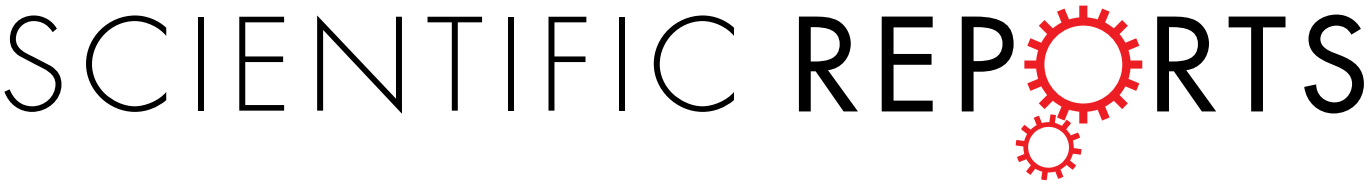

OPEN

Received: 22 August 2017
Accepted: 27 February 2018
Published online: 19 March 2018

Appearance Constrained Semi-Automatic Segmentation from DCE-MRI is Reproducible and Feasible for Breast Cancer Radiomics: A Feasibility Study

Harini Veeraraghavan[1], Brittany Z. Dashevsky[2,3], Natsuko Onishi[3], Meredith Sadinski[3], Elizabeth Morris[3], Joseph O. Deasy[1] & Elizabeth J. Sutton[3]

We present a segmentation approach that combines GrowCut (GC) with cancer-specific multi-parametric Gaussian Mixture Model (GCGMM) to produce accurate and reproducible segmentations. We evaluated GCGMM using a retrospectively collected 75 invasive ductal carcinoma with ERPR+ HER2− (n = 15), triple negative (TN) (n = 9), and ER-HER2+ (n = 57) cancers with variable presentation (mass and non-mass enhancement) and background parenchymal enhancement (mild and marked). Expert delineated manual contours were used to assess the segmentation performance using Dice coefficient (DSC), mean surface distance (mSD), Hausdorff distance, and volume ratio (VR). GCGMM segmentations were significantly more accurate than GrowCut (GC) and fuzzy c-means clustering (FCM). GCGMM's segmentations and the texture features computed from those segmentations were the most reproducible compared with manual delineations and other analyzed segmentation methods. Finally, random forest (RF) classifier trained with leave-one-out cross-validation using features extracted from GCGMM segmentation resulted in the best accuracy for ER-HER2+ vs. ERPR+/TN (GCGMM 0.95, expert 0.95, GC 0.90, FCM 0.92) and for ERPR + HER2− vs. TN (GCGMM 0.92, expert 0.91, GC 0.77, FCM 0.83).

Breast cancer is one of the most commonly diagnosed cancers in women and the second most common cause of cancer-related deaths[1]. Although the increasing availability of novel treatment options has helped to improve survival among patients, robust tools are critically needed to effectively monitor treatment response[2]. Miranikova *et al.*[3] have shown that tumour volumes measured on magnetic resonance imaging (MRI) predict treatment response in neoadjuvant settings. However, accurate and reproducible tumour segmentation is crucial for evaluating breast cancer response to treatments[4] and to improve surgical outcomes[5].

Accurate and reasonably fast segmentation is critical for radiomics analysis[6] which consists of extracting image features from large datasets with the purpose of identifying non-invasive image-based surrogates for diagnosis (differentiating disease aggressiveness) and for predicting treatment response. Radiomics analysis of breast cancers have been used for predicting cancer treatment outcomes[7–9] and for differentiating between breast cancers by molecular subtype[10–13] or for classifying cancers by their aggressiveness[14,15].

The first and crucial step in extracting the various texture measures is segmentation of the cancer. With the exception of[11,15], the vast majority of works have employed manual tumour segmentation for radiomics analysis due to the difficultly in ensuring accurate computer segmentations. However, manual delineation is time consuming. Therefore, majority of works[12–14] including ours[10,16] have used manual segmentation of one or a few representative slices. Recently, semi-automatic segmentations including GrowCut (GC)[17] have been reported to produce more reproducible texture features compared with features computed from manually delineated lung tumors[18], thereby, underscoring the importance and utility of computer-generated segmentations for high-throughput radiomics.

[1]Memorial Sloan Kettering Cancer Center, Medical Physics, New York, NY, 10065, USA. [2]Present address: University of Chicago, Radiology, Chicago, USA. [3]Memorial Sloan Kettering Cancer Center, Radiology, New York, NY, 10065, USA. Correspondence and requests for materials should be addressed to H.V. (email: veerarah@mskcc.org)

Interactive segmentation methods[19,20] model the user input to generate more accurate segmentations than fully automatic methods. Thus, the interactive GC method has been shown to produce reasonably accurate segmentations for brain gliomas[17] and more repeatable segmentations than expert users[21] for lung cancers. However, as an interactive method adapts its segmentation to user's inputs, it generates highly variable segmentations, thereby, introducing another source of variability for radiomics and longitudinal analysis of cancers. Previous works, which include[22–25] have incorporated machine learning to reduce segmentation variability. For example, Veeraraghavan and Miller[23] developed an active learning-based approach to improve the consistency of segmentation while reducing the number of required user interactions to generate reasonably accurate segmentations of brain cancers. However, repetitive interactions resulting either from the algorithm itself which present as queries or from users can become time consuming particularly for high-throughput radiomics analysis. This in turn limits the applicability of such methods for high-throughput analysis in comparison to fully automatic methods such as unsupervised fuzzy clustering[26].

We report an approach to improve the accuracy and reproducibility of interactive GC. Specifically, we developed an approach that combines the cancer-specific appearance modeling using multi-parametric Gaussian mixture models (GMM) with GC to constrain the GC segmentation, called GCGMM. Our approach eliminates the need for repetitive user interactions by generating a probabilistic segmentation. The user can select from among multiple segmentations by changing the segmentation probability (or confidence).

The goals of this study were to: (a) develop a reasonably accurate and reproducible approach to generate breast cancer segmentation with variable user inputs, and (b) to assess the feasibility of features extracted from computer-generated segmentation over manual delineation for radiomics-based classification of breast cancers. We compared the results of our approach with the GrowCut (GC) and fuzzy c-means (FCM) clustering[26]. FCM was chosen for benchmarking the performance of GCGMM as the former method has previously been used in radiomics analysis of breast cancers.

Results

We evaluated the reproducibility of manual delineations produced by multiple users using six consecutive cases with two from ER-HER2+, two from ERPR + HER2− and two from triple negative cancers to benchmark segmentation performance. All raters produced highly variable segmentations. The segmentation concordance measured using the various performance metrics was: Dice overlap coefficient (DSC) (0.78 ± 0.10), mean surface disance (mSD) ($1.23\,\text{mm} \pm 0.67\,\text{mm}$), 95% Hausdorff distance ($5.04\,\text{mm} \pm 5.9\,\text{mm}$), and volume ratio (VR) (0.16 ± 0.10).

