An investigation of the effect of fat suppression and dimensionality on the accuracy of breast MRI segmentation using U-nets.

Homa Fashandi¹, Gregory Kuling², YingLi Lu¹, Hongbo Wu², and Anne Martel^{1,2*}

¹Physical Sciences, Sunnybrook Research Institute, Toronto, Ontario M4N 3M5, Canada

²Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada

*Corresponding author: a.martel@utoronto.ca

Purpose: Accurate segmentation of the breast is required for breast density estimation and the assessment of background parenchymal enhancement, both of which have been shown to be related to breast cancer risk. The MRI breast segmentation task is challenging, and recent work has demonstrated that convolutional neural networks perform well for this task. In this study, we have investigated the performance of several 2D U-Net and 3D U-Net configurations using both fat-suppressed and nonfat suppressed images. We have also assessed the effect of changing the number and quality of the ground truth segmentations.

Materials and methods: We designed 8 studies to investigate the effect of input types and the dimensionality of the U-Net operations for the breast MRI segmentation. Our training data contained 70 whole breast volumes of T1-weighted sequences without fat suppression(WOFS) and with fat suppression(FS). For each subject, we registered the WOFS and FS volumes together before manually segmenting the breast to generate ground truth. We compared 4 different input types to the U-nets: WOFS, FS, MIXED(WOFS and FS images treated as separate samples) and MULTI(WOFS and FS images combined into a single multi-channel image). We trained 2D U-Nets and 3D U-Nets with this data, which resulted in our 8 studies (2D-WOFS, 3D-WOFS, 2D-FS, 3D-FS, 2D-MIXED, 3D-MIXED, 2D-MULTI, and 3D-MULTI). For each of these studies, we performed a systematic grid search to tune the hyperparameters of the U-Nets. A separate validation set with 15 whole breast volumes was used for hyperparameter tuning. We performed Kruskal-Wallis test on the results of our hyperparameter tuning and did not find a statistically significant difference in the 10 top models of each study. For this reason, we chose the best model as the model with the highest mean DSC value on the validation set. The reported test results are the results of the top model of each study on our test set which contained 19 whole breast volumes annotated by 3 readers fused with the STAPLE algorithm. We also investigated the effect of the quality of the training annotations and the number of training samples for this task.

Results: The study with the highest average DSC result was 3D-MULTI with 0.96 ± 0.02 . The second highest average is 2D WOFS (0.96 ± 0.03), and the third is 2D MULTI (0.96 ± 0.03). We performed the Kruskal-Wallis 1-way ANOVA test with Dunn's multiple comparison tests using Bonferroni p-value correction on the results of the selected model of each study and found that 3D-MULTI, 2D-MULTI, 3D-WOFS, 2D-WOFS, 2D-FS, and 3D-FS were not statistically different in their distributions, which indicates that comparable results could be obtained in fat-suppressed and nonfat suppressed volumes and that there is no significant difference between the 3D and 2D approach. Our results also suggested that the networks trained on single sequence images or multiple sequence images organized in multi-channel images perform better than the models trained on a mixture of volumes from different sequences. Our investigation of the size of the training set revealed that training a U-Net in this domain only requires a modest amount of training data and results obtained with 49 and 70 training datasets were not significantly different.

Conclusions: To summarize, we investigated the use of 2D U-Nets and 3D U-Nets for breast volume segmentation in T1 fat suppressed and without fat suppressed volumes. Although our highest score was obtained in the 3D MULTI study, when we took advantage of information in both fat suppressed and non fat suppressed volumes and their 3D structure, all of the methods we explored gave accurate segmentations with an average DSC on $> 94\%$ demonstrating that the U-Net is a robust segmentation method for breast MRI volumes.

1. BACKGROUND AND INTRODUCTION

Several large prospective studies have shown that annual magnetic resonance imaging (MRI) exams have a greater sensitivity than either mammography or ultrasound for women at high risk of breast cancer and, as a result, MRI based screening is now offered to women at high risk of cancer. In Ontario, screening is restricted to high-risk women who have a specific genetic mutation, or history of radiation to the chest at ages 10 to 30 years, or an estimated lifetime risk of higher than 20%. There is, however, increasing interest in offering MRI screening to women who do not fall into the currently defined high-risk groups but who, nonetheless, are at increased risk of breast cancer. For example, women with a personal history of breast cancer and women with dense breasts are known to be at higher risk. Contrast-enhanced MRI provides unique information about breast tissue composition and may, when used in combination with other established risk factors, help to improve our assessment of breast cancer risk in individual patients.

The ratio of fibroglandular tissue to fat in the breast and the amount of fibroglandular tissue that enhances after the injection of contrast agent, i.e., background parenchymal enhancement, have both been shown to be associated with risk. In order to generate accurate and reproducible quantitative measures of these parameters, an accurate breast segmentation that separates breast tissue from the chest wall is an essential first step.

