Title

Relationship between computer segmentation performance and computer classification performance in breast CT: A simulation study using RGI segmentation and LDA classification

Permalink
https://escholarship.org/uc/item/80c1z46f

Journal
Medical Physics, 45(8)

ISSN
0094-2405

Authors
Lee, Juhun
Nishikawa, Robert M
Reiser, Ingrid
et al.

Publication Date
2018-08-01

DOI
10.1002/mp.13054

Peer reviewed

Relationship between computer segmentation performance and computer classification performance in breast CT: a simulation study using RGI segmentation and LDA classification

Juhun Lee[1], **Robert M. Nishikawa**[1], **Ingrid Reiser**[2], **John M. Boone**[3]

[1]Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[2]Department of Radiology, The University of Chicago, Chicago, Illinois, USA

[3]Department of Radiology, University of California Davis Medical Center, Sacramento, California, USA

Abstract

Purpose: Many computer aided diagnosis (CADx) tools for breast cancer begin by fully or semi automatically segmenting a given breast lesion and then classifying the lesion's likelihood of malignancy using quantitative features extracted from the image. It is often assumed that better segmentation will result in better classification. However, this has not been thoroughly evaluated. The purpose of this study is to evaluate the relationship between computer segmentation performance and computer classification performance.

Method: We used 85 breast lesions (32 benign, 56 malignant) from breast computed tomography (CT) cases of 82 women. We prepared one smooth and one sharp iterative image reconstructions (IIR) and a clinical reconstruction for each of the 82 breast CT scans. For each reconstruction, we created 15 segmentation outcomes by applying 15 different segmentation algorithms. Specifically, we simulated 15 segmentation algorithms by changing parameters in a single segmentation algorithm. We then created 15 classification outcomes by conducting quantitative image feature analysis on the segmented image results. Using a 10 fold cross-validation, we evaluated the relationship between segmentation and classification performances.

Result: We found a low positive correlation between segmentation and classification performances for the smooth IIR (median Pearson's rho = 0.18), while a moderate positive correlation (median Pearson's rho = 0.4 – 0.43) was found between the two performances for the sharp IIR and clinical reconstruction. However, we found large variations in both segmentation and classification performances for the sharp IIR and clinical reconstruction. There were cases where segmentation algorithms resulted in similar segmentation performances, but the corresponding classification performances were different. These results indicate that an improvement in segmentation performance does not guarantee an improvement in the corresponding classification performance.

Corresponding Author: Juhun Lee, Department of Radiology, University of Pittsburgh, 3362 Fifth Ave., Pittsburgh PA 15213, Office: 1-412-641-2365, FAX: 1-412-641-2582.

Conclusion: Computer segmentation is an indirect variable affecting the computer classification. As better segmentation does not guarantee better classification, we should report both segmentation and classification performances when comparing segmentation algorithms.

Keywords

Breast CT; Computer segmentation; Computer classification; computer-aided diagnosis

1 Introduction

Many computer-aided diagnosis (CADx) procedures for breast cancer include lesion localization, lesion segmentation, feature extraction from the lesion, and lesion classification via training a classifier using the extracted features. The common belief in the above procedures is that an improvement in each step will improve the CADx algorithm's performance. The first step, i.e., lesion detection, is the least related to the CADx algorithm's performance, as correct lesion locations are typically provided to the CADx algorithm. The last two steps, i.e., feature extraction and lesion classification, are easy to validate; one can conduct receiver operating characteristic (ROC) curve analysis to check if the improvement in feature extraction (e.g., finding new/better features) and classification (e.g., finding new/better classifiers) lead to improving CADx. However, whether an improvement in lesion segmentation will result in better performance of the CADx algorithm has not been thoroughly studied.

Many previous studies on developing computer segmentation algorithms (to name a few[1–4]) of lesions limited their reporting to the segmentation performance of their algorithms and comparing them to other state of the art algorithms. The core assumption of these previous studies was that better segmentation will lead to better computer lesion classification (i.e., lesion diagnosis). They often considered whether their lesion segmentation algorithm will lead to better lesion diagnosis as out of the scope of their study. However, it is possible that the improvement in segmentation may not improve lesion diagnosis. Thus, we argue that research on computer segmentation algorithms of lesions for CADx tools should report if their algorithms improve classification performance.

Two other studies[5, 6] showed both segmentation and classification performances. For example, Kuo et al.[5] introduced a new segmentation algorithm for dedicated breast computed tomography (CT). They showed that their segmentation algorithm resulted in better segmentation, as well as better classification, than the segmentation outcomes and corresponding classification outcomes of an existing segmentation algorithm for breast CT[7, 8], using features extracted from a segmented breast lesion. From these results, however, it is difficult to generalize that improved segmentation leads to improved lesion classification.

