Automatic segmentation of regions of interest with Deep Learning for postoperative endometrial carcinoma treatment

Author: Arnau Andrés Rodríguez
Physics Faculty, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Aida Niñerola Baizan

Abstract: This project aims to evaluate deep learning algorithms' suitability to correctly delineate the regions of interest on computer tomography images for dosimetric computations, in the context of postoperative endometrial carcinoma treatment. To achieve this goal, the project includes the complete training and evaluation of two deep learning networks. Furthermore, a qualitative assessment of the predicted dosimetric computations and a post-processing of the predicted results have been conducted.

I. INTRODUCTION

A. Clinical context

Endometrial cancer[1] is the most frequent type of gynaecological tumour among women for which postoperative prophylactic vaginal brachytherapy (**VBT**) is the usual treatment modality. The chosen treatment comprises a combination of surgery and pelvic radiotherapy followed by the application of VBT.

Prophylactic VBT is a surgical process involving the placement of an applicator inside the vagina that delivers the required radiation dose. The applicator is composed of a series of interconnected cylinders that are designed to efficiently adapt to the vaginal anatomy of different patients and can be easily placed. These cylinders have diameters between 2 and 3.5 cm.

The first step of the process is the acquisition of Computer Tomography images (**CT**) or Magnetic Resonance images (**MRI**) over which the treatment planning will be prepared. Applicators are made of materials that do not cause artefacts to CT, and also, that are compatible with MRI.

The obtained images are 3D volumes made up of slices whose thickness goes from 1 mm to 3 mm. These images are afterwards sent to the planner with which the radiotherapist specialist delineates organs at risk (**OAR**) and the clinical target volume (**CTV**). Radioactive source trajectory is also reconstructed inside the applicator. CTV must include the proximal third of the vagina while it encompasses a usual length of 2.5 and 4 cm maximum. The anatomical treated volume is the mucous membrane of the vaginal vault, including the surgical scar. According to some authors, almost all of the vaginal walls are also included. It is worth noting that 90% of reappearances occur in the vaginal vault.

OAR and vagina's vault delimitation are done manually which implies a series of uncertainties: intraobserver, i.e, the same radiotherapist might delimit different volumes at different times in case the task is repeated or interobserver, which means that different radiotherapists will delimit these volumes in slightly different ways.

At this stage is when the usage of Artificial Intelligence (**AI**) algorithms may prove to be remarkably helpful owing to their faster performance and their more reproducible and systematic results as a consequence of their learning process.

The main goal of this project is to assess the viability of Deep Learning (**DL**) algorithms for this specific segmentation task. To this end, the complete training and evaluation of two DL models was undertaken as well as the arrangement of a solid data set to train both algorithms.

II. METHODS

A. Data set

The used data set was acquired and provided by the radiotherapy unit of the Clinic Hospital of Barcelona and it covered a range of time of 7 years extending from 2014 to 2021. It consists of CT images of 220 patients in DICOM format[2] (Digital Imaging Communication in Medicine), along with their corresponding RT-Struct[2] (Radiation Therapy Structure) file containing all the information related to the structures of the patients. A first process of classification, rearrangement and anonymisation of the data set was done[2] followed by the conversion to NIfTi[2] format of all the files so they could be handled with standard processing packages.

Once achieved, the next step was the manual revision of all the included segmentations, particularly those involving the vaginal vault, which is the region of interest (**ROI**). Each patient did not have the same OAR delineated nor the same label's name for the CTV region. Therefore, a manual revision was essential to assure the quality and consistency of the segmentation. Moreover, clinical criteria was offered to ensure the requirements of the radiotherapists. The indications were as follows:

- Vaginal vault's segmentation had to only cover the outer surface of the first applicator's cylinder (**FIG.1**), which is meant to be the radiated zone and where dosimetric computations are done (vaginal vault's mucous). The interior of the cylinder is excluded.

- Neither lack of segmentation amidst the ROI's slices nor delineated volumes out of the ROI were acceptable (**FIG.1**).