Segmentation of the breast is a non-trivial task as there may not be a clear boundary between the chest wall, pectoral muscle, and breast tissue, particularly in dense breasts. The shape of the breast is very variable, and it is modified by the light compression from the breast coils and by the effects of previous surgery. Different sequences and different scanners will affect the signal intensity, and image artifacts due to incomplete fat suppression or non-uniform sensitivity of the receiver coils also have an impact on accuracy. Examples of slices from MRI breast volumes with and without fat suppression are shown in Figure 1.

In Table 2 we review existing methods of MRI breast segmentation. We have divided these methods into 4 categories, namely: conventional image processing, conventional machine learning, atlas-based and deep learning approaches. Although different image data is used in each study, we include the error rates reported for these methods. This gives us an overview of the accuracy values for different methods. Various accuracy or error measures have been used for the segmentation evaluation task. Taha et al.,¹, reported on 20 different measures used in medical image segmentation. They also defined groupings of the measures and recommended selecting metrics from different groups to avoid biased results. Moreover, Doran et al.,², recommend choosing the best model based on more than one metric. Table 1 shows the error measures reported in the breast MRI segmentation papers and provides the definition of each metric.

Deep learning approaches, which learn the features that optimally represent the data for the problem at hand, have shown great promise for medical image analysis. A survey by Litjens et al.,³ carried out in early 2017, reviews the use of deep learning for medical image analysis and lists 350 papers related to this topic. Dalmiş et al.,⁴ used a 2D U-Net architecture to segment breast MRI into three classes: breast, background and fibroglandular tissue and in another study, deep learning are used to segment the pectoral muscle in 2D MR images,⁵. In this paper, our goals were to:

- 1) Determine whether carrying out segmentation using 3D U-nets rather than 2D U-Nets improves accuracy
- 2) Compare the accuracy of segmentation methods based on T1w images with and without fat suppression
- 3) Compare the accuracy obtained with smaller, manually annotated training datasets and larger training datasets with less accurate ground truth labels

Figure 1- Sample breast MRI images show the variations in these images. The top row shows T1 weighted images with no fat suppression. Bottom row are T1 weighted images with fat suppression

Error/Performance Measure	Acronym	Definition
Dice Similarity Coefficient	DSC	$2 \times S_{GT} \cap S_A / S_{GT} + S_A $
% difference in volumes	DP	$ S_{GT} - S_A / S_{GT} \times \%100$
Delineation Sensitivity	SENS	$ S_{GT} \cap S_A / S_A $
Delineation Specificity	SPEC	$1 - S_A - S_{GT} / I_m - S_{GT} $ Where I_m is the whole image or volume.
Accuracy	ACC	$(TP + TN) / (TP + TN + FP + FN)$ TP=True positive, TN= True negative, FP = False positive, FN = False negative
Average Total Overlap	ATO	
Jaccard Coefficient	JAC	$ S_{GT} \cap S_A / S_{GT} \cup S_A $
False Positive Volume Fraction	FPVF	$N_{S_A - S_{GT}} / N_{S_{GT}}$, where N_S is the number of voxels in the segmentation(S)
Hausdorff Distance	HD	$\max_{x \in S_{GT}} \{ \min_{y \in S_A} \ x - y\ \}$
Mean Surface Distance	MSD	$\frac{1}{n_{GT} + n_A} \left(\sum_{p=1}^{n_{GT}} d(p, S_A) + \sum_{p=1}^{n_A} d(p, S_{GT}) \right)$ Where n_{GT} and n_A are number of points on the outer surfaces of ground truth and automatic segmentations, respectively. $d(p, S) = \min_{q \in S} \ p - q\ _2$, which is the distance of point p on surface S to the surface S_1 .

Table 1- Error or performance measures reported in MRI breast volume segmentation. SGT and SA represent ground truth segmentation and the automatic segmentation, respectively.

Figure 1- The 3D U-Net architecture with depth 4. The number of feature maps is written below each convolutional layer. The 2D U-net architecture is identical except that the convolutional, max pooling and up-sampling layers are replaced with 2D versions.

Table 2 -Breast volume segmentation in MRI images, *databases are different, and the error is the best's model average reported rate on the test database, DSC = Dice Similarity coefficient reported in percentages. NA= Not applicable. T1-NFS= T1-weighted, non-fat suppressed. N_T = a number of training whole breast volumes, N_V = number of validation whole breast volumes, N_T = Number of whole breast test volumes. RADGRASE = Radial gradient echo and spin echo. T1-FS = T1 weighted Fat suppressed images.

