Leveraging Epistemic Uncertainty to Improve Tumour Segmentation in Breast MRI: An exploratory analysis

Smriti Joshi[a], Richard Osuala[a], Lidia Garrucho[a], Apostolia Tsirikoglou[b], Javier del Riego[c], Katarzyna Gwoździewicz[d], Kaisar Kushibar[a], Oliver Diaz[a], and Karim Lekadir[a]

[a]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain
[b]Karolinska Institutet, Sweden
[c]Women's Imaging. Department of Radiology, Institut d'Investigació i Innovació Parc Taulí (I3PT). Universitat Autònoma de Barcelona, Sabadell, Spain
[d]Medical University of Gdansk, Poland

ABSTRACT

Medical image segmentation has improved with deep-learning methods, especially for tumor segmentation. However, variability in tumor shapes, sizes, and enhancement remains a challenge. Breast MRI adds further uncertainty due to anatomical differences. Informing clinicians about result reliability and using model uncertainty to improve predictions are essential. We study Monte-Carlo Dropout for generating multiple predictions and finding consensus segmentation. Our approach reduces false positives using per-pixel uncertainty and improves segmentation metrics. In addition, we study the correlation of model performance to the perceived ease of manual segmentation. Finally, we compare the per-pixel uncertainty with the inter-rater variability as segmented by six different radiologists. Our code is available at https://github.com/smriti-joshi/uncertainty-segmentation-mcdropout.git.

Keywords: Uncertainty estimation, Breast imaging, MRI, Segmentation

1. INTRODUCTION

Breast cancer is now the most common cancer worldwide.[1] It is responsible for almost 30% of all cancers in women. The latest reported statistics in 2023 indicate that the rate of incidence increases by 0.5% every year in the US. However, the increasing awareness and screening programs have decreased the mortality rates by 43% from 1989 to 2020. With the emergence of deep learning-based techniques, we are observing vast progress in the development of automatic methods for the diagnosis and treatment of this disease. One such task is the automated segmentation which is critical for expediting the localization of the tumor and numerous downstream tasks such as the use of radiomics for patient's treatment response[2–5] or recurrence prediction.[6]

In the literature, it is common to assume a one-to-one mapping between the tumor and the corresponding automated segmentation. However, many emerging studies[7–9] show high inter-rater variability in the manual segmentation of tumors. This is also demonstrated in our work in Figure 1. Therefore, predicting one segmentation with a deep learning model, without accounting for any ambiguity, gives a false sense of certainty to the end-user. Several techniques have been proposed in the literature to quantify the uncertainty including Ensembling[10,11] Bayesian Neural Networks[12] and Monte-Carlo Dropout[13] (MCDropout). However, we rarely analyze the sources of these uncertainties and whether it concurs with the perceived difficulty of manual segmentation in clinical practice. In this work, we provide a deep analysis of MCDropout as a method to capture real-world uncertainty in segmentation. More specifically, our contribution is as follows:

1. Pixel-wise uncertainty estimation of predicted segmentation and its comparison with real inter-annotator variability.
2. Comparison of different combinations of the model predictions to obtain highly certain consensus segmentation and boost the performance by reducing false positives.
3. Analysis of correlation between segmentation performance and perceived difficulty of manual segmentation of breast tissue in magnetic breast resonance images (MRI).

2. METHOD

This section elaborates further on the dataset and methods used for training the segmentation network, estimating uncertainty, and using it to obtain improved segmentations.

2.1 Dataset

We use the publicly available Duke-Breast-Cancer MRI dataset[14] for the segmentation task. The description of the dataset and corresponding annotations is as follows:

Imaging data: This dataset contains MRI scans of 922 patients, with one non-fat saturated T1 sequence and three to four phases of dynamic contrast-enhanced (DCE) T1 sequences per case. In this work, we only use the first phase for the DCE series for 254 cases for which the segmentation masks were available. The MRI have voxel dimensions of 448 x 448 or 512 x 512, with different numbers of slices along the third axis and variable pixel resolutions. These images are acquired from two vendors with magnetic field strength of 1.5 T and 3 T, respectively.

Segmentation masks: The ground truth (GT) segmentations are provided by the authors of a related work[2] and were segmented for the purpose of extracting radiomics features from the tumor to predict its response to neoadjuvant chemotherapy. The MRI (DCE phase one) were automatically segmented with fuzzy means algorithm in MATLAB, corrected by a medical physicist and visually assessed by a radiologist (with 20 years of experience).

