Impact of network parameters on a U-Net based system for rectal cancer segmentation on MR images

Impact of network parameters on a U-Net based system for rectal cancer segmentation on MR images

Jovana Panic
Dept. of Electronics and Telecommunications
Polytechnic of Turin
Torino, Italy
jovana.panic@polito.it

Valentina Giannini
Dept. of Surgical Science
University of Turin and
Candiolo Cancer Institute, FPO-IRCCS
Torino, Italy
valentina.giannini@unito.it

Arianna Defeudis
Dept. of Surgical Science
University of Turin and
Candiolo Cancer Institute, FPO-IRCCS
Torino, Italy
arianna.defeudis@unito.it

Daniele Regge
Dept. of Surgical Science
University of Turin and
Candiolo Cancer Institute, FPO-IRCCS
Torino, Italy
daniele.regge@unito.it

Gabriella Balestra
Dept. of Electronics and Telecommunications
Polytechnic of Turin
Torino, Italy
gabriella.balestra@polito.it

Samanta Rosati
Dept. of Electronics and Telecommunications
Polytechnic of Turin
Torino, Italy
samanta.rosati@polito.it

***Abstract*— The use of Deep Learning (DL) algorithms in the medical imaging field is increasing in recent years. However, they require the selection of a set of parameters to properly perform. In this study we evaluated the impact of three factors (the construction of the training set, the number of network layers and the loss function) on the performance of a U-Net system in the segmentation of *Locally Advanced Rectal Cancer* (LARC) on Magnetic Resonance Imaging (MRI). Images from 3 different institutions and 4 different scanners were used to this scope, for a total of 100 patients. All images underwent a pre-processing step to normalize and to highlight the tumoral area. The sequences of two scanners were used to construct the networks while the remaining sequences were employed for validating the best performing systems. From our results, it emerged that Dice Similarity Coefficient is not affected by any of the evaluated factors. Conversely, the choice of loss function could bias the results towards either precision or recall and, thus, it should be properly performed according to the scope of the network. Moreover, a slightly improvement of the performances was observed using a training set based on clustering, maybe due to a better representation of the heterogeneity characterizing medical images.**

Keywords—Deep Learning, U-Net, tuning, segmentation, MRI, colorectal cancer

I. Introduction

Artificial Intelligence (AI) has shown promises in the development of Computer Aided Diagnosis (CAD) systems able to detect suspicious areas on medical images, characterize tumors and predict patient's response to therapy, thus allowing more personalized treatments [1]–[5]. Despite the promising preliminary results obtained by these signatures, the translation of this approaches into clinical practice is still limited by many reasons, including the lack of automatic methods to segment tumors.

In the last few years, Deep Learning (DL) algorithms have been used in the medical imaging field to segment and detect anatomical structures (cells, tumors, organs, etc,) [6]–[8]. All DL systems require the choice of a set of parameters, such as number of layers and loss function. Moreover, also the selection of the appropriate examples to be used for the net training and the structure of the network itself could influence the network performances. Although some studies faced the problem of tuning [9]–[11], to the best of our knowledge no studies assessed the impact of each specific parameter on the final performances of the system.

In the wide field of DL, recently has gained interest the U-Net architecture [12], that has been presented to overcome some limitations of previously developed structures, i.e, Fully Convolutional Networks (FCNs) and Convolutional Neural Networks (CNNs). The main advantage of the U-Net structure is the absence of the fully connected layer, replaced by the up-sampling layer and the deconvolutional layer, which allows to obtain a probability score map with the same size of the input, classifying each pixel instead the whole image [12].

The aim of this study is to evaluate how the construction of the training set, the number of network layers and the loss function influence the performance of a U-Net system in the segmentation of *Locally Advanced Rectal Cancer* (LARC) on Magnetic Resonance Imaging (MRI).