GCGMM segmentations were significantly more accurate compared with other methods. Figure 1(a) shows segmentations produced using the grow-cut (GC), GCGMM, and FCM methods together with expert delineation for two different tumours. As shown, GCGMM segmentations closely corresponded to the expert delineation while the GC and FCM methods resulted in under- and over-segmentations, respectively. Overall, GCGMM produced significantly higher DSC; significantly smaller mSD, smaller HD95 and lower VR compared with other methods (Fig. 1(b), Table 1).

Only the GCGMM method achieved a better segmentation performance than the inter-rater segmentation concordance using all the performance measures. Furthermore, GCGMM segmentations were more accurate compared with GC and FCM methods for both mild and marked background parenchymal enhancements (Table 1), and for cancers that presented as masses. Finally, GCGMM produced more accurate segmentation of ER-HER2+ cancers compared with both FCM and GC ($P < 0.001$) using all performance metrics.

Fifty one percent of all tumours generated using GCGMM had volumes similar to expert delineation ($-0.1 <= VR < 0.1$) with 8% under- ($VR < -0.1$) and 41% over-segmented. In comparison, GC and FCM resulted in 11% and 14% close to expert delineation; 33% and 18% under-segmentations and 56% and 68% over-segmentations, respectively.

GCGMM produced reproducible segmentations. GCGMM resulted in the most reproducible segmentations (Table 2) using all the performance metrics, including segmented volumes. The precision errors computed using GCGMM segmentations were smaller for all the performance metrics compared with manual delineations. Additionally, FCM that requires minimal user input such as a region of interest (ROI) placed around the tumor still resulted in higher precision errors compared with GCGMM. Similarly, GC, an interactive segmentation method resulted in the largest precision errors shown by both larger $\%CV_{RMS}$ and SD_{RMS} using all the performance metrics.

Figure 2(a) shows the inter-rater segmentation variability for an example case. Computer generated segmentations for GC, FCM, and GCGMM computed using three different user inputs are also shown for comparison. As seen, the GCGMM and FCM segmentations show lower variability compared with either the GC or multi-rater segmentations. As shown in Fig. 2(b), (Table S1), overall, GCGMM achieved more consistent segmentation performance compared with all the analyzed methods.

We measured the reproducibility of the textures extracted from the various segmentations generated using the various methods and with multiple user inputs by computing the intraclass correlation coefficient (ICC) between the texture features. The inter-rater manual segmentations were the least reproducible and achieved the lowest ICC with a median of 0.65 (IQR 0.550.79). The features computed from GCGMM segmentations were the most reproducible with highest ICC with a median of 0.89 (IQR 0.790.925) compared with ICC of features computed from GC median of 0.72 (IQR 0.680.78) and FCM median of 0.73 (IQR 0.660.82). Thirty four out of the 36 features computed using GCGMM method had higher ICC compared with inter-rater manual delineations with the exception of MRI pre-contrast intensity and pre-contrast standard deviation features. Similarly, 33 and 31 features computed using GCGMM had higher ICC compared with FCM and GC method, respectively.

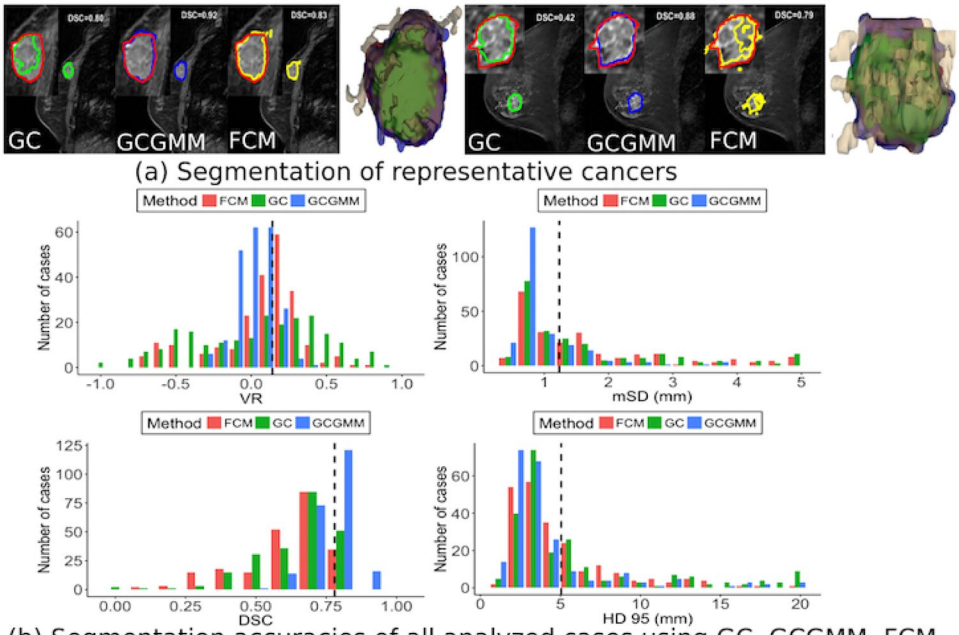

Figure 1. Performance of segmentation methods. (**a**) Example segmentations produced using GrowCut(GC), GC combined with Gaussian mixture models (GCGMM), fuzzy c-means clustering method (FCM) and volumes produced using all methods overlaid with expert delineated volume and (**b**) overall performance of the segmentation methods for all analyzed tumours. The inter-rater segmentation concordance computed using the various metrics is shown for reference using dashed lines.

The time required for generating segmentations using GCGMM was 148 secs ± 108 secs compared with FCM (38 secs ± 12 secs) and GC (55 secs ± 25 secs) methods using a HP Z820 PC. Only the GC algorithm was optimized for speed using multi-threading using implementation in C++. The tensor computation was also implemented in C++ for speed. The rest of the algorithm, particularly, Gaussian mixture modeling is implemented in Matlab.

Classifiers trained using features extracted from computer-generated segmentations were comparable to classifiers trained using features extracted from expert delineations. Classifiers trained using features extracted from GCGMM segmentations achieved the best accuracy for differentiating between the breast cancer molecular subtypes (Table 3, Fig. 3). Furthermore, GCGMM-based classifiers outperformed classifiers that used features computed from expert delineated tumors.

The ranking of features varied across classifiers (Table 4). Only the features extracted using GCGMM and expert delineation showed significant differences between ERPR + HER2− vs. TN (Table 4). When using the expert delineations, TN cancers had a significantly higher contrast texture compared with ERPR + HER2− cancers (Fig. 3(b)). TN cancers also had a significantly lower first-post contrast MRI correlation (Fig. 3(b)). Four of the top five features computed using GCGMM were significantly different between the two cancers. The TN cancers had significantly lower kurtosis from the second, and third post-contrast MRI, and significantly higher skewness from the second post-contrast MRI (Fig. 3(c).

Discussion

We developed an appearance constrained interactive segmentation method, which generated accurate for breast cancers with three different molecular subtypes as well as with different tumour presentations (mass and non-mass) and background parenchymal enhancement (mild and marked). GCGMM produced reproducible segmentations with least precision errors compared to manual, FCM, and GC segmentation methods. Our method was significantly more accurate than GC[20] and FCM[26] both of which have been used in various radiomics applications including the lung[21] and breast cancers[11,15].