Reference	Method details	Sequence	N_T , N_V , N_T	View	Error rate * (mean $\pm$ std)
Conventional Image Processing					
Wu et al., 2013, ⁶	For the chest wall-breast boundary, edge detection followed by edge linking is applied. Thresholding and morphological opening are used for the breast-air boundary.	T1-NFS	NA, NA, 60	Sag	DSC = 0.95 ± 0.019 DP = $10.1\% \pm 3.8\%$
Lin, et al., 2013, ⁷	Manually created contour placed on the chest region and refined by k-means clustering. Registration applied to get the subject-specific chest template. Body landmarks are manually placed to segment the breast region.	T1-NFS	1, NA, 30	Axial	DP = $3.05 \pm 1.93(\%)$
Ivanovska et al., 2014, ⁸	A level set approach is proposed to simultaneously correct in-homogeneity in 2D slices and segments the image into breast and FGT. The breast volume further refined by morphological operators and detection of anatomical structures in the chest area.	T1-NFS	NA, NA, 37	Axial	DSC = 0.96 ± 0.0172 SENS = 0.98 ± 0.005 SPEC = 0.99 ± 0.002
Marrone et al., 2016, ⁹	4D DCE-MRI is converted to a single 3D volume by taking the minimum values along the time axis. Otsu's threshold is applied to the volume, and anatomical structures are identified to enclose the breast volume. Morphological operators and Fuzzy C-Means clustering are used to refine the results.	DCE-MRI, T1-NFS	NA, NA, 30	Axial	DSC = 92.66 ± 1.48 SENS = 98.44 ± 0.52 SPEC = $95.83 \pm .49$
Jiang, et al., 2017, ¹⁰	Dynamic programming is used to trace the chest wall in the gradient transformed 2D slices. Dynamic programming also used to delineate the breast skin line.	T1-FS	NA, NA, 20	Axial	DSC = 0.95 ± 0.12 DP = $8.4 \pm 2.4 (\%)$
Conventional Machine Learning					
Ni et al., 2008, ¹¹	Initial segmentation of the breast area is obtained by manually localizing body landmarks, followed by Fuzzy C-means clustering. User interventions are required throughout the process.	T1-NFS	NA, NA, 11	Axial	Inter-observer reproducibility
Ertas et al., 2017, ¹²	Bias-corrected fuzzy c-means algorithm, morphological operations, and connected components analysis are utilized to segment the breast region.		50, NA, 32	Axial	RO = 0.94 ± 0.05 SENS = 0.97 ± 0.03 FPVF = 0.04 ± 0.06
Rosado-Toro, 2015, ¹³	RADGRASE sequence images contain water and fat image. k-means clustering applied to the water image. Pectoral muscle is extracted from the fat image using dynamic programming to refine the k-means segmentation	RADGRASE	NA, NA, 14	Axial	DSC = 0.92 ± 0.047
Martel, et al., 2015, ¹⁴	Random forest classifier to identify candidate edges in the volume and then applies a Poisson reconstruction to define the three-dimensional volume	T1-NFS	7, NA, 1	Sag	DSC = 0.96 ± 0.02 AHD < 0.1mm
Atlas-Based Techniques					
Gubern-Merdia et al., 2015, ¹⁵	Breast volume is preprocessed by N3 bias field correction and then normalized. The sternum is localized to define a volume of interest (VOI). Probabilistic Atlas built on a single volume is used for the breast-chest wall surface detection. Breast air boundary is determined by a region growing method.	T1-NFS	1, NA, 27	Axial	DSC = 0.94 ± 0.04 ATO = 0.96 ± 0.02
Ortiz et al., 2012, ¹⁶	Phase congruency detects the edges, and the Poisson surface reconstruction extracts the breast volume. Breast volume is further refined using an atlas. Atlas was built on all the images	Dixon	NA, NA, 409	Sag	DSC= 0.88 ± 0.04 JAC = 0.79 ± 0.06
Khalvati, et al., ¹⁷	Intensity-based atlas is generated for the inter-sequence processing; phase congruency map used to create Atlas for intra sequence processing. Atlas built on 350 volumes. [*] Inter-sequence images. [^] Intra-sequence images	Dixon, T1-NFS	NA, NA, 50 [*] NA, NA, 17 [^]	Sag	DSC ^{+*} = 0.94 ± 0.04 DSC ^{+^} = 0.87 ± 0.065
Deep Learning					
Dalmış et al., 2017, ⁴	U-Net to classify the 2D MRI slices into the breast, non-breast, and fibro-glandular tissue ** error rates are reported for the whole data set including the training images	T1-NFS	39,5,22	Axial	DSC = 0.944 ± 0.026 **