This study examined how computer segmentation performance is related to computer classification performance. Specifically, we tried to answer the following question: does better segmentation performance lead to better classification performance in dedicated breast CT? To do so, we needed breast CT cases with different segmentation outcomes. We simulated 15 different segmentation algorithms by changing internal parameters in a single segmentation algorithm. We applied 15 segmentation algorithms on breast lesions to create 15 segmentation outcomes per each lesion. We then trained and tested a classifier to create 15 classification outcomes per each lesion. Using 15 segmentation and classification outcomes, we checked how these changes in computer segmentation performance affected computer classification performance. In addition, we repeated the analysis for three different reconstructions to examine how the relationship between computer segmentation and classification differs for different image qualities.

2 Methods

2.1 Dataset

For this study, we used a total of 88 biopsy proven breast lesions (56 malignant and 32 benign lesions) in 82 non-contrast breast CT images of women aged 18 or older. Under an institutional review board (IRB) approved protocol, the prototype dedicated breast CT system at the University of California at Davis[9] was used to acquire all breast CT images of recruited women. We describe the details of the cases elsewhere[10].

2.2 Image reconstructions

We considered three reconstructions in this study using an iterative image reconstruction (IIR) algorithm[11] and a clinical reconstruction algorithm (Feldkamp-Davis-Kress (FDK) reconstruction[12]). The IIR algorithm consists of two sub reconstruction algorithms[11]. The first sub reconstruction algorithm provides smooth gray-scale information, the second sub algorithm adds sharp edge information. By combining these two sub reconstruction results with different weights, one can create breast CT cases with different image qualities or appearances. We prepared one smooth and one sharp reconstruction by changing parameters in the IIR algorithm. Specifically, we placed zero weight for the second sub algorithm to create the smooth IIR reconstruction. For the sharp IIR, we gave a three times higher weight to the second sub algorithm compared to the first sub algorithm. Figure 1 shows the image qualities, or appearances, of the three reconstructions considered in this study.

2.3 Computer segmentation of breast lesions

We first selected a cubic shaped volume of interest (VOI) with sides of 35 mm, where its center is located at a lesion center. These VOIs in IIR were isometric, while those in FDK were not isometric, since they had a different slice thickness in the z-direction. Then, we utilized a semi-automated segmentation algorithm[7, 8] , called RGI segmentation, to segment breast lesions of all reconstructions. The segmentation algorithm was originally developed for mammograms[7] and extended for tomosynthesis and bCT[8]. The algorithm is semi-automatic and requires a seed point (i.e., lesion center) only to automatically segment a given lesion volume. A research specialist, with over 15 years of experience in mammography, provided the seed point for the algorithm. As segmenting a 3D lesion is time-consuming and laborious, the research specialist manually segmented three cross-sectional views of the lesion at its center (i.e., coronal, transverse, sagittal views). This was done only for the FDK reconstruction. We then applied this segmentation to the IIR reconstructed images, as all the reconstructed images are co-registered by default.

The RGI segmentation algorithm regulates the search boundary for a lesion by applying predefined weights on the bCT image. One can change the shape and type of weights to change the segmentation outcomes and therefore simulate different segmentation algorithms. For this study, we used seven 3D normal Gaussian weights with standard deviations (SD) of 4 to 10 mm with 1 mm increments, and seven 3D cube shapes with widths of 4 to 10 mm with 1 mm increments, where the weight gets higher in the center and gradually fades out to the edge, and one cone shape weight, where its weight is the highest at the center and gradually fades out to the end of the VOI (Table 1). Note that we did not consider weights with SDs or widths less than 4 mm, as they are too small to segment large lesions. Likewise, we did not consider weights with SDs or widths larger than 10 mm, as they can be too large to effectively capture small lesions. Figure 2 illustrates the coronal view at the center for a few selected weight types. Note that the example in Figure 2 contains normal breast parenchyma located outside of the green lesion border outlined by the expert. In addition, it should be noted that Figure 2 illustrates the range of segmentation qualities by changing the weights applied to the algorithm. Specifically, the weights shown in Figure 2, except the cone weight, are at two extreme points of the range of widths and standard deviations that we considered, which was [4mm, 10mm]. Since the given lesion is large, some weights, especially cone weights and Gaussian weights, with a 10 mm standard deviation provided good segmentation outcomes, while some, especially square and Gaussian weights, with a 4 mm width or standard deviation did not.

We used the DICE coefficient[13] to evaluate segmentation results by comparing the algorithm’s output to the lesion border manually drawn by the research specialist. We averaged DICE values from cross-sectional views of each case. Then, we computed the mean of those averaged (cross-sectional view) DICE values of cases as the measure of the segmentation performance of each segmentation algorithm.

2.4 Computer classification of breast lesions

Computer classification requires a choice of a statistical classifier and input features. We considered a total of 23 quantitative image features from the segmentation results (Table 2, adopted from[10]). Previous studies utilized these features for lesion detection and classification[5, 10, 14–19].