GCGMM resulted in lowest $\%CV_{RMS}$ and lowest SD_{RMS} using all performance metrics compared with other segmentation methods. The volume precision errors using GCGMM were the lowest ($\%CV_{RMS} = 14.5\%$) compared with all methods including inter-rater segmentations. Similarly, the Hausdorff distance errors were also the lowest with ($\%CV_{RMS} = 20.7\%$) using GCGMM compared with ($\%CV_{RMS} = 48.6\%$) when using manual delineations. The precision errors computed using the GC method were high and more comparable to the inter-rater delineations than the FCM or GCGMM methods, clearly underscoring the fact that an interactive method such as GC is impacted by variability in user inputs. Finally, texture measures computed from GCGMM were more

	FCM				GC				GCGMM			
Analysis	DSC	mSD	HD95	|VR|	DSC	mSD	HD95	|VR|	DSC	mSD	HD95	|VR|
Overall mean	0.66	1.85	5.55	0.27	0.69	2.97	7.38	0.21	0.81***,***	1.08**,***	4.82**,***	0.12***,***
SD	0.15	1.31	3.41	0.16	0.15	12.29	14.18	0.18	0.07	0.59	3.67	0.08
Mild BPE mean	0.65	1.89	5.43	0.29	0.70	1.73	8.58	0.20	0.80***,***	1.11*,***	5.27 *ns*,*ns*	0.13*,***
SD	0.15	1.19	3.15	0.16	0.12	1.41	20.98	0.15	0.06	0.62	4.61	0.08
Marked BPE mean	0.68	1.74	5.73	0.25	0.68	3.48	6.41	0.24	0.81***,***	1.01*ns*,***	4.44*ns*,*ns*	0.10***,***
SD	0.15	1.41	3.88	0.17	0.15	14.84	4.53	0.19	0.07	0.58	2.82	0.07
Mass mean	0.66	1.93	5.63	0.27	0.70	2.34	5.73	0.21	0.82***,***	1.02***,***	4.24*ns*,*ns*	0.12***,***
SD	0.16	1.39	3.66	0.17	0.14	8.49	4.10	0.16	0.07	0.45	2.49	0.08
Non-mass mean	0.68	1.64	5.32	0.27	0.66	4.57	11.64	0.23	0.78*ns*,*	1.24*ns*,*ns*	6.31*ns*,*ns*	0.11***,*ns*
SD	0.14	1.06	2.66	0.15	0.17	18.84	25.63	0.21	0.07	0.84	5.42	0.08
ER-HER2+ mean	0.67	1.77	5.36	0.27	0.69	2.78	7.94	0.22	0.81***,***	1.03***,***	4.92*ns*,*ns*	0.10***,***
SD	0.16	1.32	3.07	0.16	0.14	11.90	16.6	0.17	0.06	0.61	3.88	0.07
TN mean	0.65	2.03	5.15	0.29	0.73	5.41	6.78	0.19	0.82*ns*,*ns*	1.21*ns*,*ns*	4.63*ns*,*ns*	0.14*,*
SD	0.14	1.26	2.34	0.16	0.19	20.65	5.85	0.22	0.09	0.52	2.50	0.09
ERPR + HER2− mean	0.65	2.06	6.59	0.29	0.69	2.01	5.54	0.22	0.79*ns*,*ns*	1.18*ns*,*ns*	4.55*ns*,*ns*	0.15*ns*,*ns*
SD	0.14	1.27	4.91	0.17	0.15	2.06	3.72	0.19	0.07	0.52	3.52	0.09

Table 1. Segmentation accuracies generated using GC, GCGMM, and FCM presented using mean and standard deviation (SD). FCM Fuzzy c-means clustering; GC GrowCut; GCGMM GrowCut with Gaussian Mixture Models. DSC Dice coefficient; mSD mean surface distance; HD95 95^{th} percentile of Hausdorff distance; |VR| absolute volume ratio. Significant differences between GCGMM vs. FCM and GCGMM vs. GC are indicated above each metric for the corresponding analysis after adjusting for multiple comparisons using Bonferroni-Holm correction. ns $P \geq 0.05$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

	SD_{RMS}					$\%CV_{RMS}$				
Method	DSC	mSD (*mm*)	HD95 (*mm*)	|VR|	Volume (cc)	DSC	mSD	HD95	|VR|	Volume (cc)
Manual	0.084	0.063	4.6	0.10	1.08	11.1	48.3	48.6	62.6	29.4
FCM	0.06	0.91	2.38	0.06	2.46	13.6	31.9	29.7	33.5	36.1
GC	0.10	12.3	13.5	0.14	37.6	19.6	50.0	26.7	64.2	43.8
GCGMM	0.038	0.31	1.33	0.057	1.75	5.07	21.2	20.7	54.3	14.5

Table 2. Reproducibility of segmentations generated using multiple raters and by algorithms (GC, FCM, GCGMM) using different user inputs. SD_{RMS} Root mean square of standard deviation; $\%CV_{RMS}$ Percentage coefficient of variation in the RMS value for a specific metric FCM Fuzzy c-means clustering; GC GrowCut; GCGMM GrowCut with Gaussian Mixture Models. DSC Dice coefficient; mSD mean surface distance; HD95 95^{th} percentile of Hausdorff distance; |VR| absolute volume ratio.

reproducible compared with GC and FCM segmentations as well as inter-rater delineations and resulted in the highest ICC. Ultimately, features computed using the GCGMM segmentations produced the best classification accuracy in a radiomics classification task involving cancer molecular subtypes and only the features computed using GCGMM besides the expert delineation were able to capture significant differences between the studied breast cancer molecular subtypes. Our results demonstrate that GCGMM is a feasible method for generating accurate and reproducible segmentations for breast cancer radiomics analysis. GCGMM method took longer to compute compared with the GC or the FCM method. However, the computation time on average was under 3 mins. We did not perform any code optimization while computing the run times.

Our method resulted in fewer over- or under-segmentations compared with either GC or FCM. We developed an in-house GUI for interactive selection of the appropriate volumetric lesion segmentation, which enables simultaneous radiologist validation. Given the evidence of the importance of tumour volumes in assessing treatment response in neoadjuvant chemotherapy[3] and for improving surgical outcomes[5], an approach such as ours can potentially benefit the translation of computer-aided techniques into clinical settings. We are currently evaluating our approach among a different cohort of breast cancer patients imaged prior to and following treatment with neoadjuvant chemotherapy.