Inter-observer variability experiment: We requested six breast radiologists to segment one mass and one non-mass tumor to analize their differences. This task was performed using the software tool ITKSnap (version 3.8.0).

Easy vs. Hard tumors: We requested two breast radiologists, with experience of 10+ years, to rank the ease of delineating the boundaries of the tumor on a scale of one to five, where one and five indicate very easy and very hard respectively. Furthermore, with each case, we also requested them to describe the tumor in various aspects, e.g. (i) is the tumor localized?, (ii) is the boundary of the tumor defined or fuzzy? and (iii) is there uneven enhancement within or around the tumor? We use this information to analyze the performance and limitations of the method in Section 3.

2.2 Evaluation

Overlap-based metrics are widely used in the literature for evaluating segmentation tasks due to ease of interpretability and usage. However, they are shape-unaware and limited when the structures are small and show high size variability.[15] Therefore, it is recommended to complement them with a boundary-based metric. We use a combination of dice coefficient (overlap-based) and modified Hausdorff distance (boundary-based) to evaluate our method. Finally, we use Wilcoxon signed rank test to verify the statistical significance of our results. They are detailed in the following sections:

Dice coefficient indicates the extent of spatial overlap between the prediction and the ground truth. The range of this metric is [0, 1], where 0 indicates no overlap and 1 indicates complete overlap. This can be mathematically expressed as:

$$Dice = 2\frac{\|A \cap B\|}{\|A\| + \|B\|}, \tag{1}$$

where A and B represents the predicted and ground truth segmentations, respectively.

Modified Hausdorff distance(MHD) indicates how close is a point in P to a point in Q, where P and Q are the coordinates of the contours of the prediction and GT segmentations. This can be mathematically expressed as:

$$MHD = max(h(P,Q), h(Q,P)), \tag{2}$$

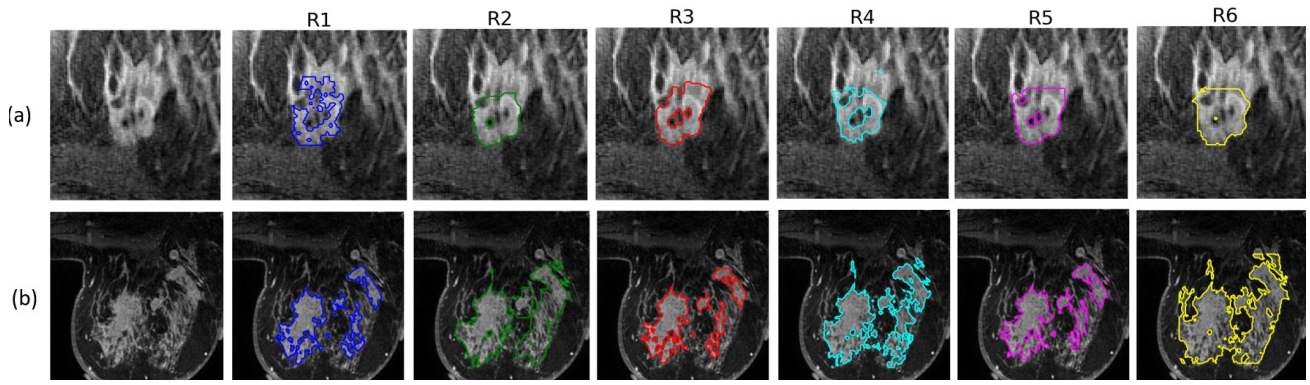

Figure 1. Annotations by six different radiologists for (a) mass and (b) non-mass tumors.

$$h(P,Q) = \frac{1}{M} \sum_{p_i \in P} \underset{q_j \in Q}{min}\, d(p_i; q_j), \tag{3}$$

where $d(p_i; q_j)$ represents the Euclidean distance between two points p_i and q_j. For this metric, lower values indicate a higher level of agreement between the prediction and GT segmentations.

Wilcoxon signed-rank test[16] is used to evaluate the statistical significance of our results. It is a non-parametric method that tests if two paired samples come from the same distribution. The minimum number of samples required is 6. In our analysis, we use it to test a maximum of 30 and a minimum of 10 cases. If the p-value is less than 0.05, then we consider the result statistically significant.

