A validation model for segmentation algorithms of digital mammography images

Kenneth A. Byrd
Howard University
k_byrd@howard.edu

Jianchao Zeng
Howard University
jzeng@howard.edu

Mohamed Chouikha
Howard University
mchouikha@howard.edu

ABSTRACT

We present a comprehensive validation analysis to evaluate the performance of three existing digital mammography segmentation algorithms against manual segmentation results produced by two expert radiologists. This is an improvement of an early methodology used for the evaluation of boundary algorithms on medical images. The mammography images used were acquired from the Digital Database for Screening Mammography (DDSM) and subsequently used as ground truth. We assessed three existing segmentation algorithms: (a) the Region Growing Combined with the Maximum Likelihood (ML) Model, (b) the Gradient Vector Flow (GVF) Model, and (c) the Standard Potential Field (STD) Model. We applied a comprehensive statistical metric. We concluded that the Region Growing Combined with the Maximum Likelihood (ML) Model yielded not only the best accuracy, specificity, percent error, and algorithm ranking, but also the greatest ratio of average computer-to-observer agreement and average inter-observer agreement (WI'). We also noted that the upper limit of the 95% Confidence Interval (CI) was greater than 1.0, and thus each individual observer is a reliable member of the group. These studies are especially important for the development of computer-aided diagnosis (CAD) systems for cancer.

INDEX TERMS

Computer-aided diagnosis, Mammography, Segmentation, Validation

I. INTRODUCTION

A standard comprehensive metric is needed to assess the robustness and effectiveness of existing medical image segmentation algorithms such as [2]-[3]-[4]-[5]. There is a need for one standard evaluation procedure that will correlate multiple data measurements and synthesize them as a single output, both quantitatively and qualitatively. Scientists and engineers in [6]-[7]-[8]-[10] have conducted evaluation studies using different criteria and statistical sets. This fact alone makes it difficult to compare the performances of their algorithms against each other's.

Chalana and Kim [1] proposed a methodology for evaluating and comparing boundary detection algorithms for medical image segmentation. We improve upon this model in our research with the addition of four "validation measures": overlap; accuracy; sensitivity and specificity; and testing on a different imaging modality, digital mammography. The validation measures were added to verify and further support the findings of the Chalana-Kim Methodology. Also, we made use of contour centroid data as a distance metric instead of the pixel-by-pixel comparison noted in [1].

The demand for effective and efficient CAD systems is at an all-time high, especially in the medical field and more importantly in the area of digital mammography. It is however necessary to have systems and protocols that can aid in the diagnosis decision of patients by physicians. Figure 1 illustrates our validation model.

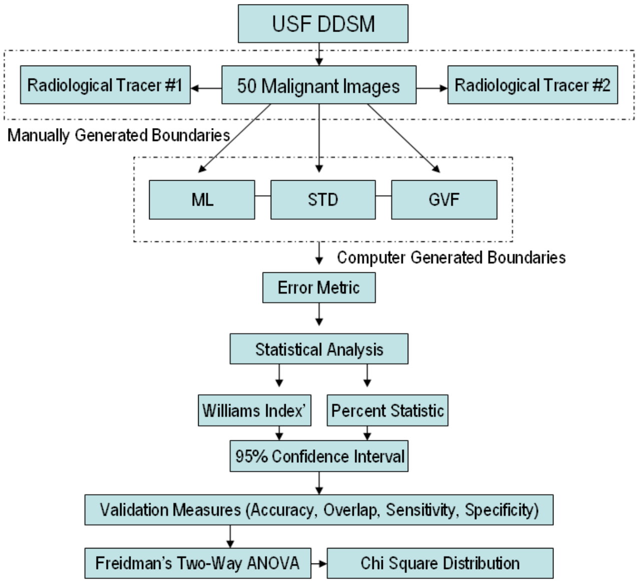


Figure 1. Digital mammography segmentation algorithm validation protocol

II. METHODOLOGY

We chose a dataset of 50 cancerous mammography images from the DDSM, which currently has 2,620 “normal,” “cancer,” “benign,” and “benign without callback” cases organized into 26 volumes [10]. Each volume is a collection of cases of the corresponding type. Cancer cases are formed from screening exams in which at least one pathologically proven cancer was found.

After the 50 DDSM images were obtained, the cancerous mass borders were outlined by expert radiologists. We designated the expert-outlined boundaries (EOBs) as our “ground truth” and compared them with the computer-generated boundaries (CGBs) of the same 50 images. Two radiologists supplied the expert traces. We extracted the mass borders using a software tool and superimposed them onto a 512 x 512-pixel black image, for a total of 100 ground truth images. Figure 2 demonstrates a segmentation of a digital mammogram by the two radiologists.