The 23 quantitative image features included four histogram (feature #1-#4), seven shape (feature #5-#11), five margin (feature #12-#16), four texture (feature #17-#20), and three surface curvature descriptors (feature #21-#23). Histogram descriptors characterize the gray-scale information within the lesion and its relationship to the surrounding background. Shape and margin descriptors represent the morphological information of the segmented lesions. Texture descriptors are a 3D version of the 2D gray-level co-occurrence matrix[20], which quantify the characteristics of the segmented lesion texture. Curvature descriptors summarize the local lesion surface variations of the segmented lesion. All features were obtained from the 3D volumes.

Under a 10 fold cross-validation, we selected the best features (via *sequentialfs* function in MATLAB) and trained an LDA classifier using the selected features and a training set, and evaluated their classification performances on 15 segmentation outcomes for each of the three reconstructions on a held out testing set.

Using the area under the receiver operating characteristic curve (AUC) of the resulting final classifiers, we compared the relationship between the classification and segmentation performances. For each test set in the 10 fold cross-validation, we computed the AUC value and the averaged DICE value (averaged over the lesions in the test set). Then, we conducted a correlation analysis (Pearson correlation coefficients) between the AUC value and the averaged DICE value for 15 segmentation outcomes, and repeated 10 times for each test set.

3 Results

The segmentation performances in terms of the averaged DICE coefficient under the 10 fold cross-validation of 15 different segmentation algorithms for smooth IIR, sharp IIR, and FDK reconstruction cases ranged from [0.63, 0.73], [0.63, 0.75], and [0.66, 0.82], respectively. Corresponding classification performances in terms of the averaged AUC under the 10 fold cross-validation ranged from [0.64, 0.81], [0.61, 0.88], and [0.66, 0.81], respectively.

Figure 3 shows the feature selection frequency under the 10 fold cross-validation for each segmentation outcome and each reconstruction. Feature #13, radial gradient index, was consistently selected by all three reconstructions. Although there were variations in selection frequency among segmentations, the feature sets of [#8, #10, #21], [#10, #21], [#10, #17, #21] were also frequently selected by smooth IIR, sharp IIR, and FDK reconstructions, respectively.

Figure 4 shows the correlation coefficient values between the segmentation and classification performances under the 10 fold cross-validation. Note that each correlation coefficient value was based on 15 AUC and averaged DICE data point pairs per each held-out cross-validation portion. The median correlation coefficient values for the IIR smooth, the IIR sharp, and the FDK reconstruction under the 10 fold cross-validation were 0.18 with median absolute deviation (MAD) of 0.24 and (min, max) = [−0.53, 0.69], 0.43 with MAD of 0.22 and (min, max) = [−0.18, 0.57], and 0.4 with MAD of 0.33 and (min, max) = [−0.56, 0.74], respectively. The median correlation coefficient values show that there was a low correlation (according to Cohen[21] as the correlation coefficient value < 0.3) between the segmentation performance and the classification performance for the smooth IIR reconstruction, while moderate correlations (according to Cohen[21] as the correlation coefficient values were between 0.3 and 0.5) were found for the sharp IIR and FDK reconstructions. Thus, one can generally expect that better segmentation results in better classification for sharper reconstructions, (i.e., sharp IIR and FDK reconstruction), than for smoother reconstructions, (i.e., smooth IIR reconstruction).

However, we observed the large variability in both DICE and AUC values for sharp reconstructions (i.e., sharp IIR and FDK reconstructions). Figure 5 shows the scatter plot of the AUC and DICE coefficient value pairs averaged over 10 held-out cross-validation sets for each reconstruction. Two segmentation algorithms for the sharp IIR reconstruction, marked as a filled square in Figure 5.B, resulted in similar segmentation performances, but their corresponding classification performances are statistically different to each other (Table 3). Similarly, another two segmentation algorithms for the FDK reconstruction cases, marked as a triangle in Figure 5.C, resulted in statistically different segmentation performances (Table 3), but their corresponding classification performances were similar to each other.

These results show that segmentation outcomes and classification outcomes are positively correlated, but there is large variability between the two performances. Thus, we cannot reliably predict that an improvement in segmentation performance will always result in an improvement in classification performance.

4 Discussion

Segmentation outcome may be one of many indirect variables that can affect the classification performance of CADx tools. Changes in DICE values indicate changes in segmentation outcomes. In this study, it was observed that changes in segmentation outcomes resulted in changes in quantitative image features in classifiers such that their effects dominate changes in classification performance. In fact, most frequenly selected features quantify the morphological information of the lesion; two shape (features #8, #10), one margin (feature #13), and one surface curvature (feature #21) descriptors, except one texture (feature #17) for the FDK reconstruction. Thus, changes in weight for the RGI segmentation algorithm resulted in the variations in segmentation and these variations resulted in changes in classification outcomes.