2.3 Segmentation

To obtain the baseline method for the tumor segmentation, we use a 3D UNet[17] with dropout inspired by the nnunet framework.[18] After each convolutional block, we added a 3D dropout layer with a probability of 0.2. The value $p = 0.2$ is chosen empirically, $p = 0.3$ showed similar results while $p = 0.5$ overfit on the training dataset. We do not use any post-processing techniques (e.g. selecting the largest prediction) or model ensembling (of 2D, 3D low resolution, and 3D networks) inherent in the framework so that we can fairly evaluate the utility of incorporating uncertainty-based post-processing.

Train-test split: We have 153 cases for training, 38 cases for validation, and 30 cases for testing.

Preprocessing: Before feeding the data into the preprocessing pipeline of the nnunet framework, we convert the data to NIfTi from DICOM and perform bias field correction.[19] Furthermore, the GT in our dataset only contains a segmentation mask for the main lesion. As a result, the segmentations for bilateral and multifocal cases are not available. To reduce the ambiguity in the training set, we excluded the multifocal cases (33 cases) and use only single breast for which the tumor is segmented.

2.4 Uncertainty Estimation

The dropout layers[20] are commonly used to reduce overfitting in models by multiplying the output of each neuron by a binary mask that is drawn following a Bernoulli distribution, randomly setting some neurons to zero in the neural network, during training time. During inference, the full model without dropouts is used to predict the result. Gal and Ghahramani[13] showed that using dropout during test-time can act as an approximation of probabilistic Bayesian models in deep Gaussian processes and can help to quantify the uncertainty of the models. Following this idea, we activated the dropout layer during inference to drop neurons with a specified probability and predict a segmentation mask along with pixel-wise probabilities. This was repeated as many times as the number of distinct predictions were required.

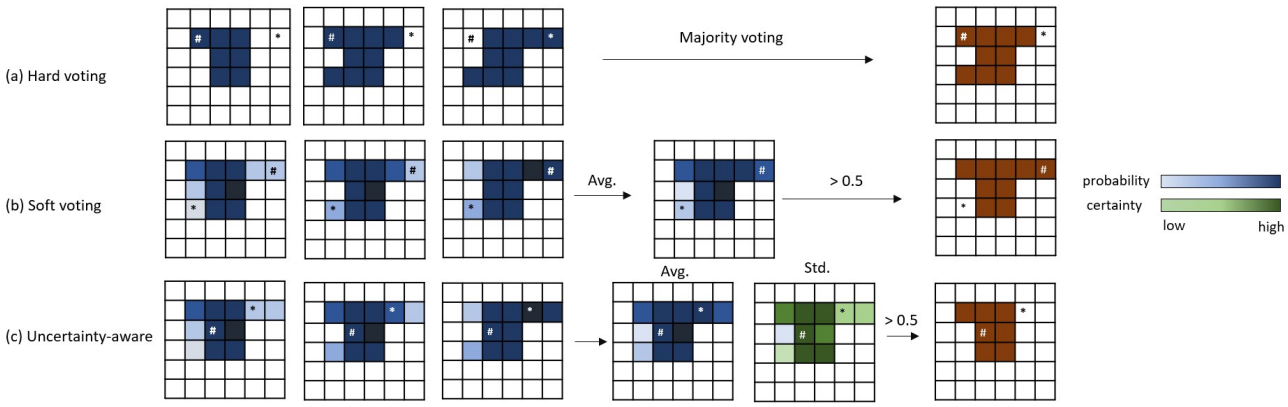

Figure 2. Method of combining predictions obtained from n different runs of MCDropout. The above example demonstrates these techniques with n=3. In each case, two pixels are marked for reader's understanding where '#' and '*' represent the pixels which are classified as foreground and background respectively in the resulting prediction. The lighter to darker blue indicates low to high predicted probability per pixel. The lighter to darker green indicates high (low) to low (high) standard deviation (certainty).

2.5 Combining predictions with high certainty

As explained in the previous section, we use MCDropout to obtain n different predictions for the same case. To combine these predictions and obtain the final segmentation, we use three different techniques, as demonstrated in Figure 2. In *hard voting*, if the pixel is predicted as foreground in the majority of the models, then it is assigned as foreground in the final prediction. In *soft voting*, predicted probabilities are averaged over n predictions per pixel. If the average is more than 0.5, the pixel is assigned foreground in the final segmentation. In *uncertainty-aware voting*, the standard deviation is also calculated per pixel along with the average pixel probability. If the average probability is more than 0.5 and the standard deviation is lower than a specified threshold, then the pixel is assigned as foreground in the final segmentation.