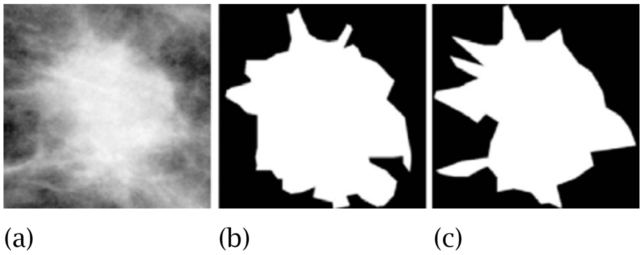


Figure 2. (a) Original mammogram (b) Radiologist #1 ground truth trace (c) Radiologist #2 ground truth trace

A. Algorithm Design

1. Region growing combined with the Maximum Likelihood Model (ML): The ML Model makes use of three operations/properties to find the contour which best represents a mass and its extended borders. This model combines region growing, maximum likelihood, and area analysis. The contours are grown using a 4-neighbor region growing model.

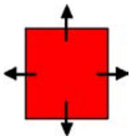


Figure 3. Example of a pixel

Figure 3 represents a seed pixel with an intensity of 100 evaluated utilizing the 4-neighbor region growing method. If the intensities of the pixels surrounding the seed pixel are greater than or equal to the seed pixel intensity (threshold), the surrounding pixels are included in the region of interest (ROI). This procedure uses the highest intensity as the seed point and decreases the intensity value in successive steps [2]. Each pixel in the graph will at some point in time be the seed pixel.

Figure 4 shows a visual picture of the seed pixel and its surrounding pixels. The pixel to the right end of the seed pixel and the pixel below it were included in our region because their intensities were greater than the intensity of the seed pixel.

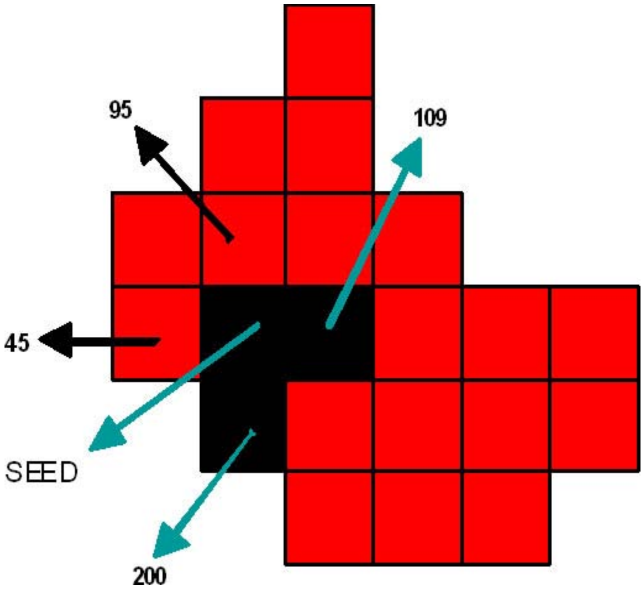
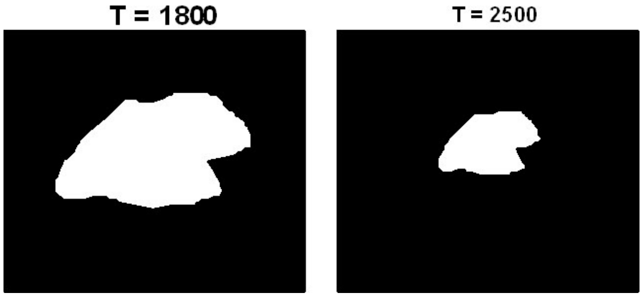


Figure 4. 4-Neighbor region growing example

This procedure generates a sequence of contours that represents the mass but does not determine the boundaries obscured by the other tissue. We cannot use the procedure to choose the contour that is most highly correlated with the experts' perceptions. In addition, an adaptive region growing technique was applied to find the contour that accurately represents the mass body contour (for a particular shadow size) and distribution used to model the intensity values. This model is able to delineate the mass body contour as well as its extended borders. This model combines region growing, likelihood function analysis and our function analysis for choosing this contour.

The resultant image is then multiplied by a two-dimensional trapezoidal membership function. This multiplication yields what we call the "fuzzified" image [2]. As the threshold for intensity increases, the size of the contour decreases. This is demonstrated in Figure 5.



(a) (b)

Figure 5. (a) Threshold for intensity image = 1800 (b) Threshold for intensity image = 2500

The next step is to find the histogram inside and outside of the contour. This is done by projecting the fuzzified image onto the original image, as demonstrated in Figure 6.

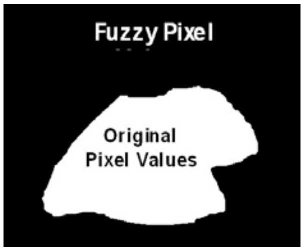
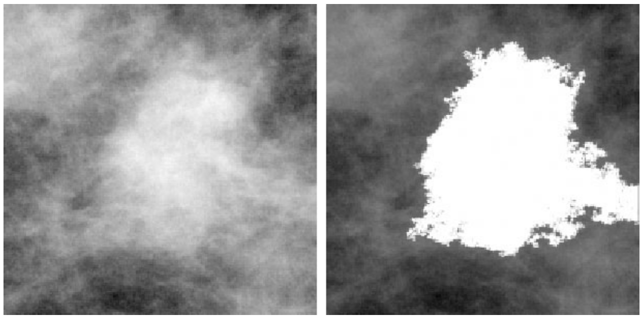


Figure 6. Fuzzy pixel value masked image

The Base-10 logarithm of the composite probability of the two regions is then computed. The computation consists of taking the log of the summation of the histogram outside the contour plus the summation of the histogram inside the contour. The likelihood that the contour represents the mass body is determined by assessing the maximum likelihood function. The likelihood is found by assessing the maximum value of the likelihood values as a function of the intensity threshold. This gives us the optimal density needed to delineate the mass body contour.