Our findings suggest that any lesion segmentation algorithm should be evaluated not only for its segmentation performance, but also for the resulting classification performance from features extracted from the segmented outcomes. This is because a wide range of classification outcomes can be obtained from similar segmentation outcomes, and the similar segmentation outcomes can yield statistically different classification performance. This confirms the conclusion of the previous study by Kuo et al.[5]; the DICE coefficient, as well as other similar measures, such as overlap ratio, is not a sufficient evaluation metric for segmentation algorithms for bCT images.

A possible limitation of the study is that segmentation performances are based on manual segmentations made by a single research specialist. A future study with additional manual segmentations is required in order to test if our finding will hold on the updated segmentation performances by the new manual segmentations.

5 Conclusion

In this study, we found that computer segmentation is an indirect variable affecting the computer classification. As better segmentation does not guarantee better classification, it is important to report both segmentation and classification performances for any segmentation algorithms to properly show the usefulness of the algorithms.

Acknowledgement

The author would like to acknowledge Alexandra Edwards for establishing ground truth for breast lesions in breast CT images. This work has been supported in part by grants from the National Institutes of Health R21-EB015053 and R01-CA181081. Drs. Lee, Nishikawa, and Reiser have nothing to declare. Dr. Boone has a research contract with Siemens Medical Systems, and receives royalties from Lippincott Williams and Wilkins (book).

7 Reference

1. Liu H, Liu Y, Zhao Z, Zhang L, and Qiu T, A new background distribution-based active contour model for three-dimensional lesion segmentation in breast DCE-MRI, Med Phys 41(8), 082303 (2014). [PubMed: 25086552]
2. Pons G, Martí J, Martí R, Ganau S, and Noble JA, Breast-lesion Segmentation Combining B-Mode and Elastography Ultrasound, Ultrason Imaging 38(3), 209–224 (2016). [PubMed: 26062760]
3. Rahmati P, Adler A, and Hamarneh G, Mammography segmentation with maximum likelihood active contours, Med Image Anal 16(6), 1167–1186 (2012). [PubMed: 22831774]
4. Tao Y, Lo S-CB, Freedman MT, Makariou E, and Xuan J, Multilevel learning-based segmentation of ill-defined and spiculated masses in mammograms, Med Phys 37(11), 5993–6002 (2010). [PubMed: 21158311]
5. Kuo H-C, Giger ML, Reiser I, et al., Impact of lesion segmentation metrics on computer-aided diagnosis/detection in breast computed tomography, J. Med. Imag 1(3), 031012–031012 (2014).
6. Marcomini KD, Carneiro AAO, and Schiabel H, Application of Artificial Neural Network Models in Segmentation and Classification of Nodules in Breast Ultrasound Digital Images, Int J Biomed Imaging 2016, 7987212 (2016). [PubMed: 27413361]
7. Kupinski MA and Giger ML, Automated seeded lesion segmentation on digital mammograms, Medical Imaging, IEEE Transactions on 17(4), 510–517 (1998).
8. Reiser I, Joseph SP, Nishikawa RM, et al., Evaluation of a 3D lesion segmentation algorithm on DBT and breast CT images, in Proc. SPIE 7624, Medical Imaging 2010: Computer-Aided Diagnosis(2010), pp. 76242N-76242N–7.
9. Lindfors KK, Boone JM, Newell MS, and D'Orsi CJ, Dedicated breast computed tomography: the optimal cross-sectional imaging solution?, Radiol. Clin. North Am 48(5), 1043–1054 (2010). [PubMed: 20868899]
10. Lee J, Nishikawa RM, Reiser I, and Boone JM, Optimal reconstruction and quantitative image features for Computer-Aided Diagnosis tools for breast CT, Med. Phys 44(5), 1846–1856 (2017). [PubMed: 28295405]
11. Antropova N, Sanchez A, Reiser I, Sidky EY, Boone JM, and Pan X, Efficient iterative image reconstruction algorithm for dedicated breast CT, in Proc. SPIE 9783, Medical Imaging 2016: Physics of Medical Imaging(2016).
12. Feldkamp LA, Davis LC, and Kress JW, Practical cone-beam algorithm, J. Opt. Soc. Am. A 1(6), 612–619 (1984).
13. Dice LR, Measures of the Amount of Ecologic Association Between Species, Ecology 26(3), 297–302 (1945).
14. Ray S, Prionas ND, Lindfors KK, and Boone JM, Analysis of breast CT lesions using computer-aided diagnosis: an application of neural networks on extracted morphologic and texture features, in Proc. SPIE 8315, Medical Imaging 2012: Computer-Aided Diagnosis(2012), pp. 83152E-83152E–6.
15. Kuo H-C, Giger ML, Reiser I, et al., Development of a New 3D Spiculation Feature for Enhancing Computerized Classification on Dedicated Breast CT, in Radiological Society of North America 2014 Scientific Assembly and Annual Meeting(Chicago IL, n.d.).
16. Wang X, Nagarajan MB, Conover D, Ning R, O'Connell A, and Wismueller A, Investigating the use of texture features for analysis of breast lesions on contrast-enhanced cone beam CT, in Proc. SPIE 9038, Medical Imaging 2014: Biomedical Applications in Molecular, Structural, and Functional Imaging(2014), pp. 903822–903822–8.