Repeated interactions as needed in GC[20] can be especially cumbersome when segmenting large datasets. Fully automatic methods[3,8,9,26,27] need little to no user interaction but may lead to less accurate results as they fail to match the expert's assessment of tumour boundary. In this report, we improved the performance, in both accuracy and reproducibility of an interactive method while limiting user input (brush strokes or rectangular ROI enclosing the tumour) by using a simple cancer-specific appearance modeling approach in favor of voxel-wise shallow learning[28–30] and more recent deep learning methods[31–33]. Our approach takes advantage of the temporal variability in the lesion appearance and derived image representations such as the temporal difference[13]

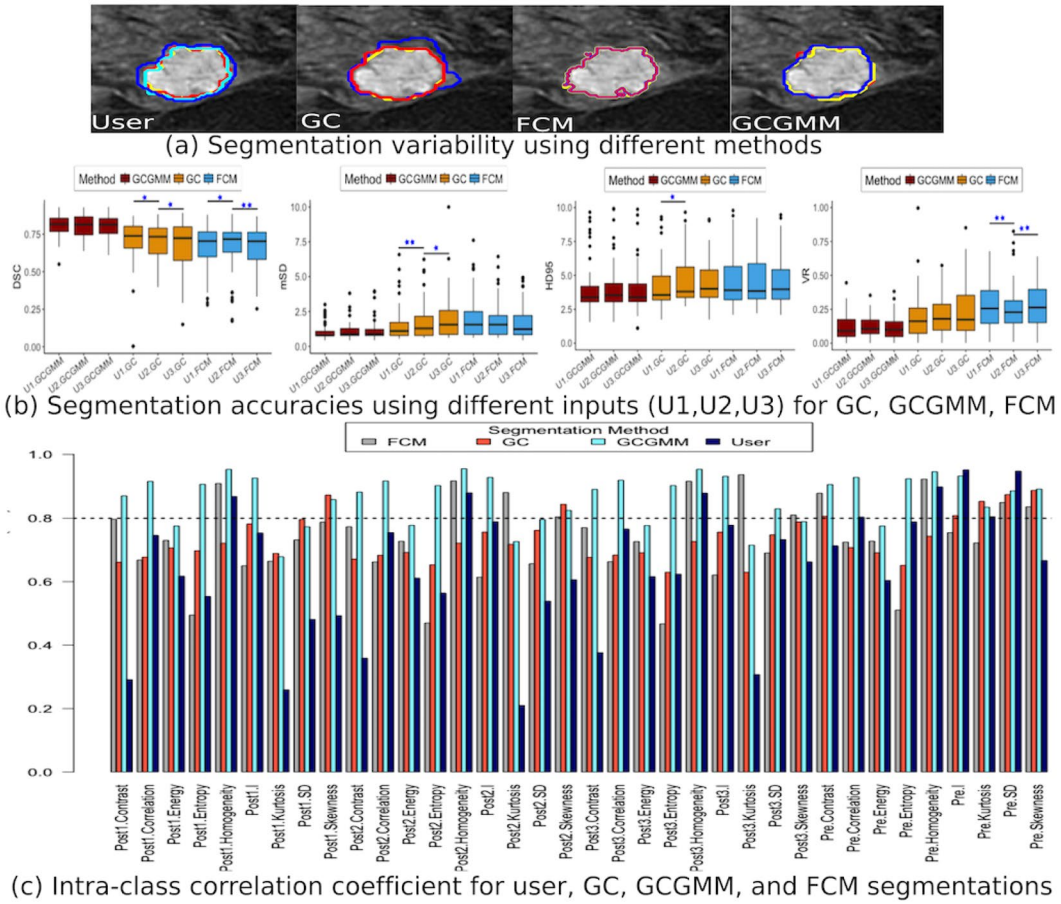

Figure 2. Segmentation variability for the different methods. The inter-rater delineations, and the segmentations generated using three different user inputs are shown in (**a**). The segmentation accuracies achieved by the different methods for the three different user inputs is shown in (**b**) and the segmentations with significantly different accuracies using a given measure are identified, where $*P < 0.05$ and $**P < 0.01$. The p-values are reported after adjusting for multiple comparisons using Bonferroni-Holm method. The intra-class correlation coefficient (ICC) of the texture measures computed from the generated segmentations are shown in (**c**).

Method	ER-HER2+ vs. ERPR+HER2−/TN					ERPR+HER2− vs. TN				
	TPR	TNR	FPR	FNR	AUC (95% CI)	TPR	TNR	FPR	FNR	AUC (95% CI)
Expert	0.85	0.91	0.09	0.15	0.95 (0.91–0.97)	0.78	0.91	0.09	0.22	0.91 (0.79–0.97)
FCM	0.85	0.85	0.15	0.15	0.92 (0.87–0.96)	0.74	0.83	0.17	0.26	0.83 (0.67–0.91)
GC	0.79	0.79	0.21	0.21	0.90 (0.86–0.94)	0.70	0.78	0.22	0.30	0.77 (0.61–0.90)
GCGMM	0.93	0.81	0.19	0.07	0.95 (0.92–0.98)	0.83	0.96	0.04	0.17	0.92 (0.82–0.97)

Table 3. Classifier accuracies using features computed from different segmentations. TPR - true positive rate, TNR - true negative rate, FPR - false positive rate, FNR - false negative rate, AUC - area under the curve.

and tensor-derived scalar images inspired by[34,35] that seek to differentiate the tumour's appearance from its background. Our results show that our approach generates consistently accurate segmentations for a variety of tumour molecular subtypes, patterns of enhancement, and BPE. Prior works on breast cancer segmentation typically focused on specific tumour types such as ER(+), node negative tumours as in[28] or tumours with specific appearance including mass and non-mass enhancing patterns as in[30], datasets with malignant and benign breast cancers[34,36].

Prior works including[18,21] showed that GC segmentations were more repeatable than manual delineations produced by different users both in terms of segmentation variability and texture feature reproducibility. Our work went a step further to improve the reproducibility of GC using GCGMM and assessed the performance difference in a radiomics task when using features computed from the different segmentations. Our results show that features computed from any of the analyzed algorithmic methods produced similar results as manual delineations and can in fact yield better results, as in the case of GCGMM. Furthermore, our work illustrates the utility of using volumetric measurements for improving classification accuracy. Previously, we used a different cohort of