2. METHODS

a. Network architecture

The original U-Net introduced in ¹⁸ consists of a contracting encoder path or analysis path and a successive expanding decoder path also called a synthesis path. In the original U-net implementation, all convolution, max pooling, and upsampling operations were carried out in 2D. This was later extended to a 3D U-NET in ¹⁹. Each analysis layer consists of two sets of convolutional layers and an activation layer followed by a max pooling layer. The number of convolutional feature maps doubles when moving from one layer to the next in the analysis path. In the synthesis path, the up-sampling layers are used to increase the resolution moving from one layer to the next. Also, the number of convolutional feature maps decreases in half moving up in the synthesis path. The connections from layers of equal resolution in the analysis path to the synthesis path provide detailed information required for the up-sampled output. The depth of the network is defined as the number of layers in the analysis path. Based on this definition, the network shown in Figure 2 has a depth of 4. We added dropout layers to the synthesis path and used rectified linear units (ReLU) as our activation function and Sigmoid at our last activation layer. Loss function based on the Dice similarity coefficient has been used. We used Keras 2.1.2 with TensorFlow 1.4.1 running on a Nvidia GeForce TITAN X GPU with 12 Gb memory to train and test the models.

b. Data Set

Bilateral T1-weighted (T1W) MR images were acquired on a 1.5-T magnet (Signa, General Electric Medical Systems, Milwaukee, WI) using a dedicated breast coil in a sagittal orientation with an average resolution of 0.39 mm by 0.39 mm in plane and 3.0 mm between slices. T1W images with fat suppression (FS) and without fat-suppression (WOFS) were both acquired before contrast agent was administered and were co-registered using Elastix^{20,21}. The whole breast volumes were then divided into left and right breasts and resampled to 2 mm isotropic voxel size. Finally, by using cropping and/or padding, all the volumes were resized to 64×128×128 pixels.

We have used 4 different types of MRI data to train our network for comparison purposes:

- WOFS: Pre-contrast T1W volumes without fat suppression
- FS: Pre-contrast T1W volumes with fat suppression
- MIXED: both WOFS and FS volumes are used
- MULTI: a 2-channel volume is generated using the aligned WOFS and FS volumes

The input to the 3D U-Net is 3D volumes for WOFS, FS and MIXED data and 2-channel volumes for the MULTI. The input to the 2D U-NET is 2D slices for WOFS, FS and MIXED data and 2-channel slices for MULTI. We perform eight studies on the combination of input data (i.e., WOFS, FS, MIXED, and MULTI) and U-Net operation types (i.e., 3D U-Net, 2D U-NET). For instance, the 3D-WOFS study refers to the 3D U-Net trained on WOFS data or 2D-MULTI refers to 2D U-Net trained on 2-channel slices of WOFS and FS images.

The number of whole breast volumes we used for the train, validation and test sets are shown in Table 3. For the 2D-MULTI, 3D-MULTI, 2D-MIXED and 3D-MIXED, the number of volumes is twice that of the other studies as both the FS and WOFS volumes are used.

Table 3- Information on a number of whole breast volumes used for training, validation and test purposes. All the sets are disjoint and do not have patient overlap. For each case, registered WOFS and FS volumes are available.

Data Set	# whole breast volumes	# Readers	Note
Training Data Set	70	1	%33 of the data used for validation during the training
HP Validation Data	15	1	
Test Data Set	19	3	

c. Ground Truth Generation

The process of generating ground truth (GT) segmentations is time-consuming and expensive, especially for medical images. To facilitate the GT generation in this application, we build a semi-automatic pipeline to generate the data. We

first create breast masks of our dataset by applying a previously developed method in our lab,¹⁴ to the WOFS images. The method uses a random forest classifier to identify candidate edges in the volume and then applies a Poisson reconstruction to define the three-dimensional volume. The approach produces a good separation of the air-breast boundaries but fails in the breast-chest wall boundary in some of the volumes. Moreover, the method includes the fatty region below the breast and the regions beyond the lateral boundaries of the breast. We manually edited the automatic segmentation using ITK-SNAP,²² and the co-registered WOFS and FS images were available to guide this process. Then our proposed algorithm divides the breast volume into the left and right breasts. By using resampling, cropping, and padding, all the volumes are resized into isotropic volumes of 64×128×128 pixels. It then removes any remaining disconnected regions from the 2D slices, and cuts the fat tissue below the breast and beyond the lateral boundaries of the breast. Figure 3 shows the proposed pipeline. The dashed line block is the only manual step of the proposed pipeline.

Figure 2- Semi-automatic ground truth generation pipeline. The automatic steps are shown in solid lines, and the manual step is shown in dashed lines. *The method is previously developed in our lab,¹⁴. **Manual editing is mainly happening in the breast-chest wall boundary. *** This step was removed for the test batch creation for one of the readers, as the reader removed the fat region below the breast manually

To remove the fatty area under the breast, we find the concave points on 2D slices of the breast mask and select the one on the left side of the contour that has the largest concavity measure. We define the concavity measure by the distance of the concave point on the contour and the convex hull of the contour. Once all the cutting points are chosen in all the slices, they are smoothed, and the fat region below the breast is cut. Figure 4 shows the process for a sample volume. Lateral slices of the mask volume are determined based on the size of the mask in each slice. We defined a discrete function, $y = f(x)$, where y is the size of the breast mask in slice number x . We used the maxima and minima of the derivative of f , $f' = \partial f / \partial x$, to detect the lateral boundaries of the breast. A few samples of the random forest generated masks, and the output of our ground truth generation pipeline are shown in Figure 5.