II. Materials and Methods

A. Population and Reference Standard

In this study, we retrospectively included 100 patients (61 men, 39 women; mean age: 64 years; range: 34-86 years) with histologically-confirmed LARC from three different Italian hospitals: Candiolo Cancer Institute, FPO-IRCCS of Candiolo (Center A); Molinette hospital A.O.U. Città della Salute e della Scienza of Turin (Center B); Mauriziano hospital of Turin (Center C). All patients underwent multiparametric (mp)MRI before neoadjuvant chemoradiotherapy (nCRT) after October 2010, including at least the axial fast spin-echo T2 weighted (T2w) and the axial EPI-SE Diffusion-Weighted Imaging (DWI) sequences. The sequences were acquired by four different scanners: two from center A (A.1 and A.2), one for center B and one for center C. All standard sequences were collected according to MRI guidelines for reporting rectal cancer staging [13].

All tumours were manually segmented by a resident radiologist and revised by a second radiologist with more than 10 years' experience in mpMRI. These masks have been used as the reference standard for the U-Net based system.

This was a multi-center retrospective study approved by the institutional review boards (IRBs) in each institution, with a waiver for requirement of informed consent as de-identified data were used.

B. *Pre-processing*

The pre-processing phase consists of three steps: the evaluation of the Apparent Diffusion Coefficient (ADC), the cropping and the normalization of the images.

First, the ADC is calculated from the DWI sequences of each patient according to the mono-exponential equation:

$$ADC = -\frac{1}{b}\ln\left(\frac{S_b}{S_0}\right), \quad (1)$$

where S_0 is the acquisition with b-value equal to 0, while S_b is the one corresponding to the highest b-value [14]. The second step consists in the automatic crop of the images around the bounding box containing the tumor, to reduce the amount of irrelevant information and to minimize the differences among patients during the normalization step, as previously described [15]. Finally, all cropped images were normalized to account for differences arose from different scanners. Both T2w and ADC sequences are standardized using the *z-score normalization*, as follows:

$$img_{std} = \frac{img-\mu}{\sigma}, \quad (2)$$

where μ is the mean pixel intensity, and σ the standard deviation.

C. *U-Net Tuning*

The U-Net structure is characterized by two sections: the contracting path and the expansive path [12]. The contracting path follows the typical CNN architecture with five convolutional blocks, each characterized by two subsequent convolutional layers followed by the max pooling layer. This path aims to extract the features maps related to the identification of the object of interest in different levels of resolution. The expansive path is symmetrical to the contracting one. So, each convolutional block is characterized by two convolutional layers followed by the up-sampling layer. This path aims to localize on the image the features extracted in the corresponding level of the contracting path. Thanks to these two paths, the neural network is able to learn how the object is characterized and where it could be localized in the image. Fig. 1 shows the developed U-Net structure, characterized by a 3x3 kernel and the Rectified Linear Unit (ReLU) activation function [16], except the output layer which is defined by 1x1 kernel and the sigmoid activation function [17].

Starting from this basic structure, we assessed the effect of three factors: training set, number of descending levels and loss function.

- *Training Set*

Patients from center A.1 and from center C (n=58) were used to construct the system (*construction set*) while cases from center A.2 and center B (n=42) were used as external validation set.

The construction set was divided into training set, to be used for training the system, and test set, adopted for the system tuning and parameter selection. To obtain a balanced training set, we collected the same number of slices with and without tumor. All tumoral slices were included, while the non-tumoral ones were chosen randomly among all slices of all patients. Two different methodologies were applied to obtain the training set:

a. Random sampling (*rnd*): this method is based on a random selection of 70% of the patients from the center A.1 and 70% from the center C, while all the remaining cases are included in the test set. The training set *rnd* was then composed of 41 patients, and in particular 222 tumoral and 222 non tumoral 256x256 slices.

b. Sampling based on clustering (*dend*): this approach is based on an agglomerative hierachical clustering method that organizes data in a hierarchical tree (called dendrogram) based on a proximity measure. Then, the final clusters are obtained by cutting the tree at a certain level [18]. To apply this approach, first we extracted from each patient of center A.1 and center C the following 20 features: mean, standard deviation, median, 25th and 75th percentile of both the LARC volume and the whole cropped volume in the T2w and ADC sequences. Then, we applied the hierarchical clustering to these patients and cut the tree in order to create two clusters. The training set was created by randomly collecting the same number of patients (70% of the less numerous cluster) from each cluster. The discarded patients of the clusters were included in the test set. The training set *dend* was