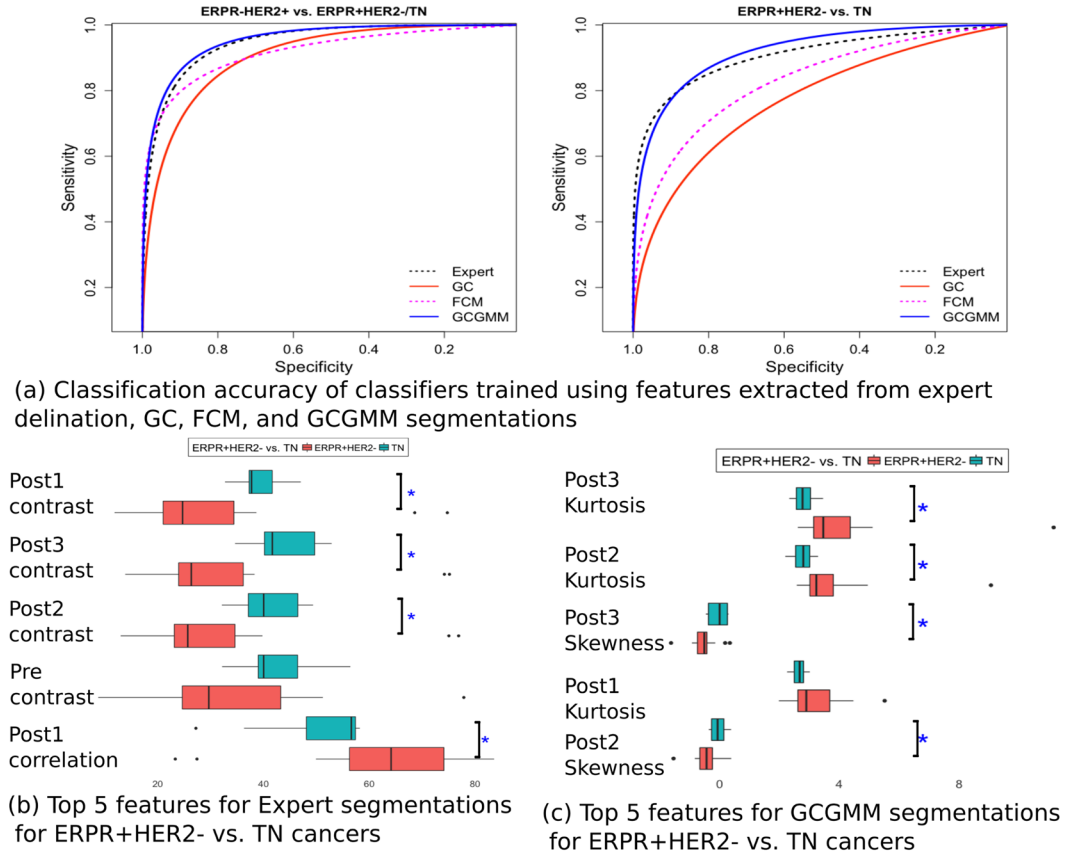

Figure 3. Performance of classifiers trained with textures extracted from different segmentations. (**a**) ROC curves for classifiers trained using features extracted from various segmentations for distinguishing between ER-HER2+ vs. ERPR + HER2−/TN and ERPR + HER2− vs. TN cancers. The five most relevant features and their differences between ERPR + HER2− vs. TN cancers for expert delineated (**b**) and GCGMM segmented tumors (**c**) are also shown.

Expert	p-Value	FCM	p-Value	GC	p-Value	GCGMM	p-Value
ER-HER2+ vs. ERPR+/TN							
Post2 I	0.74	Post3 Kurt	0.56	Post1 Skew	1.00	Post1 I	1.00
Post2 Skew	0.31	Pre Kurt	0.36	Pre Kurt	1.00	Post3 I	1.00
Post1 I	1.00	Post2 Kurt	0.56	Pre Contrast	1.00	Post2 I	1.00
Post1 Corr	1.00	Post1 Kurt	0.56	Post1 Kurt	1.00	Pre Energy	0.19
Post1 Entropy	1.00	Post3 SD	0.56	Post2 Skew	1.00	Post3 Skew	1.00
ERPR + HER2− vs. TN							
Expert	p-Value	FCM	p-Value	GC	p-Value	GCGMM	p-Value
Post1 Contrast	0.04	Post3 Kurt	0.27	Post3 Homogeneity	0.71	Post3 Kurt	0.01
Post3 Contrast	0.02	Post3 SD	0.65	Post3 Skew	0.71	Post2 Kurt	0.01
Post2 Contrast	0.04	Post2 Skew	0.65	Post2 Skew	0.58	Post3 Skew	0.01
Pre Contrast	0.08	Post1 Kurt	0.32	Pre SD	1.00	Post1 Kurt	0.16
Post1 Corr	0.04	Post1 Skew	0.65	Post2 I	1.00	Post2 Skew	0.01

Table 4. Results of Wilcoxon test to assess the difference between ER-HER2+ vs. ERPR + HER2−/TN and ERPR + HER2− vs. TN cancers using top five-most relevant (determined using Gini importance) features extracted using RF classifiers and trained using features generated from the different segmentation methods. P-values are reported after adjusting for multiple comparisons using Bonferroni-Holm method. FCM: Fuzzy c-means; GC: Grow-Cut; GCGMM: Grow-Cut with Gaussian Mixture Models Pre: Pre constrast MRI; Post1: first post-contrast MRI; Post2: second post-contrast MRI; Post3: third post-contrast MRI I: intensity; skew: skewness; corr: correlation; kurt: kurtosis; SD: standard deviation.

patients[10] to differentiate between the breast cancer subtypes and our results clearly demonstrate the performance improvement.

Four out of five top ranked features extracted using GCGMM and expert delineation were significantly different between ERPR+ and TN cancers. Similar to the findings from[11,12] which found TN cancers to be more heterogeneous, our results show that using both expert delineated and GCGMM segmentations, TN cancers were associated with higher heterogeneity, namely, larger contrast and lower kurtosis. Finally, it is interesting to note that classifiers trained using different segmentations resulted in different ranking of features.

Our work has the following limitations. First, the dataset was imbalanced between the different molecular subtypes which required data balancing using the SMOTE technique[43]. Second, experts generated delineations in consensus which prevented us from studying the variability of auto-generated segmentation with respect to inter-rater variability. We tried to address this issue by benchmarking the inter-rater variability using a small number of randomly chosen cases. Nevertheless, we evaluated our approach on a reasonably diverse set of tumours and performed a systematic evaluation starting from auto-generated segmentation to assessing feasibility of features extracted from such segmentations in a radiomics task.

Methods

Study design and patients. Our institutional review board approved our HIPAA-compliant retrospective study. A retrospective cohort of 75 patients diagnosed with pathologically-proven invasive ductal breast carcinoma between 2006–2011 were analysed. Tumour subtypes were identified through immunohistochemistry with known ER, PR, and HER2/neu receptor status. Inclusion criteria were: (i) preoperative bilateral breast MRI, (ii) no prior history of cancer, (iii) no known BRCA mutation, and (iv) no current use of hormonal therapy. Our study population consisted of 56 HER2 receptor positive (HER2+, n = 56), 15 estrogen positive (ER) and progesterone receptor (PR) positive, and 9 triple negative (TN, n = 9) tumours. Thirty-six patients used in this study overlapped with those used in[10] and all the 15 ERPR+ patients overlapped with those used in[16].

Sagittal T1-weighted, fat-suppressed 2D multi-slice (40–50 slices) images were acquired with a 1.5-T MRI system (Signa or Signa HDX; GE Medical Systems) using a dedicated 8-channel surface breast coil before and continuously at three times after the intravenous administration of 0.1 mmol gadopentetate-dimeglumine per kilogram body weight (Magnevist) using the following scan parameters: repetition time (ms)/echo time (ms), 7.4/4.2; flip angle, 10°; bandwidth, 32 kHz; field of view 18–22 *cm*; acquisition matrix 256 × 192; slice thickness, 3 mm; temporal resolution 90 s.

A radiologist (EJS) with six years of experience reading breast MRIs who was blinded to cancer molecular subtype classified all tumors as having mass or non-mass enhancement (NME). BPE was also assessed as mild or marked BPE. Tumours classified by the radiologist as having both mass and non-mass enhancement were classified as NME for the purpose of analysis. Two radiologists (EJS, BZD) generated volumetric manual delineation of the tumours using the first post-contrast T1w MRI in consensus using ITK-SNAP[37] software which served as the ground truth segmentation.