3. RESULTS

3.1 Inter-annotator segmentations vs. network outputs

In this section, we analyze the sources of variability in different segmentations in manual as well as automated annotations and evaluate if they model similar behaviour.

3.1.1 Variability in real practice:

Figure 1 shows segmentations of one mass and one non-mass tumor by six different radiologists (Rs). Usually, tumors present as high-intensity areas due to the injection of contrast agent during imaging. Considerable differences between the segmentations are observed and can be divided into the following categories:

(a) **Coarse vs. fine:** Some annotators precisely mark the boundaries of the tumor while others tend to follow it more loosely. This can be seen in Figure 1a between R1 and R6.

(b) **Conservative vs. complete inclusion:** Segmenting non-mass cases is a bit tricky because the tumor is highly unlocalized. In these cases, some annotator only mark the enhanced areas while others include the whole region affected by cancer in the segmentation. An example can be seen in Figure 1b between R5 and R6.

(c) **Excluded areas:** Tumors usually present in various shapes, sizes, and characteristics. In addition to being localized or scattered, they can also have necrosis and cysts which present as darker areas within the tumor. The annotators deal with these cases differently. An example can be seen in Figure 1a. where R4 excludes the necrosis area in the middle while R6 includes it in the region of interest.

3.1.2 Variability in network outputs:

We study the variability generated with MCDropout in comparison with the manual annotation variability in the previous section.

(a) **Fixed Dropout:** We generate six different predictions with a fixed dropout probability of 0.2 (same as training) during test time. Unlike manual annotations, the network depends on only one ground truth label for learning the segmentations. Therefore, the sources of uncertainty are limited to various representations of tumors and additional structures in the image as opposed to the techniques and underlying hypotheses of the multiple manual annotators. The results can be seen in Figure 3 (column 3). For the mass case, all predictions are quite certain and show very low disagreement. This is in contrast to the radiologist's segmentations (column 2) which show higher uncertainty in the middle due to necrosis and on top due to ambiguous breast tissue. For the non-mass case, the largest area annotated as the tumor shows agreement for both radiologist and network predictions. However, similar to radiologists, the network shows higher uncertainty for smaller undefined regions containing cancer on the right.

(b) **Variable Dropout:** In addition to using the fixed dropout probability of 0.2, we generate predictions by five different dropout probabilities (0.1, 0.2, 0.3, 0.4, and 0.5) during inference to see evolution in the prediction capability of the network as we decrease its capacity. The results can be seen in Figure 3 (columns 4 to 7). In the mass case, with dropouts of 0.1 and 0.3, it segments the whole area of the tumor with only slight differences at the pixel boundaries. However, with a high dropout of 0.5, it segments the tumor without the darker pixels in the middle, which indicates the network is more certain about the enhanced areas of the tumor when used at about half of its capacity. Similarly, for the non-mass case, we see that as we decrease the capacity by increasing dropout, the network predicts the largest tumor with clearly defined boundaries and does not overestimate the region of interest. Using variable dropouts is arguably a better way to model the manual inter-annotator variability as it represents different levels of knowledge and understanding of the task. This can be seen in the last column of the figure where the agreement maps show similar behavior as column 2.