Figure 3 - Removing the fat tissue below the breast in the breast masks generated by the Random Forest Approach. (a) Volume generated by the random forest method,¹⁴ (b) one slice of the volume shows a small disconnected region. (c) the small disconnected region is removed from the 2D slice, and the concave points are shown with red circles. The cutting point is automatically selected from the set of concave points and shown with a blue arrow. (d) The fat region below the breast is removed based on the cutting points. [Color figure can be viewed at wileyonlinelibrary.com]

The 19 whole breast volumes that were used as the final test set were each annotated by three readers and the STAPLE algorithm was used to combine the segmentations to get the final ground truth masks,²³. The 85 training and validation sets were each annotated by just one of the three readers.

Figure 4- Automatic Removal of the fatty regions below the breast and the regions beyond the lateral boundaries of the breast is part of the proposed ground truth generation pipeline. (Top-row) The masks are generated with the random forest method, ¹⁴ (Bottom-row) The output of the steps 4 and 5 of the pipeline shown in Figure 3. [Color figure can be viewed at wileyonlinelibrary.com]

d. Hyper-Parameter Search

We limited our search of hyperparameters to the depth of the network, dilation rate for the dilated convolution, the type of the optimizer, and dropout rate. The values for each of these hyperparameters are shown in Table 4. For each of our studies (2D-WOFS, 2D-FS, 2D-MIXED, 2S-MULTI, 3D-WOFS, 3D-FS, 3D-MIXED, and 3D-MULTI), we build models from our hyperparameter search space and train the models on the appropriate training dataset. We chose a separate dataset of 15 whole breast volumes and will refer to this as the hyperparameters validation set (HP-Validation set).

For each of the studies, the trained models on the grid search hyperparameter space were tested on the HP-validation set. After ranking the DSC values for each study, a Kruskal-Wallis 1-way ANOVA test was used to evaluate the differences between the trials ²⁴. The outcome of the test indicated significant differences among the trials in each study. We then perform a Dunn's multiple comparison tests using Bonferroni p-value correction and found no significant difference between the median of the top 10 grid search results for each study. Therefore, we chose the model with the highest average DSC as an output of our hyperparameter optimization for each study. Each plot in Figure 6 shows the mean DSC values of each trial per study on the HP-validation set. The color bar shows the hyperparameter values.

Table 4- Hyperparameter search space

Hyperparameter	Values	Comment
Network's Depth	3, 4, 5	Number of layers in the analysis path of the U-Net
Dilation Rate	1, 2	Dilation rate for dilated convolution
Dropout Rate	0.0, 0.2, 0.5	Randomly dropping input units at each update during training time
Optimizer	ADAM, SGD	SGD (Stochastic Gradient Descent) with Nesterov momentum, ²⁵ . ADAM optimizer, adaptive moment estimation, ²⁶

Figure 5-Mean DSC measure for individual trials on HP-Validation set. Each plot shows the grid search results in each of our studies.

3. EXPERIMENTS

a. Test Set Results

Table 5 shows the performance/error measure of the best models in each study. Figure 8 shows plots of our test set. Each plot represents the top model of each study, and each point is one test volume (left or right breast in sagittal view). The test volumes are sorted based on the breast volumes, and the size of the points are related to the breast volume of the test set. The histogram of the DSC values is shown on the right side of each plot. The highest DSC is obtained for the 3D-MULTI study with 0.96 ± 0.02 . We performed the Kruskal-Wallis 1-way ANOVA test with Dunn's multiple comparison tests using Bonferroni p-value correction on the results of the selected model of each study and found that 3D-MULTI, 2D-MULTI, 3D-WOFS, 2D-WOFS, 2D-FS and 3D-FS models were not statistically different in their distributions. Therefore, similar results are expected by all of these studies. Consequently, availability of the training data may be the decisive factor in adopting one of these studies at different institutions. The results of the paired comparisons are graphed in Figure 7. The squares in black are pairs where the medians are significantly different (therefore, rejecting H_0 hypothesis), and the white ones are pairs that do not have significantly different medians. It is interesting to note that the median DSC of our top performer (i.e., 3D-MULTI) is statistically higher than median DSC of the 3D-MIXED and 2D-MIXED models. This suggests that the networks trained on one sequence of data or multiple sequences of data arranged in multi-channel images, perform better than the models trained on the mixed set of sequences.