User inputs for segmentations. The goal of the user input experiment was to study the robustness of the algorithms in generating volumetric segmentations with varying user inputs. Therefore, we used the following strategy to evaluate the segmentation performance. Three users (two radiologists and computer scientist) produced inputs for the segmentation method. User EJS traced a contour delineating the tumor on a single slice. The second user input was placed to roughly enclose the tumor. The main difference between the first and second input was that while the first user carefully followed the tumor boundary including spiculations, the second input was a rough polygonal region of interest (ROI) that did not follow the exact tumor boundary and simply enclosed the tumor. The third input (tumor/background) consisted of a contour drawn within the tumor. Additionally, the third user placed a background contour outside the tumor. The users' inputs are shown in (Fig. 4(i)).

GC[20] employs competitive region growing starting from user interactions to produce segmentations according to user preference. Our implementation available in 3DSlicer[17] for scalar images can use multiple rounds of user inputs to produce a final segmentation. We restricted the user inputs to be presented once during initialization to a single representative slice to make the inputs as close to a fully automatic method as possible. Furthermore, we implemented an automatic background stroke extraction to limit user effort to providing only a rectangular ROI enclosing the tumour.

Our method automatically converted the ROI and contour inputs to extract foreground and background strokes as follows. Foreground strokes were computed from the user contour by extracting the morphological skeleton using $r-1$ iterations, where r corresponds to the half of maximum equivalent contour diameter. The background labels were extracted by subtracting two sets of automatically extracted ROIs computed by dilating the original user-drawn ROI (or contour) using ($d_1 = r$) and ($d_2 = max(2, r-2)$) iterations. The user input enclosing the tumor for the contour and ROI inputs were subjected to one iteration of morphological erosion to ensure that the extracted foreground strokes were contained within the tumour. Next, the foreground strokes were drawn as perpendicular lines extending from the centroid and till the minor axis length of the eroded ROI. The three inputs for an example case are shown in Fig. 4(i).

The inputs for the FCM method consisted of a rectangular ROI extracted by computing the bounding box enclosing the background strokes.

Segmentation Method. Eight feature images consisting of pre, and three post-contrast MRI, three temporal difference images (computed per voxel as, $\varepsilon_t = (I_t(x) - I_0(x))^2$, where, I_t was the post-contrast image at time t and x the voxel location), and a trace image computed from tensor representation of the DCE-MRI were used in the analysis. A voxel-wise tensor was computed from a voxel-wise covariance matrix

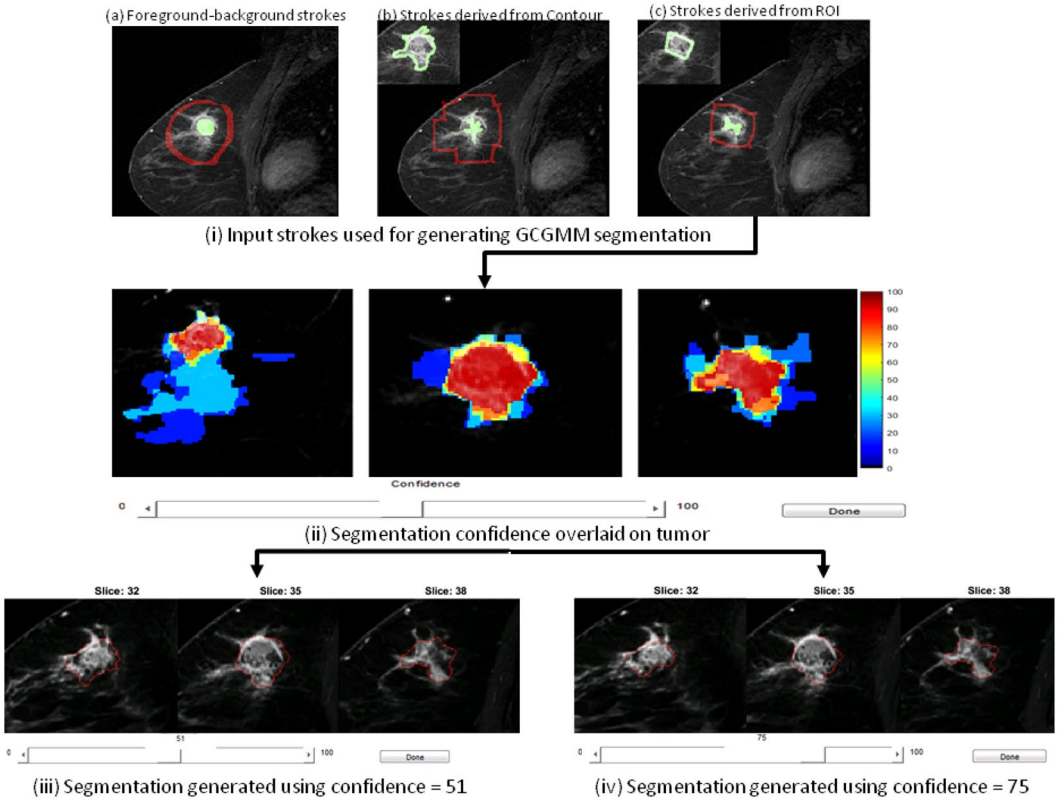

Figure 4. Workflow diagram. (i) Inputs used for generating segmentations, (ii) confidence map computed from GCGMM using region of interest refined input from (i) c, and segmentations generated using two different confidence thresholds (iii,iv) for a triple negative breast cancer.

$$A = \begin{bmatrix} i_0^2 & (i_0 - i_t)^2 & \dots & (i_0 - i_k)^2 \\ \vdots & \dots & & \vdots \\ (i_k - i_0)^2 & \dots & & i_k^2 \end{bmatrix},$$

where, t_i was the intensity of a voxel at time t. Eigen decomposition of A using the top three eigenvalues produced the temporal tensor at each voxel from which the trace image was computed. The trace image summarized the variation in the contrast uptake within the tumour and in the normal parenchyma.

All eight feature images were used for producing segmentation using GC, FCM, and GCGMM methods. FCM clustering used the same parameter settings as used in[26].

The GCGMM method produced tumour segmentation through a weighted combination of GC segmentations from individual feature images with GMM-based voxelwise classification using:

$$L = \frac{1}{N}\sum_{i=1}^{N} S_i \times (1 - \gamma) + G \times \gamma > \omega, \tag{1}$$

where, S_i is the GC segmentation for feature image i, G the GMM model-based segmentation, N the number of feature images, and $\omega = 0.6$ is an empirically chosen default confidence threshold. The parameter γ weights the contribution of GMM and GC segmentation. It corresponds to the $F_{\beta=0.5}$ measure[38] that emphasizes precision over recall to account for large data imbalance between cancer and normal voxels. One GMM model is trained per tumor where the GMM model contains all the features as a vector. Therefore, the γ values were chosen per tumor. In general, the γ values ranged between 0.09 to 0.75 with mean value of 0.37 ± 0.16 for all the analyzed cases.

The final segmentation was produced by the weighted sum of GC segmentations for each feature image with the GMM-based voxel-wise classification. An alternative approach would be to produce a single GC segmentation by using all the feature images simultaneously (with equal weights) and combining that with the GMM-based classification. We chose the former approach as we hypothesized that the latter approach where all features are weighted equally would result in an under-segmentation as only voxels that are highly similar to the user-labeled tumor voxels and with largest feature distances from background voxels would be labeled as tumour.

We developed a graphical user interface in Matlab (Fig. 4(iii,iv)) that allows a user to dynamically change the confidence threshold ω and produce the desired segmentation.