Finally, steep changes in the likelihood function were examined. The steepest changes correspond to intensities which will produce contours that are most likely to correlate with our gold standard radiological traces. It was proven in [9] that the intensity corresponding to the second steep change always yields contours with higher sensitivity values; however, the intensity corresponding to the first steep change is the wisest choice among the three variable choices as seen in Figure 7. The first steep change frequently appears in the range of intensities for which the contours do not experience substantial flooding.



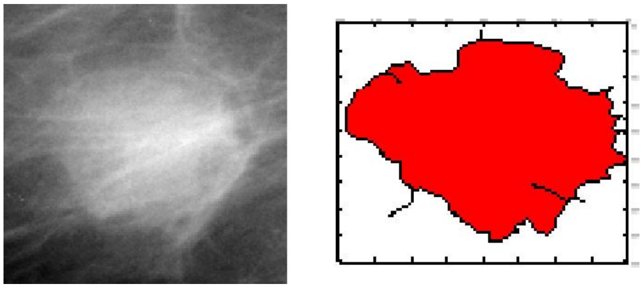
(a) (b)

Figure 7. (a) Original mammogram (b) Computer-generated segmentation using the ML Model

2. Standard Potential Field Model (STD): The STD Model uses a parametric snake as its basis. A traditional snake is a curve x(s) = [x(s), y(s)], (element in [0]-[1]) that moves through the spatial domain of an image to minimize the energy function:

$$E = \int_0^1 \frac{1}{2}\left[\alpha|x'(s)|^2 + \beta|x''(s)|^2\right] + F_{ext}(x(s))ds \qquad (1)$$

where α and β are weighting parameters that control the snake's tension and rigidity and x'(s) and x"(s) denote the first and second derivatives of x(s) with respect to s. The external energy function, F_{ext} , is derived from that image so that it takes the smaller values at the features of interest, such as boundaries. The internal force F_{int} discourages stretching and bending while the external potential force pulls the snake toward the desired image edges. Figure 8 shows an example of a segmentation using the STD Model.



(a) (b)

Figure 8. (a) Original mammogram (b) Computer-generated segmentation using the Standard Model

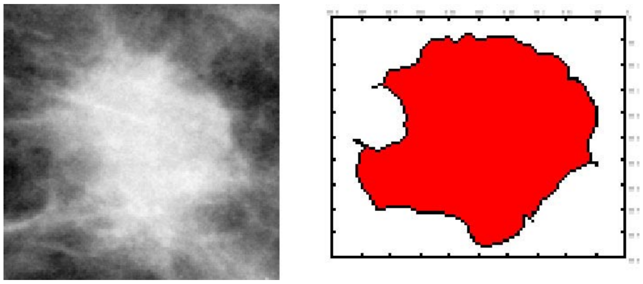
3. Gradient Vector Flow Model (GVF): The GVF Model makes use of a modified force balance condition as its basis. A new external force field is defined in the model by v(x,y) which is known as the gradient vector field or GVF. The corresponding dynamic snake equation is obtained by replacing the potential force with v(x,y), yielding:

$$x_t(s,t) = \alpha x''(s,t) - \beta x''''(s,t) + v \quad (2)$$

The parametric curve solving the above dynamic equation is called the GVF Snake and can be calculated numerically by discretization and iteration. This GVF Field is defined as the vector field $v(x,y) = [u(x,y), v(x,y)]$ that minimizes the energy function:

$$\varepsilon = \iint \mu(\mu_x^2 + \mu_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |v - \nabla f|^2 dxdy \quad (3)$$

Figure 9 depicts a segmentation using the GVF Snake Model.



(a) (b)

Figure 9. (a) Original mammogram (b) Computer-generated segmentation using the GVF Model

B. Error Metric Definition

An error metric is necessary to help find the absolute difference in their corresponding points for the intra-observer study (EOBs vs. CGBs) as well as the inter-observer studies (EOBs vs. EOBs and CGBs vs. CGBs). We used a centroid-based comparison to measure the distance between two boundaries. Each point on the computer and manually segmented contours is used to calculate the centroid of each image. The algebraic means were found for both the x and y coordinates of the centroid. We calculated the Euclidean distances between subsequent measurements (CGB to CGB, EOB to EOB, and EOB to CGB).

C. Statistical Analysis

We used two statistical tests to complete this daunting task: (1) a modified version of the Williams Index (WI'), which compares the ratio between the average computer-to-observer agreement and the average inter-observer agreement, (2) the Percent Statistic Model, which computes the percentage of observations for which the CGB lies within the inter-observer range.