For the MIXED models, we were interested in testing whether there is a significant difference in accuracy for the WOFS and FS volumes. To investigate this, we used a Wilcoxon test to compare the DSC scores of FS, and WOFS volumes segmented using the 3D-MIXED model and found a small but statistically significant improvement in the DSC for the WOSF volumes compared to the FS volumes, (p-value = $1.76e-06 < 0.05$)

Table 5- Mean error values on the Test set are reported for the selected best models.

U-Net	Data Type	DSC	SENS	Mean Surface Distance(mm)	JAC
2D	WOFS	0.96 ± 0.03	0.95 ± 0.03	1.36 ± 0.57	0.92 ± 0.05
	FS	0.94 ± 0.07	0.95 ± 0.03	1.56 ± 0.86	0.90 ± 0.09
	MIXED	0.95 ± 0.05	0.95 ± 0.03	1.52 ± 0.85	0.91 ± 0.08
	MULTI	0.95 ± 0.04	0.95 ± 0.03	1.54 ± 0.7	0.91 ± 0.06
3D	WOFS	0.95 ± 0.02	0.93 ± 0.04	1.6 ± 0.53	0.91 ± 0.04
	FS	0.95 ± 0.04	0.94 ± 0.03	1.66 ± 0.63	0.90 ± 0.06
	MIXED	0.95 ± 0.02	0.94 ± 0.03	1.56 ± 0.53	0.91 ± 0.04
	MULTI	0.96 ± 0.02	0.96 ± 0.02	1.3 ± 0.43	0.92 ± 0.04

Figure 6-The Kruskal-Wallis test on the selected model of each study. The black means the H_0 is rejected, i.e., there is a significant difference in DSC, and white means the H_0 is not rejected.

Figure 7- Plots for the DSC values of the test volumes using the grid search best models. Histograms of the DSC values are shown on the y-axis on the right of each plot. The test volumes are sorted based on breast volumes. The number of test volumes in 3D-MIXED and 2D-MIXED is twice that for the rest of the studies because we used both WOFS and FS volumes to test the models. For the 2D-MULTI and 3D-MULTI, the FS and WOFS volumes are combined to form a multi-channel image. Therefore the number of test samples is the same as other studies.

b. Size of the Training Set

We varied the size of the training set to see the effect of the number and quality of the ground truth segmentations on performance. For this experiment, we chose the 3D-WOFS study. We considered high-quality training data to use the manually edited segmentations. The low-quality segmentations were those obtained by our previously developed Random Forest method ¹⁴. Both sets of segmentations were post-processed using the pipeline shown in figure 3. Using the hyperparameters of the best model of the 3D-WOFS study, we trained the 3D U-Net with a different number of training instances. The results on the test dataset are shown in Figure 9. We observed that, for the high-quality segmentations, the results of the models trained on 49 and 70 volumes are not significantly different (p-value = 0.188 > 0.05, Mann-Whitney U-Test). The quality of the segmentations used in training does, however, influence performance. We found that using manually edited segmentations in training gave significantly better results than using the poorer quality RF segmentations when the number of training volumes was the same. We also trained the model with 686 volumes using automatically generated low quality ground truth segmentations and found that accuracy was better than that obtained using 35 manually segmented volumes (p-value = 0.0039 < 0.05) but worse than that obtained using 70 manually segmented volumes (p-value = 0.0038). There was no significant difference between results obtained using 49 manually segmented volumes and 686 low-quality ground truth segmentations (p-value = 0.084).

Figure 8- Effects of the size of the training set and the quality of training set on the performance of the model. The model’s architecture is the top model of the 3D-WOFS study. Error bars represent the standard deviation of DSC scores and, as they are symmetrical, the lower bars for the high-quality set and the upper bars for the lower quality set have been omitted for clarity.

c. Sample Outputs

Figure 10 and Figure 11 show samples of the predicted masks overlaid on the ground truth masks for different studies. For most of our test cases, the errors happen on the lateral boundaries of the breast. As seen in the plots of Figure 8, the lowest DSC values are consistently obtained in the patient with the smallest breast volumes in our test dataset. Figure 12 shows sample slices from these test samples. The errors are mostly happening in the lateral boundaries of the breast and the fatty region below the breast. As seen in Figure 12-(h), our ground truth generation pipeline failed for the right breast of this sample as it did not detect the best concave points due to the size of the breast and therefore cut the fatty region below the breast extensively. The performance of the 3D-MULTI model was, however, still acceptable.

Figure 9- Sample output of different studies. The STAPLE generated ground truth mask is green, the predicted mask is red, and the overlap between the two is yellow. For demonstration purposes, the output mask of the 3D-MULTI study is shown on the WOFS image. The input to the 3D-MIXED model is the FS volume.

Figure 10-Sample output of different studies. The STAPLE generated ground truth mask is green, the predicted mask is red, and the overlap between the two is yellow. For demonstration purposes, the output mask of the 3D-MULTI study is shown on the WOFS image. The input volume to the 3D-MIXED model is the FS volume.