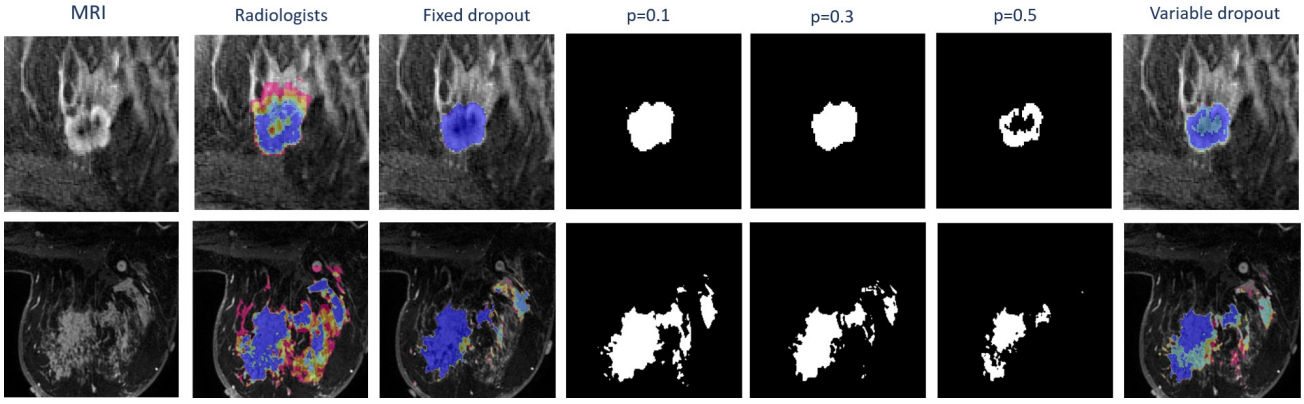

Figure 3. Difference in inter-segmentation variability in mass (top) and non-mass (bottom) tumors. From left to right, cropped MRI 2D slice with tumor; agreement map between radiologists; agreement map with six models with fixed dropout probability of 0.2; prediction with dropout of 0.1; prediction with dropout of 0.3; prediction with dropout of 0.5; agreement map with five models with variable dropouts. In the agreement maps, blue and red regions indicate agreement and disagreement between segmentations respectively.

3.2 Improving uncertainty and comparison for different cohorts

In this section, we look at the results of using uncertainty to improve our predictions with methods discussed in Section 2. Additionally, we also see how the results differ in easy and challenging cohorts, as annotated by two radiologists on a scale of 1 (very easy) to 5 (very hard). We average the scores by the two radiologists and

Table 1. Metrics obtained by combination methods on all cases, easy cases only and hard cases only. A fixed value of $p = 0.2$ is used. '*' denotes statistically significant results ($p < 0.05$) compared to the baseline model (predictions without any dropout). For Dice, higher values are better while for MHD, lower values are better.

Cohorts	**Number of predictions**	**Baseline model**		**Hard voting**		**Soft voting**		**Uncertainty**			
								Std. dev: 0.05		**Std. dev: 0.02**	
		Dice	MHD	Dice	MHD	Dice	MHD	Dice	MHD	Dice	MHD
All cases	5	0.7867	11.1653	0.795*	7.4398*	0.7924	8.5293*	0.7973	5.0867*	0.7987	**4.3075***
	10			**0.796***	7.2835*	0.7935*	8.5961*	0.7981	4.5294*	0.799	4.3187*
	15			0.7952*	7.5031*	0.7935*	8.6*	0.7974	4.5896*	0.7979	4.3472
Easy cases	5	0.8223	12.0801	0.8368*	7.4423*	0.8329*	8.9212*	0.8415*	4.953*	**0.8449***	**4.0955***
	10			0.8364*	7.3994*	0.8325*	9.1633*	0.845	4.1131*	0.845	4.1131*
	15			0.8353*	7.62*	0.8326*	9.1765*	0.8423*	4.2325*	0.8446	4.1381*
Hard cases	5	**0.7155**	9.3357	0.7113	7.4347	0.7116	7.7456	0.7087	5.3542	0.7063	4.7315
	10			0.7152	7.0516	0.7156	7.4617	0.7091	5.2162	0.7069	4.7299
	15			0.7151	7.2693	0.7152	7.4469	0.7078	5.3038	0.7046	4.7652

consider 3 and above to be hard tumors to delineate. With a total of 30 test cases, there are 20 cases in the easy and 10 cases in the hard cohorts.

(a) **Combining predictions:** Table 1 shows the segmentation metrics for baseline as well as combination methods as compared with the ground truth. Baseline prediction is generated without any test-time dropout. With MCDropout, we generate 5, 10, and 15 predictions and combine them using hard voting, soft voting, and uncertainty-aware voting. For uncertainty-aware voting, we choose two different standard deviations, 0.05 and 0.02, where the former allows pixels with higher uncertainties. We see that for all test cases, we see an improvement over dice ($0.7867 \rightarrow 0.796$) and MHD ($11.1653 \rightarrow 4.3075$). The best results are obtained with 5 and 10 predictions and increasing the number to 15 does not improve the metrics any further. Figures 4 and 5 show these methods in action. Figure 4a shows a case where there is an enhancement around the tumor. The baseline method includes it in the foreground while the combination methods remove it without affecting the actual tumor segmentation. Similarly, Figure 5a. shows a case with a lot of enhancement in the non-tumor area which is considered as foreground in the baseline while combination methods reduce these false positive segmentations to a great extent.