1. Williams Index (WI'): The proportion of agreements between two observers is equal to the reciprocal of the average disagreements, $D_{j,j'}$, between observers j and j'.

$$P_{j,j'} = \frac{1}{D_{j,j'}} \quad (4)$$

Where the average disagreement between the two observers is

$$D_{j,j'} = \frac{1}{N}\sum_{i=1}^{N} e(x_{ij}, x_{ij'}) \quad (5)$$

$$I' = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{o,j'}}}{\frac{2}{n(n-1)}\sum_{j}\sum_{j':j'\neq j}\frac{1}{D_{j,j'}}} \quad (6)$$

The CI for this index is computed using the jackknife non-parametric sampling technique. This model works by leaving out N – 1 observations during each calculation. The jackknife estimate of the standard error in the computation of WI' is given by:

$$se = \left\{\frac{1}{N-1}\sum_{i=1}^{N}[I'_{(i)} - I'_{(\cdot)}]^2\right\}^{1/2} \quad (7)$$

where

$$I'_{(\cdot)} = \frac{1}{N}\sum_{i=1}^{N} I'_{(i)} \qquad (8)$$

The 95% CI for the estimate of the modified WI is

$$I'_{(\cdot)} \pm z_{0.95} se \qquad (9)$$

where $z_{0.95} = 1.96$ is the 95th percentile of the standard normal distribution.

2. Percent Statistic (PS): The PS represents the percentage of cases for which the CGB (or measurement) lies within the inter-observer range. A CGB is defined to be within the inter-observer range if it lies within the convex hull formed by the EOB's as seen in Figure 10.

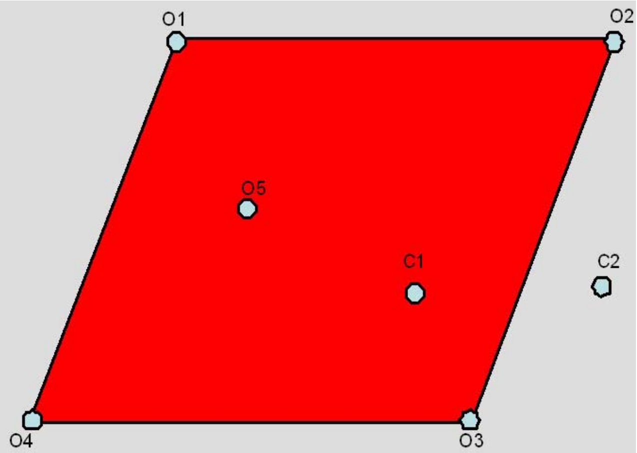


Figure 10. Convex hull of the EOB's with the CGB's

This figure represents five observer and two computer-generated boundaries (or measurements). The points O1, O2, O3, O4, and O5 represent the five observer-outlined boundaries, and the points C1 and C2 represent the two CGBs. The shaded area is the convex polygon which bounds all the EOBs. The computer-generated boundary C1 lies inside this convex polygon and is within the inter-observer range, whereas C2 lies outside this range.

A point C lies in the convex hull of a set of points O1, O2, …, On, if:

$$\max_{i}\{e(C,O_i)\} \leq \max_{i,j}\{e(O_i,O_j)\} \qquad (10)$$

mainly whether the maximum computer-to-observer distance is less than or equal to the maximum inter-observer distance. The expected probability that one observer's boundary lies outside the range of the other observers' boundaries is 1/(n+1). Thus, under the hypothesis that the CGBs and the EOBs are samples from the same distribution, the expected percent of times that the CGBs lie within the inter-observer range is n/(n+1). This is based on the hypothesis that n+1 observers produce boundaries which are samples from the same distribution.

This expected percentage is 67% for two human observers. We compute the 95% CI of the percentage statistic and check whether it includes the expected value to test whether the data is consistent with the hypothesis. If the data are not consistent with the hypothesis that the CGBs and observer-outlined boundaries are samples from the same distribution, then the CI will not include the expected value. The WI' provides information about averages, because it computes the ratio between the average computer-to-observer agreement and the average inter-observer agreement. The PS gives information about corresponding relationships between the computer measurements and the observer measurements. PS is useful because it tells us the number of times that the algorithm is successful, and it produces boundaries which are within the inter-observer range [1].

D. Validation Measures

We evaluated the three segmentation algorithms using four validation measures: overlap, accuracy, sensitivity, and specificity. These measures were computed for both expert radiological traces.

$$(11) \qquad \text{Overlap} = \frac{N_{TP}}{N_{TP} + N_{FP} + N_{FN}}$$

$$(12) \qquad \text{Accuracy} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}$$

$$(13) \qquad \text{Sensitivit y} = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

$$(14) \qquad \text{Specificity} = \frac{N_{TN}}{N_{TN} + N_{FP}}$$

Overlap is the amount of intersection between the EOB and the CGB. Accuracy is the ratio of correctly classified pixels to the entire area of the ROI. Sensitivity is a true positive measure in that it refers to the proportion of images that contain a cancerous mass which have been classified correctly. Specificity is a true negative measure that refers to the proportion of images containing a cancerous mass that have been incorrectly classified. Ground Truth is the drawings produced by the two radiologists. NTP is the true positive measurement, NTN is the true negative measurement, NFP is the false positive measurement (portion of the image incorrectly classified as cancerous mass), and NFN is the false negative measurement (portion of the image incorrectly classified as not a portion of the cancerous mass). Figure 11 shows each of the four positive-negative regions and relationships graphically:

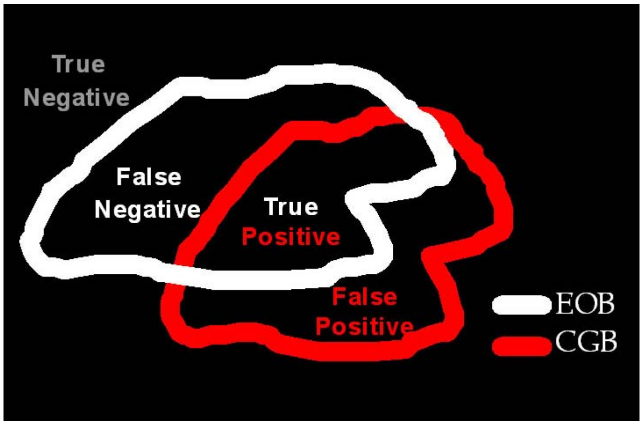


Figure 11. Positive-negative regions and relationships where the regions have the following binary relationships:
NTP = True Positive = EOB AND CGB (15)
NTN = True Negative = ~EOB AND ~CGB (16)
NFP = False Positive = ~EOB AND CGB (17)
NFN = False Negative = EOB AND ~CGB (18)

E. *Comparison of Algorithms*

Often there is a need to compare the results of applying several different algorithms on the same data set. The algorithm which results in a small overall error is preferred over the other algorithms. We carried out the comparison of errors using Friedman's two-way ANOVA test by ranks.

This test is used when either a matched-subjects or repeated-measure design is used, and the hypothesis of a difference among three or more treatments (or algorithms) is to be tested. Friedman's test is a non-parametric test, involving ranking of the errors caused by using the different algorithms for each data set. The null hypothesis is that each algorithm performs identically, and thus the average rank for each algorithm over the entire data set is the same. The Friedman test statistic (Q[r]) is:

$$Q_R = \frac{12}{Nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3N(k-1) \qquad (19)$$

which is approximated by a P^2 statistic with k-1 degrees of freedom, where:

N = the number of data sets (or images)
k = the number of algorithms
R_j = the sum of ranks for the algorithm j

III. EXPERIMENTAL RESULTS

A. *Validation Measures*

We applied a comprehensive evaluation protocol to evaluate the computer and expert-outlined segmentation results. We calculated the absolute error utilizing a centroid comparison method and a statistical evaluation of the WI', PS, and 95% CI. Comparison of the algorithms was carried out using Friedman's two-way ANOVA test and was compared to the comparable chi-square distribution. Additionally, four validation measures were employed: overlap, accuracy, sensitivity and specificity.

Boundary overlap is representative of how much a cancerous mass intersects with a gold standard trace. Utilizing the ML Model, Radiologist #2's traces produced the most overlap with a mean accuracy of 53.56%. For the STD Model, Radiologist #1 produced the most overlap, with a mean of 54.50%. Radiologist #1 yielded a higher mean overlap, with 53.10% for the GVF Model. These results are shown in row 4, columns 2 and 3, in Tables I, II, and III.

"Accuracy" represents the ratio of correctly classified pixels to the entire area of the ROI. Utilizing the ML Model, Radiologist #2's traces were most accurate yielding a mean accuracy of 77.35%. For the STD Model, Radiologist #2 was most accurate, with a mean accuracy of 71.78%. For the GVF Model, Radiologist #1's traces were more accurate, yielding a mean of 74.87%. These results are shown in row 5, columns 2 and 3, in Tables I, II, and III.

The sensitivity measure represents the probability that pixels are classified as truly diseased (true-positive). Utilizing the ML Model, Radiologist #2's traces yielded a higher sensitivity rate, with a mean of 71.13%. For the STD Model, Radiologist #2 yielded a sensitivity rate with a mean of 84.43%. Radiologist #2 yielded a higher mean sensitivity rate of 88.45% for the GVF Model. These results are shown in row 6, columns 2 and 3 in Tables I, II, and III.

The "specificity" measure represents the probability that pixels are classified as truly *not* diseased (true-negative). Utilizing the ML Model, Radiologist

#1's traces yielded a higher specificity rate, with a mean of 89.54%. For the STD Model, Radiologist #1 yielded a mean specificity rate of 77.24%. Radiologist #1 yielded a higher mean specificity rate of 77.09% for the GVF Model. These results are shown in row 7, columns 2 and 3 in Tables I, II, and III.