Multi-Parametric Gaussian Mixtures Model-based Tumour Extraction. Multi-parametric Gaussian Mixture Models (GMM) were extracted from the feature images using tumour and background input labels. The GMM model parameters, namely, the mean (μ), covariance (Σ), mixing weights (w), and the number of components (n), were automatically extracted from the data. Akaike Information Criterion (AIC) was used to select the appropriate number of mixture components for each GMM from (n = 2, 3, 4). Three was the most frequently selected number of components for tumour and background. GMM models for the tumour and background were computed using expectation maximization (EM) algorithm. The extracted GMM model was then used to produce voxel-wise labelling throughout the entire image. A voxel x was assigned tumour or background label to produce a GMM label image G using,

$$G(x) = \begin{cases} \text{if k (x, T)} > \text{k(x, B)}, & \textit{tumour} \\ \text{otherwise}, & \textit{background}, \end{cases} \tag{2}$$

where $k(x, T)$, $k(x, B)$ are the similarity distances of a voxel x computed with respect to the tumour T and the background B models. To limit the number of false positives, we required that the tumour probability $k(x, T) > \tau$, where $\tau = 0.75$.

Metrics for evaluating segmentation accuracy. Algorithm generated segmentations A were compared with radiologist delineated segmentation G using spatial overlap computed using the Dice coefficient $\left(DC = \frac{2 * A \cap G}{(A \cap G + A \cup G)}\right)$, a volume-based measure called the absolute volume difference ratio $\left(|VR| = \frac{|v(A) - v(G)|}{0.5 * (v(A) + v(G))}\right)$ and two distance measures namely, mean surface distance (mSD) and the 95% Hausdorff distance (HD95). HD95 was defined as 95^{th} percentile distance over all point distances in contour X to its closest point in contour Y:

$$HD95 = 95\%\left(\min_{y \in Y} d(x, y)\right) \forall\, x \in X, \tag{3}$$

where $d(x, y)$ is the distance between the points x and y in X and Y, respectively. The mean surface distance between two contours X and Y is defined as:

$$mSD(X, Y) = \frac{1}{|X|} \sum_{x \in |X|} \min_{y \in Y} d(x, y) \tag{4}$$

Large values of the Dice and small values of mSD, HD95, and |VR| indicate high accuracies. The 95^{th} percentile Hausdorff distance was used as this is more robust to outliers as explained in[39].

Metrics for evaluating segmentation reproducibility. Segmentation reproducibility resulting from the various methods using multiple user inputs was measured by computing the root mean square (RMS) of the coefficient of variation ($\%CV_{RMS}$) and the RMS of standard deviation (SD_{RMS}) in the segmentation metrics and as described in[40,41]. We used the $\%CV_{RMS}$ as this measure has been shown to be a conservative measure of segmentation reproducibility in[41]. CV is a measure of relative variability and is defined as the ratio of the standard deviation to the mean. The %CV measures for each method i and patient p using a segmentation metric $M^j = \{j = DSC, mSD, HD95, |VR|\}$ were computed as,

$$\%CV_p(M^j) = \frac{\overline{M_p^j}}{\widetilde{M_p^j}} \times 100, \tag{5}$$

where, $\overline{M_p^j}$ is the standard deviation in the metric M^j for the multiple user input trials in a given patient p, and $\widetilde{M_p^j}$ is the mean value of that metric for those same trials and patient. The RMS value for the %CV for each segmentation metric was then computed as,

$$\%CV_{RMS}(M^j) = \sqrt{\frac{1}{N}\sum_{p=1}^{N}(\%CV_p(M^j)^2)}. \tag{6}$$

The RMS SD for each segmentation metric was computed as,

$$RMS_{SD} = \sqrt{\frac{1}{N}\sum_{p=1}^{N}\overline{M_p^j}^2}. \tag{7}$$

Radiomics feature extraction and classification. Thirty-six texture features were computed from the DCE-MRI consisting of four first order textures (mean, standard deviation, kurtosis, and skewness) and five second order Haralick texture measures (energy, entropy, correlation, homogeneity, and contrast) from each MR image sequence. The Haralick textures were computed from a gray-level co-occurrence matrix after rescaling the images (0–255) and using 24 histogram bins. Texture measures were computed within the volumetrically segmented tumours using manual, FCM, GC, and GCGMM methods for all the trials resulting in 27000($36 \times 3 \times 3 \times 75 + 36 \times 75$) texture values. Reliability of the computed textures resulting from segmentations generated

by using multiple user inputs was measured by computing the intra-class correlation coefficient (ICC) as used in previous studies[18].

Random forest classifiers[42] (with 100 trees and default parameters) were computed using texture measures extracted using each segmentation generated from stroke inputs for distinguishing between (a) HER2+ vs. ERPR+/TN, and (b) ERPR+ vs. TN. Datasets were balanced using the synthetic minority oversampling technique (SMOTE)[43]. Classifier accuracy was evaluated using leave-one-out cross-validation (LOOCV).

Statistics. Associations between categorical measures (segmentation method, user input trial, molecular subtype, enhancement) and continuous variables (DSC, mSD, and VR) were studied using Kruskal-Wallis tests. Paired associations between continuous variables were analyzed using Wilcoxon rank sum test. P values of <0.05 were considered to be statistically significant. Bonferroni-Holm correction was applied to account for multiple comparisons. All statistical analysis was computed using R statistical software[44].

Data availability statement. All of the generated segmentation metrics and texture measures are available in supplementary data. The R code used for performing the statistical analysis is available from the github repository https://github.com/harveerar/SciRepStatAnal/.

Conclusions. We developed a cancer-specific appearance constrained interactive segmentation method for generating volumetric delineations of breast cancers from DCE-MRI. We performed a systematic evaluation of the method starting from segmentation performance, the influence of multiple user inputs on segmentation differences, and its utility for a radiomics task. Our results show that the GCGMM segmentations were accurate, reproducible and a classifier trained using features extracted from those segmentations were as good or better than classifier trained using features extracted from expert delineations for differentiating between breast cancer molecular subtypes.