(b) **Easy vs. difficult cohorts:** Table 1 shows the results separately for easy and difficult cohorts. We see that for the easy cohort, we get higher improvement over dice ($0.8223 \rightarrow 0.8449$) and MHD ($12.0801 \rightarrow 4.0955$). The best results are obtained with 5 predictions and increasing the number to 10 or 15 does not improve the metrics any further. Figure 4 shows the performance on easy cases. Figure 4b. shows a case where the baseline model falsely predicts an area of enhancement as foreground while the combination methods remove it from the segmentation while keeping the tumor area intact. On the other hand, Figure 4c. represents an case with well-defined boundaries and the baseline prediction is already very certain. Therefore, minimal improvement is seen with the combination methods. For the hard cohort, we do not see improvements in segmentation metrics, rather there is a decrease in dice. We investigated further into the cause and we see some explanations in Figure 5 where 5b. shows a very undefined tumor with low enhancement. Therefore, the predictions are not certain at the boundaries and the combination methods actually push it further away from the GT. Similarly in Fig.5c, the predictions are only certain about the biggest lesion but removes the smaller cancerous areas.

4. DISCUSSION

Firstly, we compare the variability in manual segmentation and predictions generated by MCDropout. We observe that there is higher variability in the former due to differences in techniques and underlying assumptions. When

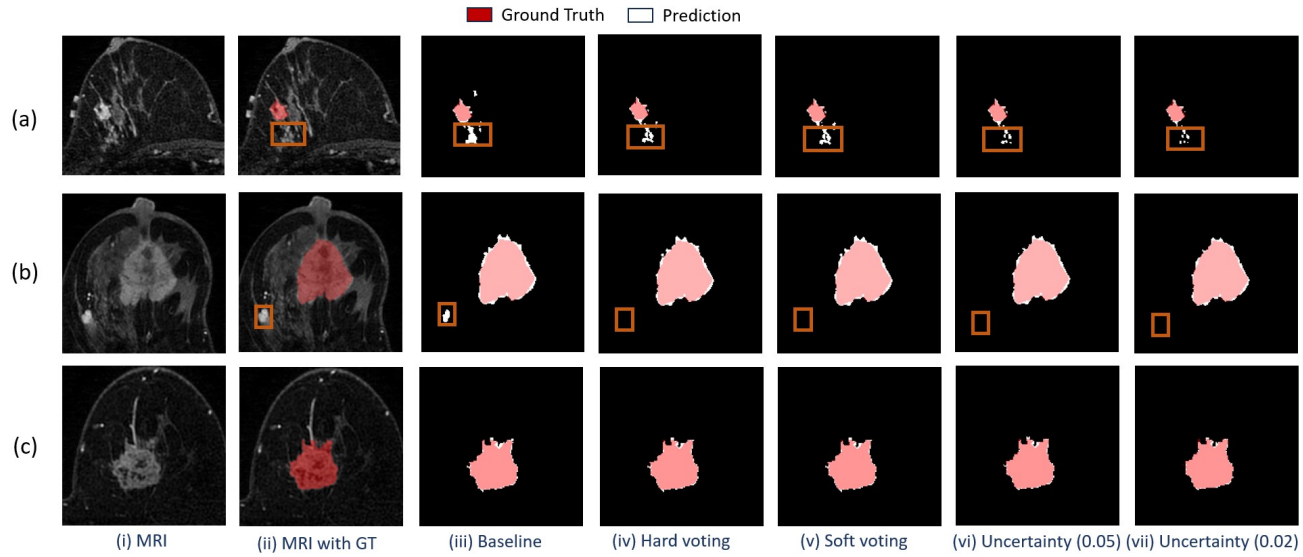

Figure 4. Segmentations obtained by baseline and combination methods are shown for the **easy cases**. The tumor is usually seen as an area with higher intensity values due to the injection of a contrast agent. (a) Some enhanced areas are seen close to the tumor (2.5; easy to medium). All methods identify the tumor as foreground. The baseline method overestimates the region of interest while the combination methods (iv) - (vii) reduce the false positives and predict areas with high certainty only. (b) A defined enhanced area is seen close to the tumour which is misidentified in the baseline. All combination methods set it as background (2; easy) (c) This is a localized tumor with well defined boundary (1.5; very easy to easy). The combination methods do not show any improvement over the baseline as the predicted areas are highly certain.