17. Lee J, Nishikawa RM, Reiser I, Boone JM, and Lindfors KK, Local curvature analysis for classifying breast tumors: Preliminary analysis in dedicated breast CT, Med. Phys 42(9), 5479–5489 (2015). [PubMed: 26328996]

18. Lee J, Nishikawa RM, Ingrid R, Margarita ZL, and John BM, Lack of agreement between radiologists: implications for image-based model observers, Journal of Medical Imaging 4(2), 025502 (2017). [PubMed: 28491908]

19. Lee J, Nishikawa RM, Reiser I, and Boone JM, Neutrosophic segmentation of breast lesions for dedicated breast computed tomography, JMI, JMIOBU 5(1), 014505 (2018). [PubMed: 29541650]

20. Chen W, Giger ML, Li H, Bick U, and Newstead GM, Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images, Magn Reson Med 58(3), 562–571 (2007). [PubMed: 17763361]

21. Cohen J, Statistical Power Analysis for the Behavioral Sciences (Routledge, 2013).

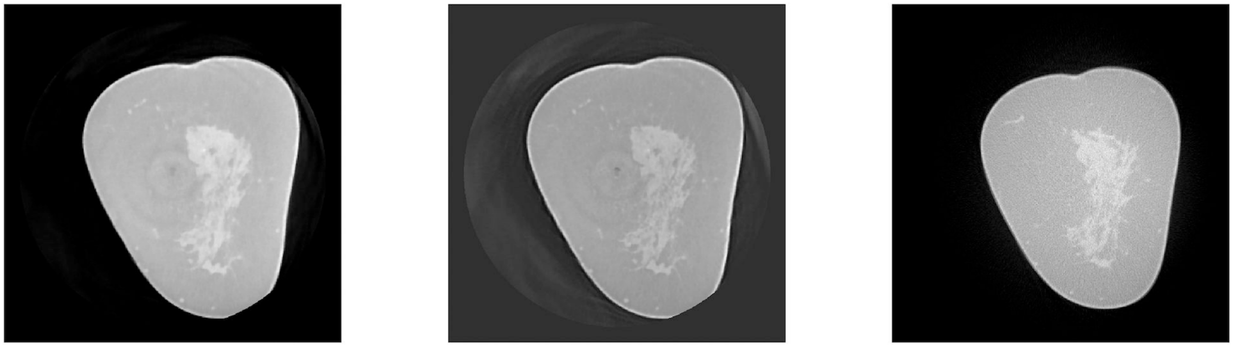

Figure 1.

This figure shows an example case with different image reconstructions in coronal view. First two columns represent the smooth and sharp reconstructions by the IIR algorithm, respectively. The last column is the case reconstructed by the FDK algorithm.