Table I. Maximum likelihood model validation measures

MAXIMUM LIKELIHOOD (ML) MODEL		
	Radiologist #1	Radiologist #2
Overlap	0.5318	0.5356
Accuracy	0.7574	0.7735
Sensitivity	0.6234	0.7113
Specificity	0.8954	0.8529

Table ll. Standard Potential Field Model validation measures

STANDARD POTENTIAL FIELD (STD) MODEL		
	Radiologist #1	Radiologist #2
Overlap	0.5450	0.5433
Accuracy	0.7142	0.7178
Sensitivity	0.7507	0.8443
Specificity	0.7724	0.6795

Table III. Gradient Vector Flow Model validation measures

GRADIENT VECTOR FLOW (GVF) MODEL		
	Radiologist #1	Radiologist #2
Overlap	0.5310	0.5030
Accuracy	0.7487	0.7307
Sensitivity	0.8052	0.8845
Specificity	0.7709	0.608

Each of the four presented validation measures is paramount in our study. Moreover, "Accuracy" which represents the ratio of correctly classified pixels to the entire area of the ROI should have a direct correlation to the measurements for the "preferred algorithm", which is given by result of the two-way ANOVA test by ranks.

Utilizing the ML Model, Radiologist #2's traces were most accurate in comparison to that algorithm, with an average accuracy of 77.35%. For the GVF Model, Radiologist #1 was most accurate with a mean of 74.07% and for the STD Model, Radiologist #2 was once again most accurate with a 71.70% accuracy measure. We may also deduce that the ML Model is the most accurate algorithm, given the average accuracy of the two radiological traces.

B. Williams Index (WI') and Percent Statistic (PS)

Table IV. (a) WI and PS for GVF (b) STD (c) ML Models

SE	0.3521
WI'	0.7040
95% CI	(0.0142, 1.3941)
PS	0.9600
95% CI	(0.5739, 1.3460)

(a)

SE	0.2890
WI'	0.7163
95% CI	(-0.1213,1.2827)
PS	0.9500
95% CI	(0.5207, 1.3790)

(b)

SE	0.4282
WI'	0.7950
95% CI	(0.0760, 1.6342)
PS	0.7200
95% CI	(-0.1691, 1.6100)

(c)

As we can see from Table IV, the WI's are 70.40%, 71.63% and 79.5% for the GVF, STD Model, and ML Models. The WI' for the three cases may be explained by the following: Let an image be selected at random and segmented by a random segmentation algorithm. If the image was also segmented by the reference algorithm, 0, the second segmentation would agree with the first algorithm at 70.40%, 71.63% and 79.5% (GVF, STD, and ML) of the rate that would be obtained by a second randomly selected reference algorithm. The upper limit of the 95% WI' CIs for (a), (b) and (c) are all greater than one (1.0), indicating that all three CGBs agree as much with the two EOBs as the two EOBs agree with each other.

C. Comparison of Algorithms

The ML, STD, and GVF Models were all tested on, N = 50 malignant DDSM images. The results were evaluated against expert traces from Radiologist #1 and Radiologist #2. The CGBs from the different algorithms were compared to the aforementioned EOBs, resulting in an error measurement for each algorithm and for each image. The algorithm which resulted in the smallest overall error was chosen as the "preferred algorithm" of segmentation. The results are displayed in Table V.

Table V. Friedman's two-way ANOVA results

SOURCE	RADIOLOGIST #1	RADIOLOGIST #2	OVERALL
R1 (GVF)	105	113	109
R2 (STD)	103	101	102
R3 (ML)	92	86	89
k	3	3	3
N	50	50	50
Q[r]	1.96	7.32	4.12
p-value	0.3753	0.0257	0.1275

To determine whether the difference in algorithm rank is a result of random chance, we compare the Friedman Statistic, Q[r], to the chi-square distribution of two degrees of freedom to. If the null hypothesis is rejected, i.e., a difference in the ranks is significant, the different algorithms are compared by the multiple comparison procedure described in the previous chapter. Analyzing the fourth column, we see that Q[r] is 4.12. In examining the chi-square distribution for k – 1 = 2 degrees of freedom (k = 3 algorithms), we find that equals 5.9915. The value of Q[r] is accepted if it is less than its corresponding value of the chi-square distribution. Therefore, the null hypothesis is accepted and there is no difference in the three segmentation algorithms which were used to segment the mammography images in question.

The p-value is another way to view the difference in the algorithms. The calculated p-value is 0.1275, and the declared value is 0.05. The calculated p-value exceeds the declared p-value, and accordingly there is no significant difference in the results generated by the three segmentation algorithms in between each other. By the design of this test, it can be concluded that differences in the mean rankings of algorithms are attributed to chance.

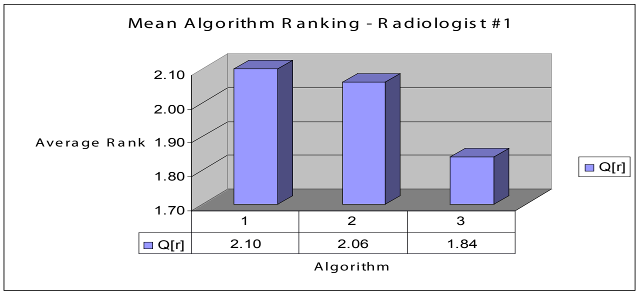


Figure 12. Mean algorithm ranking – Radiologist #1

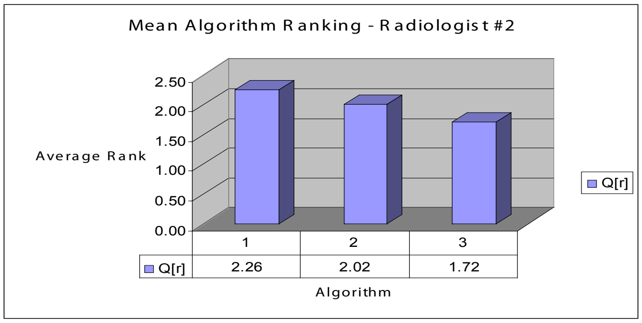


Figure 13. Mean algorithm ranking – Radiologist #2

Figures 12 and 13 show the results from the two-way ANOVA test. It is shown that algorithm 3 (the ML Model) was the preferred algorithm.