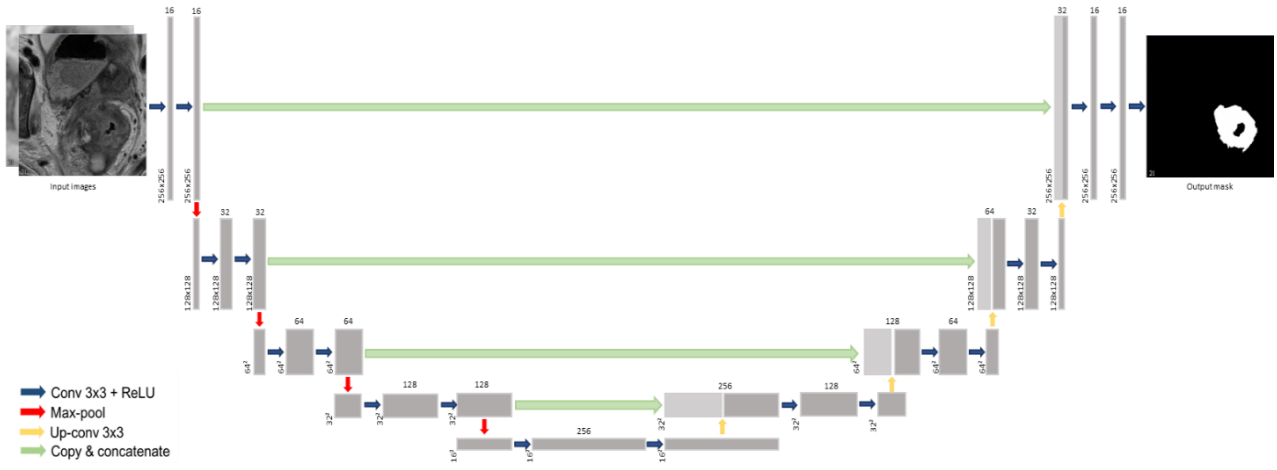

Fig. 1: U-Net structure with 4 descending levels.

composed of 36 patients, in particular 211 tumoral and 211 non tumoral 256x256 slices.

- *Number of descending levels*

The number of descending levels defines how deep the neural network can go. There are two main problems due to an high number of descending levels: the vanishing gradient problem [19] during the training and the poor generalization caused by the overfitting on the training data. The first issue is characterized by the gradient becoming so small preventing the weights of the neurons from changing their values. The second one defines the inability of the model to correctly analyze data not used during the training and generalize the results. In particular, we evaluated a number of descending levels equal to 3 (3lv), 4 (4lv) and 5 (5lv).

- *Loss function*

The loss function estimates the prediction errors of the model to be used to update neurons' weight during the training phase. Since the DL network learns a mapping from inputs to outputs, it is of key importance to choose the most suitable loss function for the task.

Two different loss functions were compared:

a. Binary Crossentropy (BC): this loss function (3), has been demonstrated useful for binary classification tasks.

$$BC = -\frac{1}{N}\sum\left(y_i \cdot \log\left(p(y_i)\right) + (1-y_i) \cdot \log(1-y_i)\right), \qquad (3)$$

where y_i is the i-th label and $p(y_i)$ is the predicted probability of the sample to belong to the i-th label class.