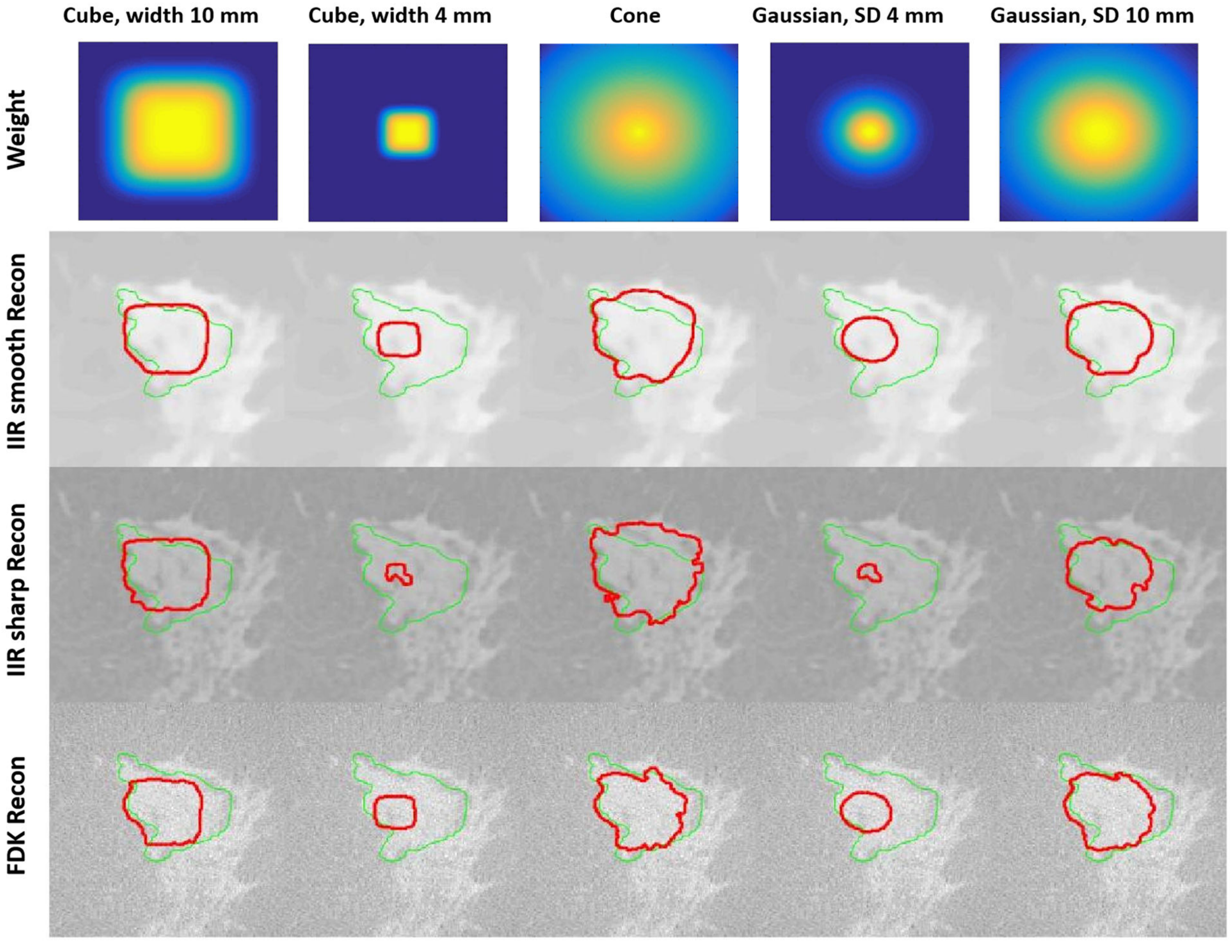

Figure 2.

The first row shows the some examples of various types and ranges of weight, which were applied to each breast CT image before applying the RGI segmentation algorithm. Images in last three rows represent the resulting segmentation outcomes in the coronal view at the lesion center, using the weights shown in the first row. Outlines in green represent the lesion outline drawn by a human expert. Note that the lesion is partially embedded in fibroglandular tissue. Outlines in red are the cross-sectional border from the segmented lesion volume by the RGI segmentation algorithms using the weights in the first row.

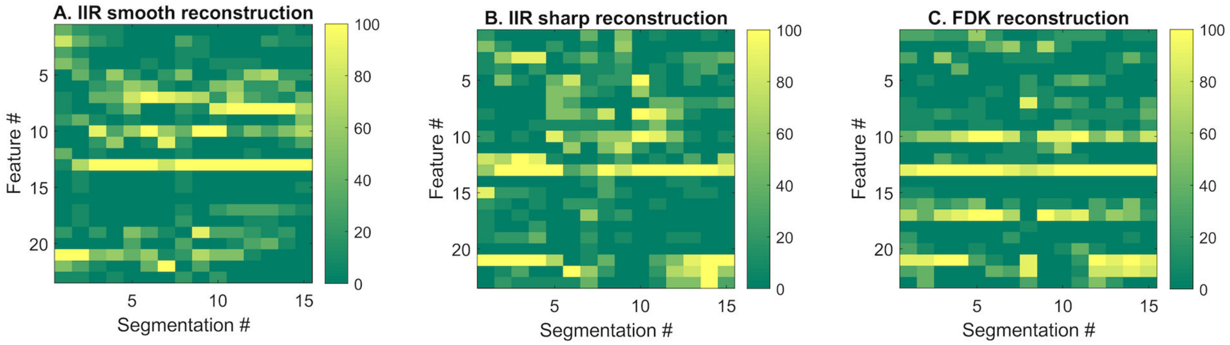

Figure 3.

This figure shows the feature selection frequency for each segmentation outcomes in the three reconstructions under 10 fold cross-validation. Feature #13 was frequently selected by most segmentations in all reconstructions. Although there are some variations in selection frequency among segmentations, feature sets of [#8, #10, #21], [#10, #21], and [#10, #17, #21] were frequently selected in addition to feature #13 by IIR smooth, IIR sharp, and FDK reconstructions, respectively.