Figure 11-Output of 3D-MULTI study on the test samples with the lowest DSC values. The samples belong to one subject. The masks are overlaid on the corresponding WOFS images. The number below each image is the slice number. As shown in these images, most of the errors occur in the lateral slices of the breast and the region below the breast. Green is the output of STAPLE algorithm on our three readers annotated masks. Red is the output of the model. Yellow is the overlap between the two.

The detection of breast and chest wall is more challenging in dense breasts than fatty breasts. The image in Figure 13-(a) shows a dense breast in our test set. The results obtained on this image showed some error in the breast-chest wall region. As seen in Figure 13-(d), the generated mask with 3D-FS study outperformed the masks generated by other 3D studies.

Figure 12- Sample results on a dense breast. The result of 3D-Multi is overlaid on the WOFS slice. Green is the ground truth mask; Red is the predicted mask and Yellow is the overlap between the two. The input to the 3D-MIXED model is the FS volume.

4. DISCUSSION AND CONCLUSION

Our results confirm the observation made by others^{19, 4} that the U-Net architecture is able to produce good results even when the number of training samples is small if it is used in a restricted domain of images. Although there is wide variation in the shape and characteristics of MRI breast images, the variations are not as extensive as other domains, for instance, ImageNet dataset²⁷. We demonstrated that with a relatively small dataset of high quality annotated images, good performances are achievable.

The results of the hyperparameter tuning shown in Figure 6 show that model performance is consistent across a wide range of parameters but that it is still important to carry out this step as in all the studies there was a significant difference between the best and worst performing parameter sets.

Our study did not show a significant difference in performance for FS and WOFS images which is encouraging as both sequences may not always be available. We did observe that the FS images were occasionally more accurate in dense breasts whilst the WOFS images were slightly better in less dense breasts; this is likely due to the contrast difference at the chest wall. It is possible that using information about breast density to combine the output from an ensemble of U-nets could be used to improve accuracy further.

We did not show any significant advantage in using a 3D U-net rather than a 2D U-net. We had to downsample all images to a voxel size of 2mm to fit batches of 4 (64×128×128) volumes into GPU memory for the 3D U-Net. With a 2D U-net, it would be possible to retain a higher spatial resolution, and this may be important if the goal is to segment fibroglandular tissue as well as the whole breast.

We also looked at how many training datasets we needed and found that, for this particular study, there was no advantage to using more than 49 manually segmented volumes. We had previously developed a segmentation algorithm based on random forests that worked well on WOFS images, but which did not do well on FS images. We tested whether we could use the (imperfect) output from this algorithm to train a new, more accurate U-net and found that, given a sufficiently large training set, we were able to match the performance of a U-net trained using up to 49 manually edited segmentations. This suggests that a “bootstrapping” approach of this kind can be viable in the right circumstances.

In this study we only segmented images acquired using 2 difference acquisition protocols and a 1.5 T MRI scanner. In practice, there is a much wider variation in acquisition protocols, breast coils, and scanner type. Further work is needed to determine the best way to handle this variation. One approach would be to train a network using a variety of imaging protocols. However, our results suggest that there may be advantages in training a U-net on a homogeneous dataset rather than trying to train on a mixture of sequence types. It may be that if we had increased the capacity of the MIXED model by adding more convolutional layers to cope with the increased variability in the data, then we could have achieved more accurate results. Another possible approach suggested by Drozdal et al.²⁸ is to use a low capacity fully convolutional network as a pre-processor to normalize the MRI data first. Alternatively, since we were able to achieve a DSC of 92% on just 14 training images, we could simply use domain adaptation and retrain a U-net using a small number of annotated volumes acquired using the desired sequence.

ACKNOWLEDGMENTS

This research received support from the Sunnybrook Research Institute, through funding from the Federal Economic Development Agency for Southern Ontario (FedDev Ontario). This work was also funded in part by the Canadian Breast Cancer Foundation.