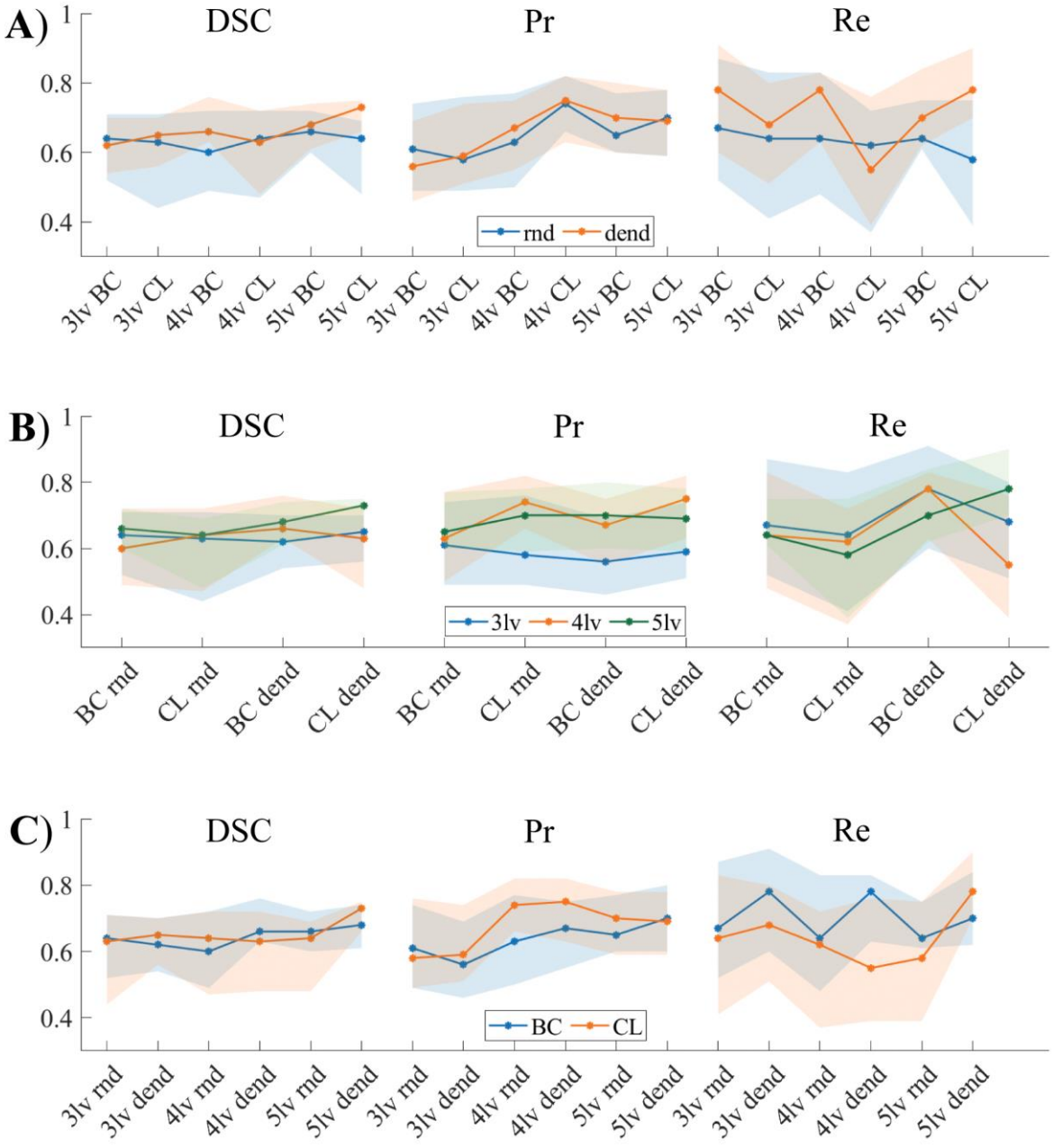

Fig. 2: Impact of the training set (panel A), number of descending levels (panel B) and loss function (panel C) on the segmentation performances obtained on the test set with different U-Net configurations. The solid lines represent the median values across all patients in the test set, whereas the colored areas represent the corresponding inter-quartile ranges.

b. Custom Loss (CL): this loss function is based on the Dice Loss (DL) function (4).

$$DL = 1 - \frac{2\sum y_i * p(y_i)}{\sum y_i + \sum p(y_i)}. \quad (4)$$

The DL function allows a better balance between classes and a network learning improvement in detecting even objects of small size and weak saliency [20].

