation boundary and the rubber band straightening transform (RBST) [19]. The features are described in Table I. Detailed references for each feature may be found in [9]. Note that all features are extracted from the original mammogram image, not the enhanced image. Feature sets A and B are feature sets I and II in [20]. There are a large number of features, and the feature set must be reduced due to the limited training samples available. Stepwise linear discriminant analysis (SLDA) using receiver operating characteristics (ROC) area under the curve ( $A_Z$ ) with forward selection and backward rejection is used to optimally select up to ten features and project them to maximize class separation. The SLDA implementation is described in [20]. The number of features are limited to ten to avoid overtraining. After feature optimization, the  $k$ -NN and ML classifiers are used to classify the tumor as malignant or benign.

Once the features are extracted, each mammogram is classified using a LOO methodology using a  $k$ -NN or ML classifier. The results are evaluated using the area under the ROC curve ( $A_Z$ ) [21-23], the  $A_Z$  confidence intervals (95% confidence level), as well as overall accuracy (OA), number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) [21].

#### IV. RESULTS AND DISCUSSION

Fig. 1 shows several example cases. In this figure, each row corresponds to one case, and the leftmost image is the original mammogram, the center image is the GTDLO enhanced image, and the rightmost image shows the physician ROI (white line), and the periphery and spiculation segmentations (black line). The GDA feature analysis is performed and all of the spiculated masses were detected.

All spiculated masses are correctly identified, with no false positives, for an overall accuracy of 100%. These results are very encouraging, and validated the intuition and stated desire of Zwiggelaar *et al.* to use a more sophisticated classifier [11]. The overall  $A_Z=1.0$ , indicating that the training data are perfectly separable, using  $\sigma=10^{-2}$ .

From Table II, it can be seen that in general, the SSSL (based on periphery plus spiculation segmentation) results outperform the corresponding ALSSM (periphery only segmentation) results in terms of higher OA, lower number of FN, and higher  $A_Z$  values. For feature set A, which contains patient age, morphological, statistical, and NRL features, the results increased the highest  $k$ -NN OA from 87% to 88%. The ML classifier had a striking difference in OA, going from 80% with ALSSM to 88%. In general, the ML classifier had more FN and fewer FP than the  $k$ -NN classifier, which is an undesirable result. Feature set B, which includes the features in feature set A as well as the SGLD texture features, showed improvements. The best results were for  $k=5$ , which improved the overall accuracies from 82% for ALSSM to 88% and 90%, and reduced the FN from 6 for ALSSM to 3. ML achieved 90% versus 87% OA for ALSSM. For feature set C, which is the patient age and the SGLD features, the results were the best. One striking