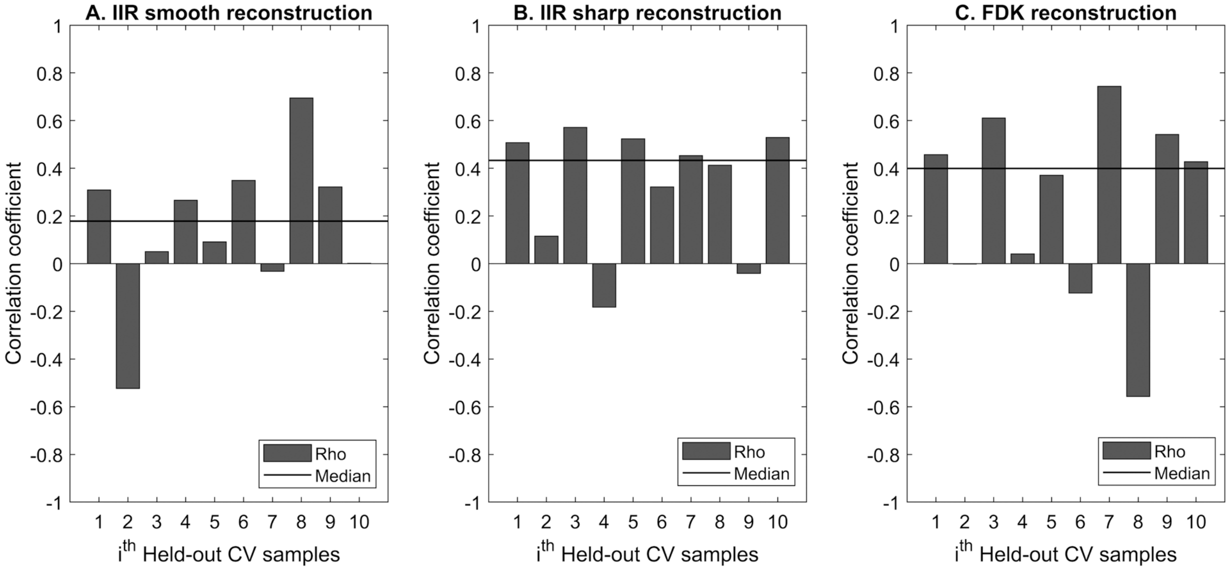

Figure 4.

This figure shows the correlation coefficient values between the segmentation performance (averaged DICE) and the classification performance (AUC) 10 fold cross-validation for each reconstruction. There were 15 segmentation and classification data point pair for each 10 held-out cross validation set. The segmentation and classification performances for all reconstructions were positively correlated, as their median correlation coefficient values were 0.18, 0.43, and 0.4 for the IIR smooth, the IIR sharp, and the FDK reconstruction, respectively. However, we found considerable variations in correlation coefficient values among 10 held-out sets; their median absolute deviation values were 0.24, 0.21, and 0.33, for the IIR smooth, the IIR sharp, and the FDK reconstruction, respectively.

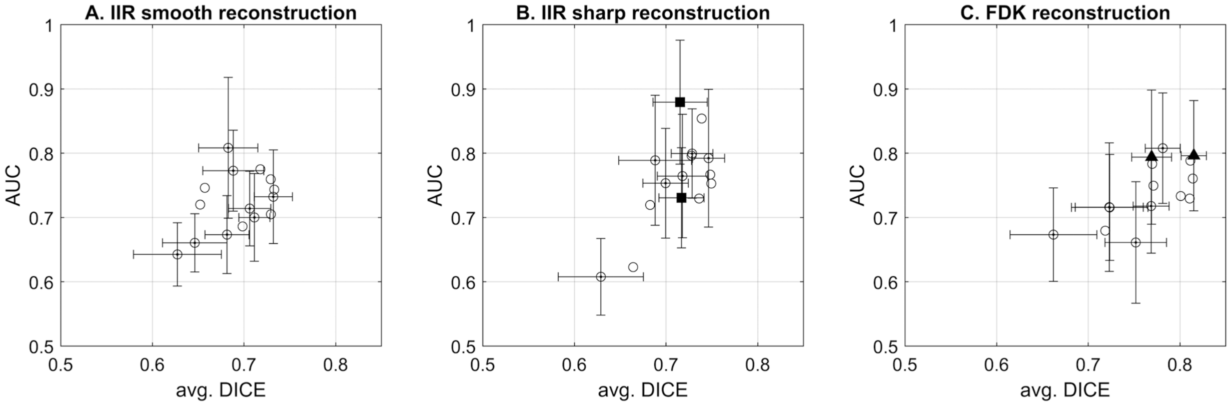

Figure 5.

Scatter plots for the segmentation and corresponding classification for the LDA classifiers under the 10 fold cross-validation. 95% confidence intervals of a few selective segmentation algorithms are shown. The list of selected segmentation algorithms are the RGI segmentation algorithm with 3D cube weight with 4 – 7, 9 – 10 mm width and 3D Gaussian weight with 5 and 9 mm width. Segmentation algorithms with narrow weight width, e.g., less than 7 mm, resulted in lower segmentation performance than wider weight width, e.g., equal to or larger than 7 mm. We compared a few selected segmentations, with square and triangle markers, in Table 3.