More specifically, our implemented loss function was obtained by merging (3) and (4):

$$CL = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(p(y_i)\right) + 1 - \frac{2\sum y_i * p(y_i)}{\sum y_i + \sum p(y_i)}. \quad (5)$$

where, in (4) and (5), y_i is the *i*-th label and $p(y_i)$ is the predicted probability of the sample to belong to the *i*-th label class. With respect to the DL function, the CL function is able to overcome the issues related to the class imbalance, since there is a higher number of non-tumoral voxels against tumoral ones.

A total of 12 U-net systems were evaluated. All networks were implemented in Keras with Tensorflow library, with the Adam optimizer [21] and a learning rate of 0.001, β1 of 0.9 and β2 of 0.999.

D. Performance Evaluation

The probability maps obtained from the U-Nets were binarized using the Otsu's Thresholding method [22]. The following metrics were computed by comparing the binary masks with the reference standard:

- *Dice Similarity Coefficient* (DSC):

$$DSC = \frac{2|MM \cap OM|}{|MM|+|OM|} = \frac{2TP}{FP+2TP+FN}, \quad (6)$$

where MM is the volume of the manual mask (reference standard), OM the volume of the mask produced by the network, TP True Positive, FP False Positive, and FN False Negative. DSC is also known as *Overlap Index*. It gives a score between 0 and 1, where 1 indicates a complete overlap, i.e. the predicted mask OM is identical to the ground truth MM.

- *Precision* (Pr):

$$Pr = \frac{TP}{TP+FP}, \quad (7)$$

it gives a value between 0 and 1, where 1 indicates the best prediction. It gives a measure related to the over-segmentation.

- *Recall (Re):*

$$Re = \frac{TP}{TP+FN}. \quad (8)$$

this parameter is also referred as the *True Positive Rate* or *sensitivity*. As the others, it gives a value between 0 and 1, where 1 indicates the best prediction. Unlike the previous parameter, it gives a measure related to the under-segmentation.

In general terms, Pr means the percentage of results which are relevant, and Re refers to the percentage of total relevant results correctly classified by the algorithm. They are often used together in order to have a complete overview of the behavior of the system. In fact, all those parameters provide information related on how much the segmentations differ, both in shape and in misclassified tissues.

The two-tailed Mann-Whitney U test was performed to compare the results of the three metrics. A p-value <0.05 was considered statistically significant.

First, we evaluate the impact of each of the three above mentioned factors on the three metrics. Then, we chose the best U-Net parameters' combinations based on DSC, Pr and Re and we applied these networks to the external validation dataset.

III. Results and Discussion

Fig. 2 shows the impact of the training set (panel A), number of descending levels (panel B) and loss function (panel C) on the segmentation performances obtained on the test set.

In general, the DSC is not particularly influenced by any of the three evaluated factors.

For what concerns Pr, no differences can be observed between the two training sets (Fig. 2A), whereas lower Pr are evident with a number of descending level equal to 3 (Fig. 2B). In particular, 4lv results are statistically higher than 3lv in the configuration with CL-dend (p-value: 0.04) and it shows a trend toward significance in combination with CL-rnd and BC-dend (p=0.07 and 0.08, respectively). Moreover,