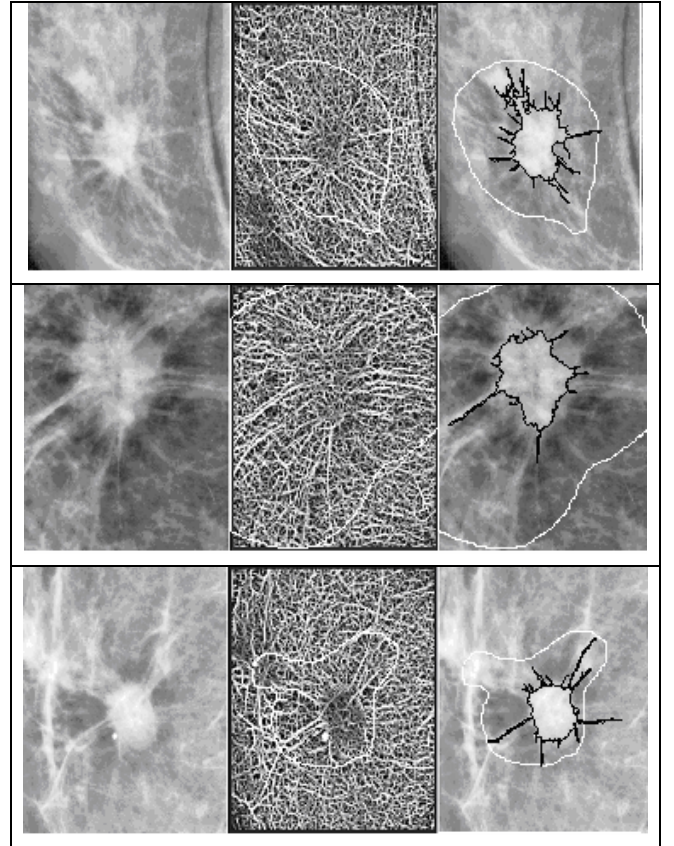


Fig. 1. Selected spiculated mammograms showing from left to right: CLAHE enhanced image, GTDLO image, ALSSM periphery segmentation, and the proposed spiculation segmentation, respectively. In the two rightmost images, the physician supplied ROI is shown in white. If the ROI is larger than the portion of the image shown, then it is clipped. The overall spiculation segmentations are shown in black.

difference is the large increase in overall accuracy for  $k=1$  and 2: ALSSM has 77% overall accuracy, and the SSLS result is 93% OA. The number of FN was reduced from 5 to 1. The ML classifier had similar results.

Comparing all of these results, some of the features selected by SLDA in feature sets A and B had some class confusion. This also points out that adding more features may not necessarily improve results, which may be a result of the limited training set size (59 images in each LOO iteration). Furthermore, SLDA was used, which did not provide an exhaustive search for best feature selection. The  $A_z$  values for SSLS analysis improved for all feature sets, indicating that the spiculation features are more powerful discriminators than the periphery features alone.

## V. CONCLUSIONS

The proposed SSLS method is based on extensions of the spiculation image enhancement of Zwiggelaar *et al.* [11], by the DTLO line enhancement algorithm of Dixon and Taylor [12], and by the spiculation segmentation methods of Sahiner *et al.* and Chan *et al.* [6,13], and provides an important extension of their work. The level set methodology provided a fast and efficient means of segmenting the spiculated regions. Fortunately, since spicules are fairly linear and grow outwards from the periphery, a simple speed function could be generated to segment the spicules.

The GDA method proved to be very effective in analyzing the images for spiculation presence. For the DDSM images in this study, each spiculated image is detected properly, and no non-spiculated images are falsely classified as spiculated. The GDTLO enhancement algorithm proved to be very effective at highlighting linear structures. This can clearly be seen in Fig. 1. The proposed method performed very well, and in almost all cases, performed better than segmentation of the periphery alone. The best results are obtained from feature set C (patient age and the SGLD features), for  $k=1$  and 2 with the  $k$ -NN classifier, where the OA are 93%, 3 FP and 1 FN. In general, the  $k$ -NN classifier outperformed the ML classifier slightly in terms of higher OA and fewer FN.

Finally, there are very few CAD systems that utilize spiculation segmentation, and many of them utilize features derived from the image gradient in order to segment spicules, and since mammographic images are very noisy, the proposed method, based on the DTLO line operator, could provide a robust alternative which may be less sensitive to noise. Future work includes running the proposed system on different datasets and on a larger dataset from the DDSM database.