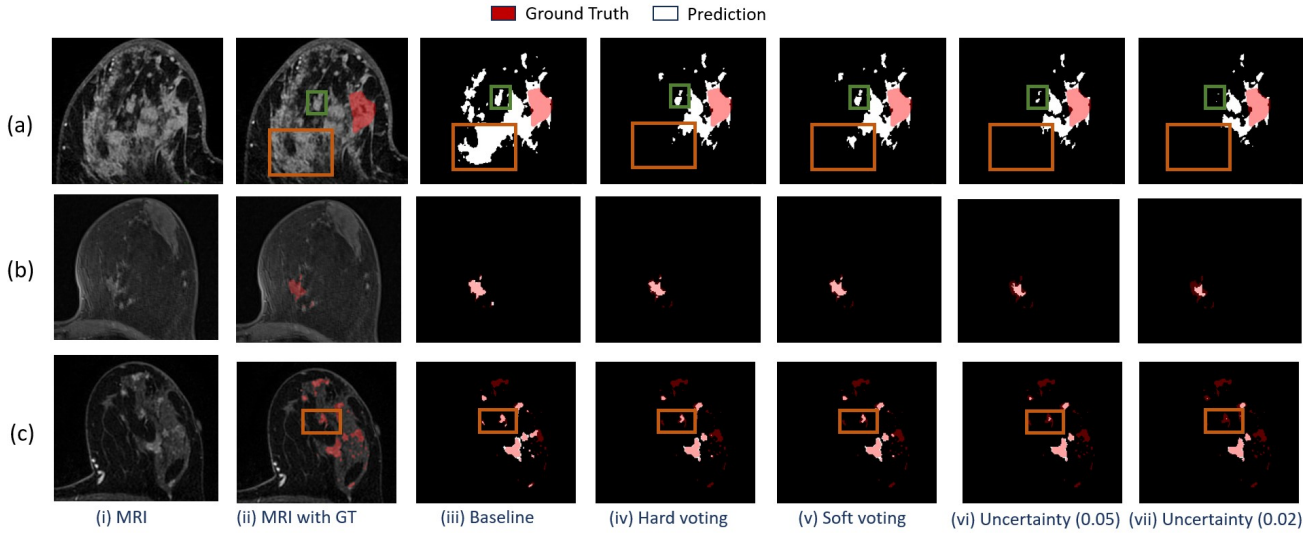

Figure 5. Segmentations obtained by baseline and combination methods are shown for the **hard cases**. The tumor is usually seen as an area with higher intensity values due to the injection of a contrast agent. (a) A lot of enhancement around the tumor makes it challenging to find its boundaries (5; very hard). All methods identify the tumor as foreground. The baseline method overestimates the region of interest while the combination methods (iv) - (vii) reduce the false positives and predict areas with high certainty only. (b) This tumor has low enhancement and fuzzy appearance increasing the model's uncertainty close to the boundaries. The baseline is closest to the GT while combination methods (iv) - (vii) underestimate the region of interest (4; hard). (c) This tumor is scattered across the breast and presents as a non-mass tumor (4; hard). The baseline is closest to ground truth while combination methods (iv) - (vii) eliminate the small regions of interest because of low certainty.

using the same dropout probability as used in the training, the uncertainties are mainly caused by structures that resemble the enhancement patterns of a tumor. For example, we see that the radiologists' segmentations are uncertain in necrosis region (Figure 3(top)) but as the model sees GT segmentation from only one radiologist while training, which includes necrosis area as part of foreground, it tends to include it in its predictions as well (Figures 4b. and 4c.). While using larger dropouts compared to the model training, the model relies more on the enhancement and shape information and gets uncertain about more complicated representations of the tumor. Secondly, we evaluate different methods of combining MCDropout predictions for improving segmentation. We obtain an overall improvement over the testset, with higher improvement on the cohort of easy cases.

One major limitation of our work is that the improvement of segmentation is only attributed to the reduction in false positives as the baseline usually overestimates the region of interest. We cannot handle cases of false negatives as the methods rely on high probability values of the foreground pixels. Furthermore, we only compare MCDropout for estimating the uncertainty. It would be interesting to evaluate other methods of uncertainty also demonstrate similar advantages and shortcomings.

5. ACKNOWLEDGMENTS

This project has received funding from Europe research and innovation programme under grant agreement No 101057699 (RadioVal) & by Horizon 2020 grant agreement No 952103 (EuCanImage) as well as the project FUTURE-ES (PID2021-126724OB-I00) from the Ministry of Science and Innovation of Spain.