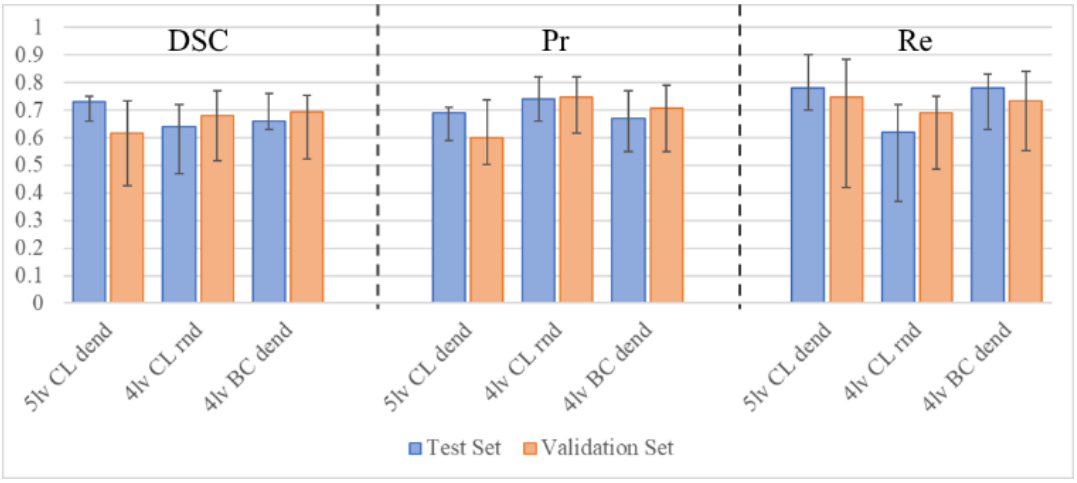

Fig. 3: Results of the three best U-Net structures applied on the test set and on the external validation set. Bars represent the median values and whiskers represent the corresponding inter-quartile ranges.

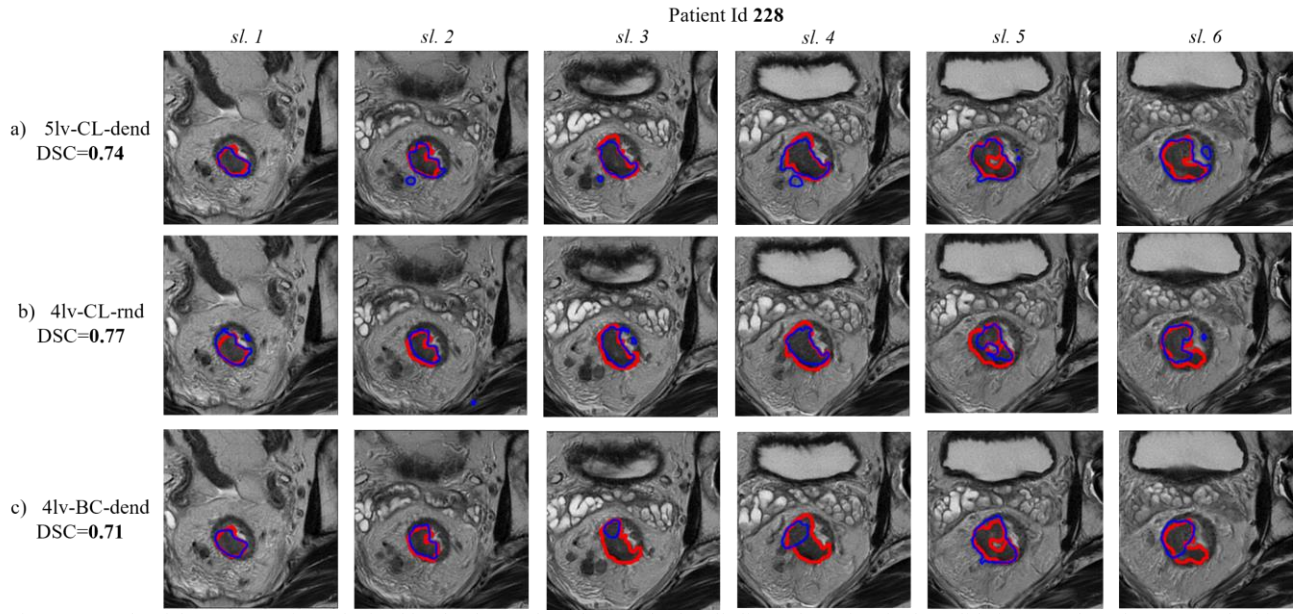

Fig. 4. Example of the segmentations obtained by the three selected networks for all slices of a patient in the validation set. The reference standard is presented by the red line, the network's segmentation by the blue line.

5lv is statistically higher than 3lv when it is used with BC-dend (p=0.03). No differences are found between 4lv and 5lv. Finally, comparing the two loss functions (Fig. 2C), a slightly higher Pr is reached for CL with respect to BC, in combination with 4lv-dend (p-value=0.09).