IV. CONCLUSION

We have presented a comprehensive validation analysis to evaluate the performance of three existing digital mammography segmentation algorithms against manual segmentation results produced by two expert radiologists. The Region Growing Combined with the Maximum Likelihood (ML) Model yielded not only the best accuracy, specificity, percent error, and algorithm ranking but also the greatest ratio of average computer-to-observer agreement and average inter-observer agreement (WI'). The upper limit of the 95% Confidence Interval (CI) was greater than 1.0, thereby making each individual observer (radiological tracer) a reliable member of the group.

Our validation model is effective and can be used as a reliable gold standard to test the robustness of digital mammography segmentation algorithms. We were able to evaluate CGBs against CGBs, EOBs to EOBs, and finally CGBs to EOBs. The WI' and PS statistics were consistent with each other, which means they are great complementary indicators of statistical significance. Friedman's two-way ANOVA test by ranks showed us that the difference in algorithm performance was attributed to chance, because the "Q-value" (Friedman Statistic, Q[r]) was less than the corresponding chi-square distribution value of p = 0.05 and two degrees of freedom.

The ML Model performed the best overall among the three algorithms evaluated in this study and is the "preferred algorithm" for the segmentation of digital mammography images. This is attributed to its design specifications being geared toward mammography images (region growing and intensity features), as the other models are used more in medical ultrasound for imaging the heart and abdomen. Results of the STD and GVF models in the previously mentioned areas have yielded promising results; however, they did not prove to be the most effective means of segmenting our malignant mammography images. The ML Model yielded an overall percent error of 2.34%, followed by the GVF Model (6.75%), and the STD Model (8.08%).

V. FURTHER RESEARCH

We would like to further this research by testing the robustness of the existing algorithms by acquiring a larger data set (number of images) and more radiological traces to use as ground truth. Having two radiological tracers, or two sets of ground truth data, has allowed us to only "touch the surface" in validating such segmentation algorithms presented in this paper. Additional expert radiological tracers will provide us with a larger inter-observer range, which will give us more dynamic opinions on how certain images should be segmented. Increasing the number of images in our data set will also increase the sensitivity of our measurements. A larger sample will help normalize our results and help us to better analyze the data qualitatively.

We would like to use the evaluation procedure for the ML Model on all three images generated from the Maximum Likelihood code, which gives us the original image, an image (contour) after one steep change, and then, finally, a contour after two steep changes of the defined probability cost function. We get different contour information for each image because of the inverse relationship of the threshold and size of the image. This will help verify the conclusion in [2]-[9] which states that the images which were a result of one steep change, yield the best segmented contours.

We would like to apply different parameters to generate the contour of the images for the STD and GVF Models. Those parameters include, but are not limited to, the initial radius of the snake, varying alpha and beta (the snake's tension and rigidity), the values of sigma (the blurring effect), and different values which will give us different edge-map data for accurate segmentation.

Finally, we wish to apply the results of this work to the design of a framework for a new digital mammography content-based image retrieval system (DMCBIR). A robust DMCBIR system is needed to accommodate the need of archiving the growing number of images being produced by radiology departments all over the world. Our results system would be to give a radiologist (or pathologist) the ability to find and retrieve multiple images, from a query image, that are similar in feature. It is necessary to have such a system available to help augment and aid in the diagnosis decisions of patients by a practicing medical doctor/physician. The application of the three segmentation algorithms used in this study will allow us to find the values of the similarity measures (SM) for the so-called DMCBIR System.

