# Segmentation of gynaecological ultrasound images using different U-Net based approaches

Sónia Marques INESC TEC Porto, Portugal Catarina Carvalho INESC TEC Porto, Portugal Carla Peixoto Centro Hospitalar de São João Faculdade de Medicina da U.Porto Porto, Portugal

Jorge Beires Centro Hospitalar de São João Porto, Portugal Jorge Silva INESC TEC Faculdade de Engenharia da U.Porto Porto, Portugal Duarte Pignatelli Centro Hospitalar de São João Faculdade de Medicina da U.Porto Porto, Portugal

Aurélio Campilho INESC TEC Faculdade de Engenharia da U.Porto Porto, Portugal

*Abstract*—Ovarian cancer is one of the most commonly occurring cancer in women. Transvaginal ultrasound is used as a screening test to detect the presence of tumors but, for specific types of ovarian tumors, malignancy can only be asserted through surgery. An automatic method to perform the detection and malignancy assessment of these tumours is thus necessary to prevent unnecessary oophorectomies.

This work explores the U-Net's architecture and investigates the selection of different hyperparameters for the ovary and the ovarian follicles segmentation. The effect of applying different post-processing methods on beam-formed radio-frequency (BRF) data is also investigated.

Results show that models trained only with BRF data have the worst performance. On the other hand, the combination of Bmode with BRF data performs better for ovary segmentation. As for the hyperparameter study, results show that the U-Net with 4 levels is the architecture with the worst performance. This shows that to achieve better performance in the segmentation of ovarian structures, it is important to select an architecture that takes into account the spatial context of the regions of interest. It is also possible to conclude that the method used to analyse BRF data should be designed to take advantage of the fine-resolution of BRF data.

*Index Terms*—B-mode ultrasound data, beam-formed ultrasound data, image segmentation, convolutional neuronal networks, ovarian cancer

# I. INTRODUCTION

Ovarian cancer is one of the most commonly occurring cancer in women. The majority of ovarian cancers grows as a cystic mass, resulting from an abnormal development of ovarian follicles [1]. Both ovarian cysts and follicles are fluidfilled structures but cysts do not contain a oocyte, and contain abnormal tissue [2].

Typically, transvaginal ultrasound B-mode images are used for diagnosis of ovarian cancer. However, for specific types of ovarian tumors, malignancy can only be asserted through surgery. The majority of the cystic masses detected by B-mode ultrasound (US) screening are benign, but such uncertainty of diagnosis leads to unnecessary oophorectomies [3]. An

FCT - Fundação para a Ciência e a Tecnologia.

automatic method to perform the detection and malignancy assessment of these masses is thus necessary.

Segmentation of ovarian follicles on US images entails several challenges, namely the low signal-to-noise ratio of US images, the high variability of the shape and size of the follicles and the presence of hypoechoic structures besides follicles, such as vessels and the bladder.

The state-of-the-art methods for the automatic segmentation of ovarian follicles are generally focused on gradient and texture analysis integrated in multi-stage algorithms [4], [5], energy minimization techniques such as active contour [6], [7] and graph cut optimization [8]. The methods proposed in [6] and [7] also take advantage of deep learning techniques through the use of a multilayered perceptron (MLP) but in both cases the MLP is only used to classify the previously selected regions of interest as follicle or non-follicle. Among these, the best performing method [7] reported an accuracy of 98.3%.

Additionally, in [9] an end-to-end deep-learning approach is proposed for the segmentation of ovarian structures, which reports a mean Dice score of 0.784 for the follicles.

The type of images used in the state-of-the-art methods are strictly B-mode US images, in some cases employing image enhancement methods for denoising and contrast enhancement. However, beam-formed radio-frequency (BRF) is a raw type of US data, which includes both structural (lowerfrequency) and textural (higher-frequency) information.

This study assesses the importance of the application of post-processing techniques to US data and of the selection of U-Net's hyperparameters on the performance of a method fully based on the U-Net for the segmentation of the ovary and the ovarian follicles.

#### II. METHODS

This section presents the methodology used in this study, including a description of the dataset, the convolutional neural network implemented, and also the selected loss function and hyperparameters.