Focusing on Re, we found that in general the training set dend outperforms the training set rnd (Fig. 2A). In particular, Re is statistically higher when dend is combined with 5lv-CL (p=0.009) and it shows a trend toward significance in combination with 4lv-BC (p=0.07). Comparing different descending levels (Fig. 2B), 5lv shows a higher Re than 4lv when combined with CL-dend (p=0.004). Regarding differences between BC and CL (Fig. 2.C), we can observe a general increase of Re for BC, in particular when it is used with 4lv-dend.

Observing the general behaviour for BC and CL (last row in Fig.2), we can notice that the two loss functions affect Pr and Re in an opposite manner: higher Pr is visible for CL whereas higher Re is obtained with BC (Fig. 2C). This is coherent with literature since CL has been developed to overcome the limitation of BC when dealing with unbalanced datasets.

From the analysis of the overall results, we selected the combinations 5lv-CL-dend and 4lv-BC-dend, since they show the highest Re values (the combination 3lv-BC-dend was excluded due to its low Pr value). Moreover, the combination 4lv-CL-rnd was selected for its high Pr value, while 4lv-CL-dend was discarded since it obtained a low Re value. The DSC was not useful in this phase since there were no significant differences among the networks.

The three selected U-Net structures were applied on the external validation set and the obtained results are showed in Fig. 3, compared with those on the test set. As it emerged from the graph, the three structures proved to be robust as they perform similarly on the test and validation sets.

Fig. 4 shows an example of all tumoral slices of a patient belonging to the validation set where 5lv-CL-dend achieves DSC=0.74, 4lv-CL-rnd DSC=0.77 and 4lv-BC-dend DSC=0.71. Even if the DSC values are around the 0.7~0.8, all models are able to correctly detect most of the tumoral volume, despite the presence of some FN and FP voxels in some slices (in particular sl. 3 and 4 and 6 for 5lv-CL-dend and 4lv-BC-dend).

IV. Conclusions

In this study we assessed the impact of three different factors (i.e. training set, number of descending levels and loss function) on the segmentation performance obtained with a U-Net.

Looking at the results, it is evident that DSC is not affected by any of the evaluated factors. Conversely, Pr and Re show different behaviors on almost all evaluated networks. In particular, the choice of loss function could bias the results towards either Pr or Re and, thus, it should be properly performed according to the scope of the system.

Moreover, we observed a slightly improvement of the performances using a training set based on clustering. This could be due to a better representation of the heterogeneity characterizing medical images using this kind of approach. In future, other clustering methods could be employed for the training set construction.

For what concerns the number of descending levels and other combinations, we cannot draw any precise guidance on which configuration is the best for each problem, but we have demonstrated the need for precise and comprehensive parameter tuning to develop DL networks. In fact, we proved that tuning allows to select structures having good generalization capability, since their results were robust also on the external validation set.

Acknowledgments

This work was funded by AIRC 5xmille Special Program Molecular Clinical Oncology - Ref. 9970, FPRC 5xmille 2013 Ministero della Salute, and FPRC 5xmille 2015 Ministero della Salute (STRATEGY), Fondazione AIRC under 5 per Mille 2018—ID. 21091 program—P.I.Bardelli Alberto, G.L. Regge Daniele.

REFERENCES

[1] P. Bulens *et al.*, "Predicting the tumor response to chemoradiotherapy for rectal cancer: Model development and external validation using MRI radiomics," *Radiother. Oncol.*, vol. 142, pp. 246–252, 2020, doi: 10.1016/j.radonc.2019.07.033.

[2] I. Petkovska *et al.*, "Clinical utility of radiomics at baseline rectal MRI to predict complete response of rectal cancer after chemoradiation therapy," *Abdom. Radiol.*, vol. 45, no. 11, pp. 3608–3617, 2020, doi: 10.1007/s00261-020-02502-w.

[3] F. Coppola *et al.*, "Radiomics and Magnetic Resonance Imaging of Rectal Cancer: From Engineering to Clinical Practice," *Diagnostics*, vol. 11, no. 5, p. 756, 2021, doi: 10.3390/diagnostics11050756.