VI. APPENDIX A - GALLERY OF SEGMENTATION RESULTS

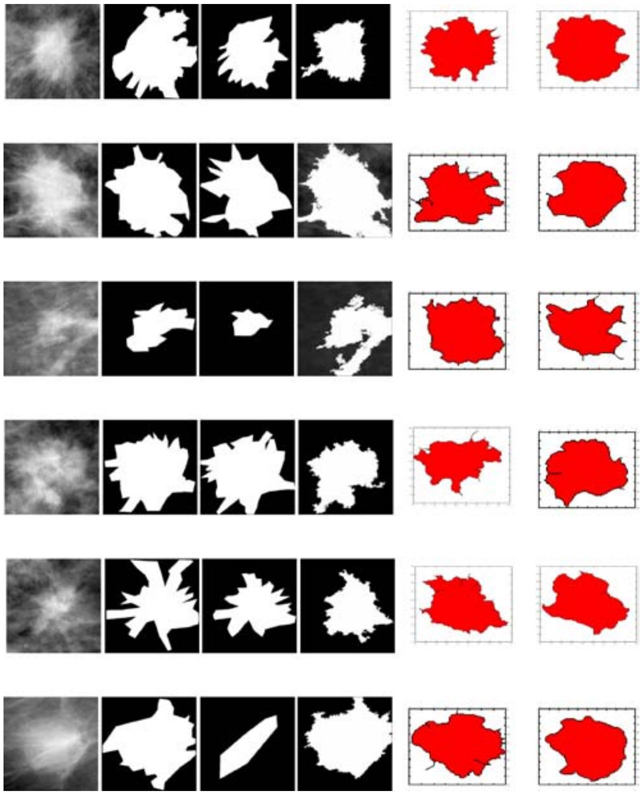


Figure A. Simulation results for all algorithms, Original mammogram, Radiologist #1 trace, Radiologist #2 trace, Maximum likelihood segmentation, Standard Model segmentation, and GVF Snake Model segmentation

VII. ACKNOWLEDGMENTS

The authors would like to thank Dr. Lisa Kinnard of the Georgetown University Medical Center ISIS Center for providing the mammography images, ground truth data (radiological traces), and the ML algorithm source code.

VIII. REFERENCES

[1] Chalana, V. and Kim, Y., "A methodology for evaluation of boundary algorithms on medical images," *IEEE Transactions on Medical Imaging,* vol. 16(5), pp. 642-52, 1997.

[2] Kinnard, L. M., "Segmentation of Malignant and Benign Masses in Digital Mammograms Using Region Growing Combined with Maximum-Likelihood Modeling," Doctoral Dissertation, Howard University, 2003.

[3] Xu, C., Prince, J. L., "Snakes, shapes, and gradient vector flow," *IEEE Trans. Image Processing,* vol. 7, no. 3, pp. 359-369, Mar. 1998.

[4] Kass, M., Witkin, A., and Terzopoulos, D., "Snakes: active contour models," *International Journal on Computer Vision,* vol. 1, no. 4, pp. 321-331, 1987.

[5] Stalib, L. H., and Duncan, J. S., "Left ventricular analysis for cardiac images using deformable Models," *IEEE Computer in Cardiology Magazine,* pp. 427-430, 1989.

[6] Wolfe, E. R., Delp, E. J., Meyer, C. R., Bookstein, F. L., and Buda, A. J., "Accuracy of automatically determined borders in digital two-dimensional echocardiography using a cardiac phantom," *IEEE Trans. Medical Imaging,* vol. MI-6, pp. 292-296, 1987.

[7] Detmer, P. R., Bashein, G., and Martin, R. W., "Matched filter identification of left-ventricular endocardial borders in transesophageal echocardiograms," *IEEE Trans. Medical Imaging,* vol. 9, pp. 396-404, 1990.

[8] Geiser, E. A., Conetta, D. A., Limacher, M. C., Stockton, V. O., Olivier, L. H., and Jones, B., "A second-generation computer-based edge detection algorithm for short-axis two-dimensional echocardiographic images: Accuracy and improvement in interobserver variability," *J. Amer. Soc. Echocardiol.,* vol. 3, pp. 79-90, 1990.

[9] Kinnard, L., Lo, S-C, Makariou, E., Osicka, T., Wang, P., Chouikha, M., Freedman, M. T., "Steepest Changes of a probability-based cost function for delineation of mammographic masses: A validation study," vol. 31, Issue 10, 2796-2810, October 2004.

[10] University of South Florida Digital Database for Screening Mammography. Available: http://marathon.csee.usf.edu/Mammography/Database.html

[11] Byrd, K., Zeng, J., Chouikha, M., "Performance assessment of mammography image segmentation algorithms," AIPR, pp. 152-157, 34th Applied Imagery and Pattern Recognition Workshop, Apr. 2005.

[12] Warfield, S. K., Zou, K. H., Wells, W. M. III, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," IEEE Transactions on Medical Imaging. vol. 23, no. 7, July 2004.

KENNETH BYRD is currently a Ph.D. student in the Department of Electrical and Computer Engineering at Howard University in Washington, DC. His research interests include: medical imaging, neural networks, cancer studies and biomedical engineering applications. He received a Master's Degree in Electrical Engineering from Howard University in 2005 and a Bachelor of Engineering Degree in Computer Engineering from the University of Delaware in 2003.

JIANCHAO ZENG, PH.D. has been with the Department of Electrical and Computer Engineering of Howard University for the past three years. He was previously with the Radiology Department of Georgetown University Medical Center. His research interests include image processing, medical imaging, image analysis, robotics, and their applications. His recent focus is on image segmentation, sensor fusion, and human detection. Dr. Zeng has published more than 100 papers in peer-reviewed journals and at international conferences.

MOHAMED CHOUIKHA, PH.D. is Chair of the Department of Electrical and Computer Engineering at Howard University. His research interests include Image and Signal Processing, Estimation Theory & Detection, Communications, and High Performance Computing. He has numerous scholarly publications and awards and has served as Principal Investigator on several substantial research and development projects including the Army High Performance Computing Research Center and the Next Generation Cancer Diagnosis Imaging Training Program. Mohamed received a Ph.D. in Electrical Engineering from the University of Colorado in Boulder.