# A. Dataset

The dataset used is based on 107 beam-formed radiofrequency (BRF) transvaginal US images of the ovary. These images were acquired at Centro Hospitalar de São João, with consent of the patients, while these attended first time appointment for fertility treatment planning. Each image contains one ovary with single or multiple follicles. The original BRF data was acquired with an EC9-5/10 Endovaginal Microconvex transducer (frequency range 9-5 MHz and image field of 124°). The acquired BRF images are 21.7 mm wide, and between 58.8 mm and 107.1 mm of depth, with the original image dimensions of  $192 \times [2000, 6200]$  px.



Fig. 1: Example of dataset images. On the left, post-processed linear B-mode image. On the right, the corresponding linear B-mode image with the ovary ground-truth superimposed in green and the follicle ground-truth in red.

Three different types of images were extracted from the original BRF data, namely, absolute BRF, B-mode and post-processed B-mode.

Absolute BRF (or BRF) is obtained by computing the absolute value of the original BRF data and normalizing the obtained intensity to the range [0, 255].

B-mode linear data was obtained by applying quadrature signal demodulation and filtering with a Hamming window of the original BRF data, followed by envelope detection and log compression. To produce the post-processed B-mode images, Contrast Limited Adaptive Histogram Equalization and a despeckling filter [10] were applied to the linear B-mode images (see Fig. 1.a).

The three versions of the data were fed as input to the network either individually, the three combined into a 3D array, or in combinations of two different versions. All images were resized to  $512 \times 512$ , to normalize the resolution of the input data.

The ground-truth (GT) of follicles and ovary were delineated by a medical expert, as shown in Fig. 1.b. The dataset division was as follows: 92 images were used for 5-fold crossvalidation and 15 for test.

#### B. Convolutional Neural Network

The Convolutional Neural Network (CNN) used is the U-Net, typically used for segmentation tasks in the biomedical field. The implemented U-Net architecture is similar to the original [11], only adding optimal zero-padding to preserve the image spatial size. The U-Net (used as baseline architecture) begins with a contracting path composed by four down-sampling steps, whose output is then passed to an expanding path composed by four up-sampling steps, resulting in 5 levels of resolution. At each down-sampling/up-sampling step, two sequences of convolution ( $3 \times 3$  kernels), batch normalization and rectified linear unit (ReLU) are computed. In each down-sampling and up-sampling step occurs, respectively, the two-fold increase and reduction of the number of features. Each step of the contracting path is followed by max-pooling with a  $2 \times 2$  kernel and stride of 2, and each step of the expanding path is followed by 2D transposed convolution with a scale factor of 2.

At the last step of the U-Net, two sequences of convolution, batch normalization and rectified linear unit (ReLU) are performed, followed by a convolution with a  $1 \times 1$  kernel and stride of 1, and the softmax function. The final output consists of a  $512 \times 512$  image with two classes (follicle/non-follicle or ovary/non-ovary).

For this study, three variations of the baseline were tested, namely:

- Kernel size of convolutions of  $5 \times 5$
- Depth of 4 levels of resolution
- Depth of 6 levels of resolution

# C. Learning framework

All models were trained using cross-entropy (CE) as loss function:

$$CE = \sum_{i}^{N} \sum_{j=1}^{C} t_{ij} \log(s_{ij})$$
(1)

where N corresponds to the number of pixels in the image, C is the number of output classes,  $t_{ij}$  is the pixel's ground-truth value for that class and  $s_{ij}$  is the pixel's predicted probability score for that class.

The batch size was set to 4 images, except for models with the following characteristics, in which it was set to 2, due to equipment limitations:

- Baseline architecture (with input of the combination of B-mode and BRF, the combination of B-mode and postprocessed B-mode and the combination of post-processed B-mode and BRF)
- Architecture with 6 levels of resolution (for all types of input data)

The initial learning rate was set to 0.001, and decreased by a factor of 4 every time the validation loss improved less than  $10^{-3}$ . For the optimization of the U-Net's parameters, the Adaptive Moment Estimation (Adam) optimizer [12] is used. The early stopping method implemented consists in stopping training after 20 epochs without improvement of the validation loss, with a maximum number of 100 epochs.

The implementation of this study was done in Python 3.6.6 using Pytorch 0.4.1.

### III. RESULTS

For the evaluation of the models' performance, test results were binarized with a threshold of 0.5 and compared against Program Digest 2019 IEEE IUS Glasgow, Scotland, October 6-9, 2019

the GT using the Dice Similarity Coefficient (DSC). For models' performance analysis the mean DSC was computed across the 5 folds of cross-validation.