5. References

1. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *Taha Hanbury BMC Med Imaging*. 2015;15. doi:10.1186/s12880-015-0068-x.
2. Doran SJ, Hipwell JH, Denholm R, et al. Breast MRI segmentation for density estimation: Do different methods give the same results and how much do differences matter? *Med Phys*. 2017;44(9). doi:10.1002/mp.12320.
3. Litjens G, Kooi T, Bejnordi BE, et al. A Survey on Deep Learning in Medical Image Analysis. <https://arxiv.org/pdf/1702.05747.pdf>. Accessed June 5, 2018.
4. Dalmiş MU, Litjens G, Holland K, et al. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Med Phys*. 2017;44(2):533-546. doi:10.1002/mp.12079.
5. Moeskops P, Wolterink JM, van der Velden BHM, et al. Deep Learning for Multi-Task Medical Image Segmentation in Multiple Modalities. *Miccai-2016*. 2017:478-486. doi:10.1007/978-3-319-46723-8_55.
6. Wu S, Weinstein SP, Conant EF, Schnall MD, Kontos D. Automated chest wall line detection for whole-breast segmentation in sagittal breast MR images. *Med Phys*. 2013;40(4):1-12. doi:10.1118/1.4793255.
7. Lin M, Chen JH, Wang X, Chan S, Chen S, Su MY. Template-based automatic breast segmentation on MRI by excluding the chest region. *Med Phys*. 2013;40(12):1-10. doi:10.1118/1.4828837.
8. Ivanovska T, Laqua R, Wang L, Liebscher V, Völzke H, Hegenscheid K. A level set based framework for quantitative evaluation of breast tissue density from MRI data. *PLoS One*. 2014;9(11):1-19. doi:10.1371/journal.pone.0112709.
9. Marrone S, Piantadosi G, Fusco R, Petrillo A, Sansone M, Sansone C. Breast segmentation using Fuzzy C-Means and anatomical priors in DCE-MRI. *Proc - Int Conf Pattern Recognit*. 2017:1472-1477. doi:10.1109/ICPR.2016.7899845.
10. Jiang L, Lian Y, Gu Y, Hu X, Li Q. Fully automated segmentation of whole breast in MR images by use of dynamic programming. *Med Phys*. 2017;44(6). doi:10.1117/12.2043343.
11. Nie K, Chen J-H, Chan S, et al. Development of a quantitative method for analysis of breast density based on three-dimensional breast MRI. *Med Phys*. 2008;35(12):5253-5262. doi:10.1118/1.3002306.
12. Ertas G, Doran SJ, Leach MO. A computerized volumetric segmentation method applicable to multi-centre MRI data to support computer-aided breast tissue analysis, density assessment and lesion localization. *Med Biol Eng Comput*. 2017;55:57-68. doi:10.1007/s11517-016-1484-y.
13. Rosado-Toro JA, Barr T, Galons JP, et al. Automated Breast Segmentation of Fat and Water MR Images Using Dynamic Programming. *Acad Radiol*. 2015;22(2):139-148. doi:10.1016/j.acra.2014.09.015.
14. Martel AL, Gallego-Ortiz C, Lu Y. Breast segmentation in MRI using Poisson surface reconstruction initialized with random forest edge detection. *Med Imaging 2016 Image Process*. 2016;SPIE 9784:97841B. doi:10.1117/12.2214416.
15. Gubern-Mérida A, Kallenberg M, Mann RM, Martí R, Karssemeijer N. Breast segmentation and density estimation in breast MRI: A fully automatic framework. *IEEE J Biomed Heal Informatics*. 2015;19(1):349-357. doi:10.1109/JBHI.2014.2311163.
16. Ortiz CG, Martel AL. Automatic atlas-based segmentation of the breast in MRI for 3D breast volume computation. *Med Phys*. 2012;39(10):5835-5848. doi:10.1118/1.4748504.
17. Khalvati F, Gallego Ortiz C, Balasingham S, Martel AL. Automated Segmentation of Breast in 3D MR Images Using a Robust Atlas. *IEEE Trans Med Imaging*. 2014;34(1):116-125. doi:10.1109/TMI.2014.2347703.
18. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Med*

19. İçek OÇ, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2016*. Springer International Publishing; 2016:424-432. <https://lmb.informatik.uni-freiburg.de/Publications/2016/CABR16/cicek16miccai.pdf>. Accessed June 4, 2018.
20. Mehrabian H, Richmond L, Lu Y, Martel AL. Deformable registration for longitudinal breast MRI screening. *J Digit Imaging*. 2018. doi:10.1007/s10278-018-0063-1.
21. Klein S, Staring M, Murphy K, Viergever MAMA, Pluim JPW. elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Trans Med Imaging*. 2010;29(1):196-205. doi:10.1109/TMI.2009.2035616.
22. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*. 2006;31(3):1116-1128. doi:10.1016/j.neuroimage.2006.01.015.
23. Warfield SK, Zou KH, Wells WM, Zou KH. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. *IEEE Trans Med Imaging*. 2004;23(7). doi:10.1109/TMI.2004.828354.
24. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *Source J Am Stat Assoc*. 1952;47(260):583-621. <http://www.jstor.org/stable/2280779>. Accessed June 11, 2018.
25. Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov Math Dokl*. 1983;27(2):372-376.
26. Kingma DP, Ba JL. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. In: *International Conference on Learning Representations (ICLR)*. ; 2015. <https://arxiv.org/pdf/1412.6980.pdf>. Accessed June 11, 2018.
27. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2009:248-255. doi:10.1109/CVPR.2009.5206848.
28. Drozdal M, Chartrand G, Vorontsov E, et al. Learning normalized inputs for iterative estimation in medical image segmentation. *Med Image Anal*. 2018;44(2):1-13. doi:10.1016/j.media.2017.11.005.