## REFERENCES

- [1] National Cancer Institute, "National Cancer Institute Fact Sheet: Improving Methods for Breast Cancer Detection and Diagnosis," 2006. Available: <http://www.cancer.gov/cancertopics/screening/breast>.
- [2] R.A. Smith, V. Cokkinides, and H.J. Eyre, "American Cancer Society Guidelines for the Early Detection of Cancer, 2006," *CA: A Cancer Journal for Clinicians*, vol. 56, no. 1, pp. 11-25, 2006.
- [3] V.F. Andolina, S.L. Lillé, and K.M. Willison, *Mammographic Imaging: A Practical Guide*. New York, NY: Lippincott Williams & Wilkins, 1992.
- [4] R.L. Egan, *Breast Imaging: Diagnosis and Morphology of Breast Diseases*. Philadelphia, PA: W. B. Saunders Co., 1988.
- [5] D.J. Winchester and D.P. Winchester, *American Cancer Society Atlas of Clinical Oncology: Breast Cancer*. London, UK: B.C. Decker, 2000.
- [6] H.-P. Chan, N. Petrick, and B. Sahiner, "Computer-Aided Breast Cancer Diagnosis," in *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, A. Jain, A. Jain, S. Jain, and L. Jain, Eds. River Edge, NJ: World Scientific, pp. 179-264, 2002.
- [7] J.M. Jeske, J.R. Bernstein, and M.A. Stull, "Screening and Diagnostic Imaging," in *American Cancer Society Atlas of Clinical Oncology* London, UK: B.C. Decker, pp. 41-63, 2000.
- [8] J.A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, 2nd ed. Cambridge, UK: Cambridge University Press, 2002.
- [9] J.E. Ball, "Three stage level set segmentation of mass core, periphery, and spiculations for automated image analysis of digital mammograms." Ph.D. in Electrical Engineering. Starkville, MS: Mississippi State Univ., May 2007.
- [10] N. Karssemeijer, "Detection of Masses in Mammograms," in *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, A. Jain, A. Jain, S. Jain, and L. Jain, Eds. River Edge, NJ: World Scientific, pp. 187-212, 2000.
- [11] R. Zwiggelaar, S.M. Astley, C.R.M. Boggis, and C.J. Taylor, "Linear structures in mammographic images: detection and classification," *IEEE Trans. on Medical Imaging*, vol. 23, no. 9, pp. 1077-1086, Sept. 2004.
- [12] R.N. Dixon and C.J. Taylor, "Automated Asbestos Fiber Counting," *1979 Inst. Physics Conference*, vol. 44, pp. 178-185, 1979.
- [13] B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, and L.M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics*, vol. 28, no. 7, pp. 1455-1465, Jul. 2001.
- [14] M. Heath, K.W. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumar, "Current status of the Digital Database for Screening Mammography," in *Digital Mammography*, N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, Eds. Boston, MA: Kluwer Academic Publishers, pp. 457-460, 1998.
- [15] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2001.
- [16] C. Hagay, "Stellate Images," in *Mammography: A Guide to Interpretation*, Translated ed. A. Le Treut and M. H. Dilhuydy, Eds. St. Louis, MO: Mosby Year Book, pp. 66-78, 1991.
- [17] V. Franc and V. Hlavac, "Statistical Pattern Recognition Toolbox for Matlab User's Guide, Document Number CTU-CMP-2004-08," version 2.05: Center for Machine Perception, Czech Technical University. Jun. 2004. available: <http://cmp.felk.cvut.cz/>.
- [18] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385-2404, Oct. 2000.
- [19] B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, and M.M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics*, vol. 25, no. 4, pp. 516-526, Apr. 1998.
- [20] J.E. Ball and L.M. Bruce, "Digital Mammographic Computer Aided Diagnosis (CAD) using Adaptive Level Set Segmentation," *IEEE Trans. on Medical Imaging*, [under review].
- [21] M. Kallergi, "Medical Image Analysis Methods," in *The Electrical Engineering and Applied Signal Processing Series*, L. Costaridou, Ed. New York: Taylor & Francis, pp. 433-472, 2005.
- [22] J.A. Hanley and B.J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29-36, Apr. 1982.
- [23] A. Agatheeswaran, "Analysis of the effects of JPEG2000 compression on texture features extracted from digital mammograms". Masters Thesis in Electrical and Computer Engineering. Starkville, MS: Mississippi State University, pp. 20-37, 42-43, Dec. 2004.