A paired two tailed t-test analysis was used to compare the performance of the different models, being the null hypothesis that their performance was equal. The alpha level for all tests was set at 0.05.



Fig. 2: Mean DSC for the test-set of follicle (a) and ovary (b) segmentation results for each trained model.

The bar plots with the mean DSC of the images in the test-set for each of the trained models for follicle and ovary segmentation are found in Fig. 2. In this document, "Bm", "Proc" and "All" refer to the B-mode, post-processed B-mode and the combination of the three types of input data, respectively. Furthermore, " $5 \times 5$  kernel", "4 levels" and "6 levels" refer to the three variations of the baseline U-Net's architecture.

Table I shows the statistical difference found when comparing the different architecture hyperparameters, according to the corresponding DSC results. Results show that when comparing the U-Net with 4 levels of depth against the others, the former has a significantly worse performance, being especially significant in the case of ovary segmentation. Among the other tested architectures, no significant difference was found. TABLE I: p-value found for comparison between the tested U-Net's hyperparameters, for each type of input (alpha level was set at 0.05). Bold identifies the rejected null hypothesis cases.

		l	o-value for	r —	p-value for			
Input type	U-Net's hyperp.	follic	e segment	tation	ovary segmentation			
		6	5x5	4	6	5x5	4	
		levels	kernel	levels	levels	kernel	levels	
All	baseline	0.919	0.448	0.976	0.722	0.217	0.01	
	6 levels	-	0.537	0.925	-	0.5	0.047	
	5x5 kernel		-	0.525		-	0.008	
Bm &BRF	baseline	0.422	0.099	0.008	0.306	0.665	0.023	
	6 levels	-	0.74	0.085	-	0.523	0.043	
	5x5 kernel		-	0.098		-	0.003	
Bm &Proc	baseline	0.503	0.41	0.744	0.341	0.151	0.045	
	6 levels	-	0.8	0.281	-	0.317	0.019	
	5x5 kernel		-	0.332		-	0.003	
BRF &Proc	baseline	0.582	0.037	0.572	0.082	0.119	0.139	
	6 levels	-	0.081	0.216	-	0.741	0.058	
	5x5		_	0.017		_	0.028	
	kernel		-	0.017		-	0.020	
Bm	baseline	0.9	0.922	0.033	0.943	0.78	0.006	
	6 levels	-	0.925	0.264	-	0.864	0.059	
	5x5 kernel		-	0.209		-	0.029	
BRF	baseline	0.649	0.17	0.003	0.146	0.079	0.029	
	6 levels	-	0.077	0.069	-	0.837	0.0	
	5x5 kernel		-	0.008		-	0.002	
Proc	baseline	0.391	0.551	0.072	0.668	0.243	0.0	
	6 levels	-	0.932	0.18	-	0.083	0.005	
	5x5 kernel		-	0.363		-	0.001	

TABLE II: p-value found for comparison between the types of input, for baseline U-Net's hyperparameters (alpha level was set at 0.05). Bold identifies the rejected null hypothesis cases.

	Input type	p-value							
		Bm &BRF	Bm &Proc	BRF &Proc	Bm	BRF	Proc		
Ovary	All	0.106	0.781	0.729	0.754	0.034	0.457		
	Bm&BRF	-	0.042	0.052	0.079	0.002	0.154		
	Bm&Proc		-	0.38	0.962	0.076	0.733		
	BRF&Proc			-	0.643	0.109	0.322		
	Bm				-	0.031	0.849		
	BRF					-	0.038		
Follicle	All	0.361	0.17	0.088	0.552	0.01	0.845		
	Bm&BRF	nan	0.112	0.155	0.601	0.001	0.454		
	Bm&Proc		nan	0.9	0.202	0.038	0.627		
	BRF&Proc			nan	0.152	0.059	0.347		
	Bm				nan	0.006	0.527		
	BRF					nan	0.017		

Table II shows the statistical difference found when comparing the different types of input data, according to the corresponding DSC results. Results show that models trained only with BRF data have the worst performance, both for the segmentation of follicles and of the ovary. Besides that, no other input type proved to be significantly better for follicle segmentation. On the other hand, the combination of B-mode with BRF data performs better for ovary segmentation.

Examples of the best and worst results for follicle and ovary segmentation are presented in Fig. 3. In these, for comparison between GT and segmentation results, true positives, false positives and false negatives are represented in yellow, red and green, respectively.