[4] V. Giannini, S. Rosati, C. Castagneri, L. Martincich, D. Regge, and G. Balestra, "Radiomics for pretreatment prediction of pathological response to neoadjuvant therapy using magnetic resonance imaging: Influence of feature selection," in *Proceedings - International Symposium on Biomedical Imaging*, 2018, vol. 2018-April, pp. 285–288, doi: 10.1109/ISBI.2018.8363575.

[5] S. Rosati, C. M. Gianfreda, G. Balestra, V. Giannini, S. Mazzetti, and D. Regge, "Radiomics to Predict Response to Neoadjuvant Chemotherapy in Rectal Cancer: Influence of Simultaneous Feature Selection and Classifier Optimization," in *2018 IEEE Life Sciences Conference (LSC)*, Oct. 2018, pp. 65–68, doi: 10.1109/LSC.2018.8572194.

[6] X. Fang, S. Xu, B. J. Wood, and P. Yan, "Deep learning-based liver segmentation for fusion-guided intervention," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 6, pp. 963–972, 2020, doi: 10.1007/s11548-020-02147-6.

[7] B. Lee, N. Yamanakkanavar, and J. Y. Choi, "Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture," *PLoS One*, vol. 15, no. 8 August, pp. 1–20, 2020, doi: 10.1371/journal.pone.0236493.

[8] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nat. Methods*, vol. 16, no. 12, pp. 1233–1246, 2019, doi: 10.1038/s41592-019-0403-1.

[9] V. S. Dhaka, G. Rani, M. G. Oza, T. Sharma, and A. Misra, "A deep learning model for mass screening of COVID-19," *Int. J. Imaging Syst. Technol.*, vol. 31, no. 2, pp. 483–498, 2021, doi: 10.1002/ima.22544.

[10] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases," *Comput. Biol. Chem.*, vol. 97, no. December 2021, p. 107619, 2022, doi: 10.1016/j.compbiolchem.2021.107619.

[11] F. Knuth *et al.*, "MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts," *Acta Oncol. (Madr).*, vol. 0, no. 0, pp. 1–9, 2021, doi: 10.1080/0284186x.2021.2013530.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[13] R. G. H. Beets-Tan *et al.*, "Magnetic resonance imaging for clinical management of rectal cancer: Updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting," *Eur. Radiol.*, vol. 28, no. 4, pp. 1465–1475, Apr. 2018, doi: 10.1007/s00330-017-5026-2.

[14] S. Shinya, T. Sasaki, Y. Nakagawa, Z. Guiquing, F. Yamamoto, and Y. Yamashita, "The efficacy of diffusion-weighted imaging for the detection of colorectal cancer," *Hepatogastroenterology.*, 2009.

[15] J. Panic *et al.*, "A Convolutional Neural Network based system for Colorectal cancer segmentation on MRI images," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Jul. 2020, vol. 2020-July, pp. 1675–1678, doi: 10.1109/EMBC44109.2020.9175804.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *ImageNet Classif. with Deep Convolutional Neural Networks*, pp. 1097–1105, 2012, doi: 10.1061/(ASCE)GT.1943-5606.0001284.

[17] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," pp. 1–20, 2018.

[18] R. O. Duda, P. E. Hart, D. G. Stork, and J. Wiley, "Pattern Classification All materials in these slides were taken from Pattern Classification (2nd ed)," no. April, 2016.

[19] S. Hochreiter and P. Frasconi, "Gradient Flow in Recurrent Neural Nets : The Difficulty of Learning Long-Term Dependencies. A Field Guide to Dynamical Recurrent Neural Network," *A F. Guid. to Dyn. Recurr. Neural Netw.*, pp. 401–403, 2001, doi: 10.1109/9780470544037.ch14.

[20] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10553 LNCS, pp. 240–248, doi: 10.1007/978-3-319-67558-9_28.

[21] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015.

[22] P. K. Sahoo, S. Soltani, and A. K. C. Wong, "A survey of thresholding techniques," *Comput. Vision, Graph. Image Process.*, vol. 41, no. 2, pp. 233–260, 1988, doi: 10.1016/0734-189X(88)90022-9.