TABLE I  
FEATURE LIST. FEATURE SETS A AND B ARE THE SAME FEATURE SETS I AND II IN [20].

Feature Set			Feature Type and Source <sup>1</sup>	Feature List	Number of Features
A	B	C			
•	•	•	Patient age (DDSM)	Age	1
•	•		Morphological (SB)	Area, Axis ratio, Box ratio, Circularity, Convex hull area, Eccentricity, Equivalent diameter, Extent, Extent ratio <sup>2</sup> , Major axis length, Minor axis length, Perimeter length, Solidity, Width to height ratio	14
•	•		Statistical (SB)	Gray level mean, Gray level std. dev, Gray level std. dev. ratio <sup>3</sup>	3
•	•		NRL (SB)	Entropy, Length, Mean, Roughness, Std. dev., Zero crossing count	6
•	•	•	GLCM (SB)	(Note <sup>4</sup> ) Energy, Variance, Correlation, Inertia, Inverse Difference Moment, Entropy	144
•	•	•	GLCM (RBST)	(Note <sup>5</sup> ) Energy, Variance, Correlation, Inertia, Inverse Difference Moment, Entropy	864

<sup>1</sup> This denotes the region from which the features were extracted. DDSM=DDSM database (there is no region, as the patient age is part of the database). SB=segmentation boundary. NRL stands for Normalized Radial Length. RBST=Rubber Band Straightening Transform [19]. GLCM stands for gray level co-occurrence matrix. GLCM is also known as spatial gray level dependence (SGLD). <sup>2</sup> The extent ratio is  $\max(\text{length}, \text{height}) / \min(\text{length}, \text{height})$ . <sup>3</sup> the Gray level std. dev. ratio is the ratio of the std. dev. of the gray levels inside the segmentation to the std. dev. of gray levels outside the segmentation boundary and within 200 pixels of the segmentation boundary. <sup>4</sup> The GLCM SB features are calculated at distances  $d=\{1,2,4,6,8,10\}$  and directions  $\theta=\{0^\circ,45^\circ,90^\circ,135^\circ\}$ . There will be a total of 6 GLCM features x 6 distances x 4 angles for 144 features. <sup>5</sup> The RBST features are the same features as the GLCM SB features. The RBST uses a parameter  $k$  to choose how many pixels before and after are used to create the normal vector to the spiculation boundary [19]. The RBST features are calculated for distances  $k=\{2,4,6,8,10,12\}$  For each value of  $k$ , there will be 144 features generated. Therefore there are  $864 = 144 \times 6$  features.

TABLE II  
OVERALL RESULTS. THE BEST RESULTS SHOWN IN BOLD (BEST = HIGHEST OVERALL ACCURACY WITH LOWEST NUMBER OF FN). THE FEATURE SETS (A, B, AND C) ARE DEFINED IN TABLE I. UNDER THE  $A_z$  COLUMN, THE LOWER NUMBERS INDICATE THE 95% CONFIDENCE INTERVAL.