We would like to thank all the radiologists for participating in the inter-annotator variability study: Gordana Ivanac (University of Zagreb Medical School, Croatia), Rosa García Dosda (La Fe University Hospital Valencia, Spain), Maria Laura Cosaka (Alexander Fleming Institute, Argentina), Meltem Gulsun Akpinar (Hacettepe University Hospital, Turkey) and Abeer Hamed (Ain Shams University Hospital, Egypt). We would like to thank Marco Caballo (Radboud University Medical Center, the Netherlands) for providing 254 segmentation masks for the Duke-Breast-Cancer MRI Dataset.

REFERENCES

[1] Global Cancer Observatory, "The global cancer observatory (gco) is an interactive web-based platform presenting global cancer statistics to inform cancer control and research." https://gco.iarc.fr/ (2023). Accessed: 2023-08-07.

[2] Caballo, M., Sanderink, W. B. G., Han, L., Gao, Y., Athanasiou, A., and Mann, R. M., "Four-dimensional machine learning radiomics for the pretreatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced mri," *Journal of magnetic resonance imaging : JMRI, 57(1), 97–110* (2023).

[3] Li, Q., Xiao, Q., Li, J., Wang, Z., Wang, H., and Gu, Y., "Value of machine learning with multiphases ce-mri radiomics for early prediction of pathological complete response to neoadjuvant therapy in her2-positive invasive breast cancer," *Cancer Management and Research* , 5053–5062 (2021).

[4] Zhou, J., Lu, J., Gao, C., Zeng, J., Zhou, C., Lai, X., Cai, W., and Xu, M., "Predicting the response to neoadjuvant chemotherapy for breast cancer: wavelet transforming radiomics in mri," *BMC cancer* **20**, 1–10 (2020).

[5] Liu, Z., Li, Z., Qu, J., Zhang, R., Zhou, X., Li, L., Sun, K., Tang, Z., Jiang, H., Li, H., et al., "Radiomics of multiparametric mri for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study," *Clinical Cancer Research* **25**(12), 3538–3547 (2019).

[6] Comes, M. C., La Forgia, D., Didonna, V., Fanizzi, A., Giotta, F., Latorre, A., Martinelli, E., Mencattini, A., Paradiso, A. V., Tamborra, P., et al., "Early prediction of breast cancer recurrence for patients treated with neoadjuvant chemotherapy: a transfer learning approach on dce-mris," *Cancers* **13**(10), 2298 (2021).

[7] Granzier, R. W., Verbakel, N. M., Ibrahim, A., Van Timmeren, J., Van Nijnatten, T., Leijenaar, R., Lobbes, M., Smidt, M., and Woodruff, H., "Mri-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability," *scientific reports* **10**(1), 14163 (2020).

[8] Visser, M., Müller, D., van Duijn, R., Smits, M., Verburg, N., Hendriks, E., Nabuurs, R., Bot, J., Eijgelaar, R., Witte, M., et al., "Inter-rater agreement in glioma segmentations on longitudinal mri," *NeuroImage: Clinical* **22**, 101727 (2019).

[9] Poirot, M. G., Caan, M., Ruhe, H. G., Bjørnerud, A., Groote, I., Reneman, L., and Marquering, H., "Robustness of radiomics to variations in segmentation methods in multimodal brain mri," *Scientific Reports* **12**(1), 16712 (2022).

[10] Lakshminarayanan, B., Pritzel, A., and Blundell, C., "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems* **30** (2017).

[11] Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D., "Learning in an uncertain world: Representing ambiguity through multiple hypotheses," in [*Proceedings of the IEEE international conference on computer vision*], 3591–3600 (2017).

[12] Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., and Ronneberger, O., "A probabilistic u-net for segmentation of ambiguous images," *Advances in neural information processing systems* **31** (2018).

[13] Gal, Y., Ghahramani, Z., Balcan, M., and Weinberger, K., "33rd international conference on machine learning, icml 2016," (2016).

[14] Saha, A., Harowicz, M. R., Grimm, L. J., Weng, J., Cain, E. H., Kim, C. E., Ghate, S. V., Walsh, R., and Mazurowski, M. A., "Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]," *The Cancer Imaging Archive* (2021).

[15] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Büttner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, A. T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C. I., Shetty, S., van Smeden, M., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Calster, B. V., Varoquaux, G., and Jäger, P. F., "Metrics reloaded: Recommendations for image analysis validation," (2023).

[16] Woolson, R. F., "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials* , 1–3 (2007).

[17] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*], 234–241, Springer (2015).

[18] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods* **18**(2), 203–211 (2021).

[19] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C., "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging* **29**(6), 1310–1320 (2010).

[20] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research* **15**(1), 1929–1958 (2014).