Table 1.

List of predefined weights for the sementation algorithm

Segmentation number	Weights
1	3D Cube shape with width of 10 mm
2	3D Cube shape with width of 9 mm
3	3D Cube shape with width of 8 mm
4	3D Cube shape with width of 7 mm
5	3D Cube shape with width of 6 mm
6	3D Cube shape with width of 5 mm
7	3D Cube shape with width of 4 mm
8	Cone shape weights
9	3D normal Gaussian with SD of 4 mm
10	3D normal Gaussian with SD of 5 mm
11	3D normal Gaussian with SD of 6 mm
12	3D normal Gaussian with SD of 7 mm
13	3D normal Gaussian with SD of 8 mm
14	3D normal Gaussian with SD of 9 mm
15	3D normal Gaussian with SD of 10 mm

Table 2.

List of image features used in this study

Feature #	Feature Type	Definition*
	Histogram descriptors	
1	Average region gray value [HU]	μ (Gray value in V)
2	Region contrast [HU]	F1 - μ (Gray value outside of V)
3	Region gray value variation [HU]	σ (Gray value in V)
4	Margin gray value variation [HU]	σ (Gray value in M)
	Shape descriptors	
5	Irregularity	$2.2\ ^{*}V^{1/3} / M^{1/2}$
6	Compactness	% of volume of V included in SP
7	Ellipsoid axes min-to-max ratio	Min to max ratio of semi-axes of the ellipsoid fitted to V
8	Margin distance variation [mm]	σ (distances from the center of V to the margin of V)
9	Relative margin distance variation	F8 / Mean(distances from the center of V to the margin of V)
10	Average gradient direction	μ (gradient direction of each voxel in M)
11	Margin volume [mm^3]	Σ (voxels in M)
	Margin descriptors	
12	Average radial gradient [HU]	μ (radial gradient of each voxel in M)
13	Radial gradient index (RGI)	F12 / μ (magnitude of image gradient of each voxel in M)
14	Margin strength 1	μ (magnitude of image gradient of each voxel in M) / F2
15	Margin strength 2	σ (magnitude of image gradient of each voxel in M) / F2
16	Radial gradient variation	σ (radial gradient of each voxel in M)
	Texture descriptors	
17	GLCM|Energy	3D version of 2D gray-level co-occurrence | Energy
18	GLCM|Contrast	3D version of 2D gray-level co-occurrence | Contrast
19	GLCM|Correlation	3D version of 2D gray-level co-occurrence | Correlation
20	GLCM|Homogeneity	3D version of 2D gray-level co-occurrence | Homogeneity
	Surface Curvature descriptors	
21	Total Curvature	$\mu\ (\lvert p_1\rvert + \lvert p_2\rvert$ over S) / $\sigma\ (\lvert p_1\rvert + \lvert p_2\rvert$ over S)
22	Mean Curvature	$\mu\ (0.5 \times (p_1 + p_2)$ over S) / $\sigma\ (0.5 \times (p_1 + p_2)$ over S)
23	Gaussian Curvature	$\mu\ (p_1 \times p_2$ over S) / $\sigma\ (p_1 \times p_2$ over S)

*V refers to the segmented lesion volume.M refers to the margin of the lesion volume. SP refers to the minimum sphere including V. S refers to the surface of V. p_1 and p_2 refer to the first and second principal component of S. μ and σ indicate mean and standard deviation.

Table 3.

Segmentation and classification performance comparison among selected segmentation algorithms for sharp reconstructions

Performance statistics for selected two segmentation algorithms					**Difference**	
Comparable segmentation performance but different classification performance (square marker in Figure 5.B)					**Figure of merit: AUC**	
Recon.	**Seg #**	**AUC [95% CI]**	**Seg #**	**AUC [95% CI]**	**$\text{AUC}_{\text{Diff.}}$ [95% CI]**	***p*-value**
IIR sharp	2	0.88 [0.78, 0.97]	10	0.73 [0.65, 0.81]	0.15 [0.07, 0.23]	0.0003[†]
Comparable classification performance but different segmentation performance (triangle marker in Figure 5.C)					**Figure of merit: DICE**	
Recon.	**Seg #**	**DICE [95% CI]**	**Seg #**	**DICE [95% CI]**	**$\text{DICE}_{\text{Diff.}}$ [95% CI]**	***p*-value**
FDK	14	0.82 [0.8, 0.83]	5	0.77 [0.75, 0.79]	0.05 [0.03, 0.07]	<0.0001[†]

[†] Statistically significant at the corrected significance level $\alpha = 0.05/2 = 0.025$.