FS	SSLS Analysis I							SSLS Analysis II						ALSSM					
	CL	OA	TP	TN	FP	FN	$A_z$	OA	TP	TN	FP	FN	$A_z$	OA	TP	TN	FP	FN	$A_z$
A	1-NN	82	25	24	6	5	0.9862 ± 0.0006	87	28	24	6	2	0.9862 ± 0.0006	82	25	24	6	5	0.9687 ± 0.0014
	2-NN	82	25	24	6	5		87	28	24	6	2		82	25	24	6	5	
	3-NN	83	25	25	5	5		88	28	25	5	2		80	25	23	7	5	
	4-NN	82	24	25	5	6		87	27	25	5	3		85	27	24	6	3	
	5-NN	<b>88</b>	<b>27</b>	<b>26</b>	<b>4</b>	<b>3</b>		<b>90</b>	<b>28</b>	<b>26</b>	<b>4</b>	<b>2</b>		85	27	24	6	3	
	6-NN	87	27	25	5	3		88	28	25	5	2		<b>87</b>	<b>28</b>	<b>24</b>	<b>6</b>	<b>2</b>	
	7-NN	87	27	25	5	3		88	28	25	5	2		<b>87</b>	<b>28</b>	<b>24</b>	<b>6</b>	<b>2</b>	
	8-NN	87	27	25	5	3		88	28	25	5	2		<b>87</b>	<b>28</b>	<b>24</b>	<b>6</b>	<b>2</b>	
	9-NN	87	27	25	5	3		88	28	25	5	2		85	27	24	6	3	
	10-NN	87	27	25	5	3		88	28	25	5	2		85	27	24	6	3	
	ML	88	25	28	2	5		90	26	28	2	4		80	21	27	3	9	
B	1-NN	82	25	24	6	5	0.9862 ± 0.0007	87	28	24	6	2	0.9862 ± 0.0007	85	25	26	4	5	0.9708 ± 0.0017
	2-NN	82	25	24	6	5		87	28	24	6	2		85	25	26	4	5	
	3-NN	83	25	25	5	5		88	28	25	5	2		82	24	25	5	6	
	4-NN	82	24	25	5	6		87	27	25	5	3		82	24	25	5	6	
	5-NN	<b>88</b>	<b>27</b>	<b>26</b>	<b>4</b>	<b>3</b>		<b>90</b>	<b>28</b>	<b>26</b>	<b>4</b>	<b>2</b>		82	24	25	5	6	
	6-NN	87	27	25	5	3		88	28	25	5	2		83	25	25	5	5	
	7-NN	87	27	25	5	3		88	28	25	5	2		85	26	25	5	4	
	8-NN	87	27	25	5	3		88	28	25	5	2		<b>87</b>	<b>27</b>	<b>25</b>	<b>5</b>	<b>3</b>	
	9-NN	87	27	25	5	3		88	28	25	5	2		<b>87</b>	<b>27</b>	<b>25</b>	<b>5</b>	<b>3</b>	
	10-NN	87	27	25	5	3		88	28	25	5	2		<b>87</b>	<b>27</b>	<b>25</b>	<b>5</b>	<b>3</b>	
	ML	88	25	28	2	5		90	26	28	2	4		87	26	26	4	4	
C	1-NN	<b>93</b>	<b>29</b>	<b>27</b>	<b>3</b>	<b>1</b>	0.9849 ± 0.0006	<b>93</b>	<b>29</b>	<b>27</b>	<b>3</b>	<b>1</b>	0.9849 ± 0.0006	77	25	21	9	5	0.9679 ± 0.0018
	2-NN	<b>93</b>	<b>29</b>	<b>27</b>	<b>3</b>	<b>1</b>		<b>93</b>	<b>29</b>	<b>27</b>	<b>3</b>	<b>1</b>		77	25	21	9	5	
	3-NN	92	29	26	4	1		92	29	26	4	1		80	25	23	7	5	
	4-NN	92	29	26	4	1		92	29	26	4	1		82	26	23	7	4	
	5-NN	92	29	26	4	1		92	29	26	4	1		83	26	24	6	4	
	6-NN	92	29	26	4	1		92	29	26	4	1		83	26	24	6	4	
	7-NN	90	28	26	4	2		90	28	26	4	2		85	26	25	5	4	
	8-NN	92	29	26	4	1		92	29	26	4	1		85	26	25	5	4	
	9-NN	90	28	26	4	2		90	28	26	4	2		85	26	25	5	4	
	10-NN	92	29	26	4	1		92	29	26	4	1		<b>87</b>	<b>26</b>	<b>26</b>	<b>4</b>	<b>4</b>	
	ML	90	27	27	3	3		92	28	27	3	2		<b>87</b>	<b>26</b>	<b>26</b>	<b>4</b>	<b>4</b>	