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics. *CA Cancer J Clin* **67**, 7–30, https://doi.org/10.3322/caac.21332 (2017).
2. Graham, L. *et al.* Current approaches and challenges in monitoring treatment responses in breast cancer. *J Cancer* **5**, 58–68, https://doi.org/10.7150/jca.7047 (2014).
3. Minarikova, L. *et al.* Investigating the prediction value of multiparametric magnetic resonance imaging at 3T in response to neoadjuvant chemotherapy in breast cancer. *Eur Radiol* 1–11, https://doi.org/10.1007/s00330-016-4565-2 (2016).
4. Hylton, N. *et al.* Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy-results from ACRIN 66571/I-SPY TRIAL. *Radiology* **263**, 663–672 (2012).
5. Houssami, N. *et al.* Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: Systematic review and meta-analysis in detection of multifocal and multicentric cancer. *J. Clin Oncology* **26**, 3248–3258 (2008).
6. Kumar, V., Gu, Y., Basu, S., Eschrich, S. & Schabath, M. B. e. QIN "Radiomics: the process and the challenges". *Magn Reson Imaging* **30**, 1234–1248, https://doi.org/10.1016/j.mri.2012.06.010 (2012).
7. Ahmed, A., Gibbs, P., Pickles, M. & Turnbull, L. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *Journal of magnetic resonance imaging: JMRI* **38**, 89–101, https://doi.org/10.1002/jmri.23971 (2013).
8. Teruel, J. R. *et al.* Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *NMR Biomed* **27**, 887–96, https://doi.org/10.1002/nbm.3132 (2014).
9. Aghaei, F. *et al.* Computer-aided breast MR image feature analysis for prediction of tumor response to chemotherapy. *Med Phys* **42**, 6520–8, https://doi.org/10.1118/1.4933198 (2015).
10. Sutton, E. *et al.* Breast cancer molecular subtype classifier that incorporates MRI features. *J. Magn Reson Imaging* **44**, 122–129 (2016).
11. Li, H. *et al.* Quantiative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *NPJ, Breast Cancer* **2**, 16012 (2016).
12. Agner, S. *et al.* Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: A feasibility study. *Radiology* **272**, 91–99 (2014).
13. Ashraf, A. *et al.* Identification of intrinsic imaging phenotypes of breast cancer tumours: Preliminary associations with gene expression profiles. *Radiol* **272**, 374–384 (2014).
14. Waugh, S. *et al.* Magnetic resonance imaging texture analysis classification of primary breast cancer. *Eur. Radiol* **26**, 322–330 (2016).
15. Bhooshan, N. *et al.* Cancerous breast lesions on dynamic contrast-enhanced MR images: Computerized characterization for image-based prognostic markers. *Radiology* **254**, 680–690 (2010).
16. Sutton, E. *et al.* Breast cancer subtype intertumor heterogeneity: MRI-based features predict results of a genomic assay. *J. Magn Reson Imaging* **42**, 1398–1406 (2015).
17. Egger, J. *et al.* GBM volumetry using the 3DSlicer medical image computing platform. *Sci. Reports* **3**, 1364 (2013).
18. Parmar, C. *et al.* Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* **9**, e102107, https://doi.org/10.1371/journal.pone.0102107 (2014).
19. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal Machine Intelligence* **28**, 1768–1783 (2006).
20. V. Vezhnevets & V. Konouchine. GrowCut - Interative multi-label N-D image segmentation. In *Proc. Graphicon*, 150–156 (2005).
21. Velazquez, E. R. *et al.* Volumetric CT-based segmentation of nsclc using 3D-Slicer. *Scientific Reports* **3**, 3529, https://doi.org/10.1038/srep03529 (2013).
22. Blake, A., Rother, C., Brown, M., Perez, P. & Torr, P. Interactive image segmentation using an adaptive GMMRF model. In *Proc. European Conference in Computer Vision (ECCV)* (Springer-Verlag 2004).
23. Veeraraghavan, H. & Miller, J. Active learning guided interactions for consistent image segmentation with reduced user interactions. In *IEEE Intl Symposium on Biomedical Imaging: From Nano to Micro*, 1645–1648 (2011).
24. Wu, J., Zhao, Y., Zhu, J.-Y., Luo, S. & Tu, Z. MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *IEEE Conf. Computer Vision and Pattern Recognition*, 256–263 (2014).
25. Triebel, R., H, S., Soiai, M. & Cremers, D. Active online learning for interactive segmentation using sparse Gaussian processes. In X. Jiang, J. Hornegger & R. Koch (eds) *Pattern Recognition*, vol. 8753 of *Lecture Notes in Computer Science*, 641–652 (Springer International Publishing, 2014). https://doi.org/10.1007/978-3-319-11752-2_53.
26. Chen, W., Giger, M. & Bick, U. A fuzzy C-Means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiology* **13**, 63–72 (2006).
27. Twellman, T., Meyer-Baese, A., Lange, O., Foo, S. & Nattkemper, T. Model-free visualization of suspicious lesions in breast MRI based on supervised and unsupervised learning. *Eng. Applic. Artif. Intell* **21**, 129–140 (2008).

28. Ashraf, A. *et al.* A multi-channel Markov random field framework for tumor segmentation with an application to gene expression-based breast cancer recurrence risk. *IEEE Trans. Med Imaging* **32**, 637–648 (2013).
29. Chang, Y.-C., Huang, Y.-H., Huang, C.-S., Chen, J.-H. & Chang, R.-F. Computerized breast lesions detection using kinetic and morphologic analysis for dynamic contrast-enhanced MRI. *Magnetic Resonance Imaging* **32**, 514–522 (2014).
30. Gubern-Mérida, A. *et al.* Automated localization of breast cancer in DCE-MRI. *Medical Image Analysis* **20**, 265–274 (2015).
31. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, 8 (2015).
32. Brosch, T., Tang, L., Yoo, Y., Traboulsee, A. & Tam, R. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE. Trans. on Medical Imaging* **35**, 1229–1238 (2016).
33. Milletari, F., Navab, N. & Ahmadi, S. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *CoRR* **abs/1606.04797**. http://arxiv.org/abs/1606.04797 (2016).
34. Agner, S., Xu, J. & Madabhushi, A. Spectral embedding based active contour (SEAC) for lesion segmentation on breast dynamic contrast enhanced magnetic resonance imaging. *Medical Physics* **40**, 032305–032312 (2012).
35. Jayender, J., Chikarmane, S., Jolesz, F. A. & Gombos, E. Automatic segmentation of invasive breast carcinomas from DCE-MRI using time series analysis. *J. Magn Reson Imaging* **40**, 467–475 (2014).
36. Zheng, Y., Baloch, S., Englander, S., Schnall, M. & Shen, D. Segmentation and classification of breast tumor using dynamic contrast-enhanced MR images. In *Med. Image Comput Comput Assist Interv* (2007).
37. Yushkevich, P. & Gerig, G. ITK-SNAP: An interactive medical image segmentation tool to meet the need for expert-guided segmentation of complex medical images. *IEEE. Pulse* **8**, 54–57 (2017).
38. van Rijsbergen, C. Foundations of evaluation. *Documentation* **30**, 365–373 (1974).
39. Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* **34**(10), 1993–2023 (2015).
40. Glüer, C. C. *et al.* Accurate assessment of precision errors: how to measure the reproducibility of bone densitometry techniques. *Osteoporosis Int.* **5**, 262–70 (1995).
41. Mastmeyer, A. & Pernelle, G. Accurate model-based segmentation of gynecologic brachytherapy catheter collections in MRI images. *Med Image Anal* **42**, 173–188 (2017).
42. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
43. Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16**, 321–357 (2002).
44. core team, R. R:A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria https://www.R-project.org/ (2016).

Acknowledgements

This work was supported in part by the MSK Cancer Center Support Grant/Core Grant (P30 CA008748).

Author Contributions

H.V. and E.J.S. conceived the idea and performed the experiments. E.M. and J.O.D. helped with formulating the experiments and analysis. H.V. developed and implemented the methods. E.J.S., B.Z.D. manually delineated and produced ground truth for experiments. M.S. and N.O. performed manual delineations for inter-rater study with E.J.S. H.V. and E.J.S. wrote the paper. All authors reviewed and edited the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-22980-9.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.