### IV. DISCUSSION

This paper presents a detailed performance comparison between models with different U-Net's architecture hyperpa-



Fig. 3: Best and worst DSC scores for (a-b) ovary and (c-d) follicle segmentation.

rameters and type of input data.

The results of the study of the type of input data demonstrate that exclusive use of unfiltered data, such as BRF, is not suitable for follicle and ovary segmentation. However, for ovary segmentation, the improved performance of BRF when combined with B-mode reveals the importance of highfrequency data for the recognition of texture information of the ovarian stroma. The use of post-processed B-mode revealed to be ineffective. This could be due to either the low performance of the U-Net or an inadequate enhancement of the characteristics such as texture and edges.

Regarding the study of U-Net's architecture hyperparameters, the lower performance of the U-Net with 4 levels of resolution demonstrates that the lower complexity of the CNN results in an insufficient analysis of the data and of the spatial context of the regions of interest. The 5 levels used in the baseline architecture are sufficient to allow a proper comprehension of the data, given that the use of the U-Net with 6 levels did not result in further improvement.

# V. CONCLUSION

This work presents a comprehensive study of the effect of the selection of the U-Net's architecture hyperparameters (such as depth and kernel size) and of the use of processing techniques on US data (including the conversion from BRF to B-mode, filtering and contrast enhancement techniques) for the segmentation of the follicles and the ovary.

It was concluded that to achieve better performance in the segmentation of these structures, it is important to select an architecture that can take into account the spatial context of the regions of interest. It is also possible to conclude that the method used to analyse BRF data should be designed to take advantage of the fine-resolution of BRF data.

#### ACKNOWLEDGMENT

This work is financed by national funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia as part of project UID/EEA/50014/2019.

#### REFERENCES

- J. S. Richards, "From follicular development and ovulation to ovarian cancers: An unexpected journey," in Vitamins and Hormones, pp. 453-472, Elsevier, 2018.
- [2] M, Furuya, "Ovarian cancer stroma: pathophysiology and the roles in cancer development," in Cancers, vol. 4, pp. 701-24, 2012.
- [3] J. A. Rauh-Hain, T. C. Krivak, M. G. Del Carmen, and A. B. Olawaiye,"Ovarian cancer screening and early detection in the general population." Reviews in obstetrics and gynecol-ogy, vol. 4, no. 1, pp. 15-21, 2011.
- [4] P. Hiremath and J. Tegnoor, "Automatic detection of follicles in ultrasound images of ovaries using edge based method" International Journal of Computer Applications, 2010.
- [5] B. Cigale, D. Zazula, "Directional 3D wavelet transform based on gaussian mixtures for the analysis of 3D ultrasound ovarian volumes,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 64-77, Jan 2019.
- [6] A. D. Usman, O. R. Isah, and A. M. S. Tekanyi, "Application of Artificial Neural Network and Texture Features for Follicle Detection," African Journal of Computing & ICT, vol. 8,no. 4, pp. 2-9, 2015.
- [7] O. R. Isah, A. D. Usman, and A. M. Tekanyi, "A hybrid model of PSO algorithm and artificial neural network for automatic follicle classification," International Journal Bioautomation, vol. 21, no. 1, pp. 43-58, 2017.
- [8] T. Chen, W. Zhang, S. Good, S. K. Zhou, and D. Comaniciu, "Automatic ovarian follicle quantification from 3D ultrasound data using global/local context with database guided segmentation," 2009 IEEE 12th Int. Conf. on Computer Vision, pp. 795-802, 2009.
- [9] D. S. Wanderley, C. B. Carvalho, A. Domingues, C. Peixoto, D. Pignatelli, J. Beires, J. Silva, and A. Campilho, "End-to-end ovarian structures segmentation," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 681-689, Springer International Publishing, 2019.
- [10] M. Nii, Y. Kato, M. Morimoto, S. Kobashi, N. Kamiura, Y. Hata, S. Imawaki, T. Ishikawa, and H. Matsubayashi, "Ovarian follicle classification using numerical and B-mode image features from ultrasound scanning devices," in 2018 Joint 7th Int. Conf. on Informatics, Electronics Vision and 2018 2nd Int. Conf. on Imaging, Vision Pattern Recognition, pp. 222–227, 2018.
- [11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," Lecture Notes in Computer Science, vol. 9351, pp. 234-241, 2015.
- [12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Int. Conf. on Learning Representations, 2014.