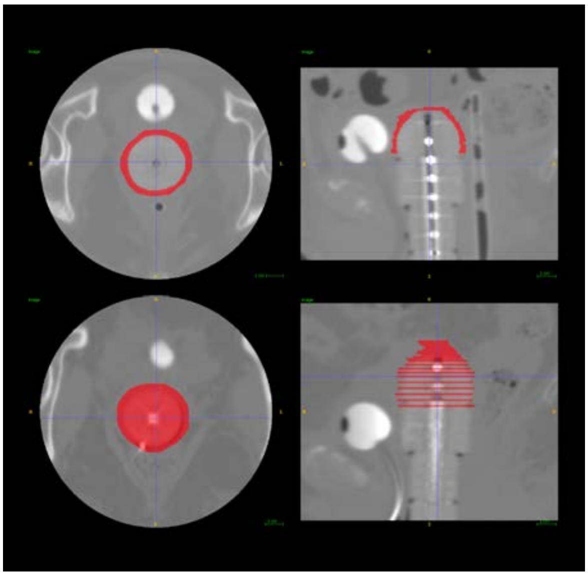

FIG. 1: Top images correspond to a correct case, while bottom ones correspond to an incorrect case. Left and right images display an axial and a sagittal view of the ROI respectively.

B. Nets

All DL networks were implemented with the use of MONAI[3] and Pytorch[8]. The used nets have been V-Net[4] and UNETR[5] that stands for UNET-TRansformers. Despite their similarity to U-Net's[4–6] architecture, they have a series of improvements and changes, including:

- The fully convolutional architecture of V-Net, which also replaces pooling operations with convolution ones. This modification produces a smaller memory footprint during training depending on the segmentation task.

- V-Net's utilisation of residual functions instead of classic stochastic gradient descent. This choice leads to a faster convergence and improved segmentation results.

- UNETR's implementation of transformers as encoders. This choice leads to a more optimal learning of global context and capturing of long-range dependencies.

- UNETR's demonstrated higher efficacy in segmenting small organs and more accurate boundary segmentation. These characteristics are of the utmost importance for this segmentation task at hand.

C. Quantitative metrics for evaluation

Evaluation metrics serve as a means to evaluate nets' performances and results. When selecting the adequate metrics to evaluate our segmentation task, there are several key properties from the ROI to consider, as outlined in [7]. These are:

- Absence of outliers, which are small wrongly delineated regions outside the ROI and that do no not belong to it.

- The significantly smaller size of the ROI when compared to the background's size. This property also helps in distinguishing and isolating the ROI from the surrounding areas.

- ROI's complex boundary regardless of the applicator's consistent and regular shape. The complexity arises because of the specific anatomical features and structures within the ROI.

Therefore, the most suitable metrics are the Dice Coefficient[7] (**DICE**) and the Hausdorff Distance[7] (**HD**).

DICE is a frequently used overlap based metric and also a reproducibility measure. It is given by the expression:

$$DICE = \frac{2\cdot|V_g^1 \cap V_t^1|}{|V_g^1| + |V_t^1|} \tag{1}$$

with V_g^1 and V_t^1 corresponding to the volume (voxels) delineated from the ground truth input (manually delineated) and the predicted output (automatically delineated) respectively. DICE tends to 1 when the degree of overlapping between both is high while it tends to 0 when they differ significantly.

HD is a spatial distance based metric and serves as a dissimilarity measure. It is specially recommended when contour is of the highest importance. Given two point sets, A and B, HD between them is given by the expression:

$$HD(A,B) = \max(h(A,B), h(B,A)) \tag{2}$$

where $h(A,B)$ is the directed Hausdorff distance and it is given by:

$$h(A,B) = \max_{a\in A}\min_{b\in B}||a-b|| \tag{3}$$

where $||a-b||$ is a norm, in this case, an Euclidean distance. HD is the maximum distance of a set to the nearest point in the other set. HD is very sensitive to outliers, hence, it is recommended the calculation of the 95th percentile as a way to mitigate their impact.

D. Training and evaluation process

Following the standardised workflow in DL segmentation tasks[3, 8], the whole process was split into three main parts: training, validation and testing. After assuring the quality and consistency of the data set, it was divided into three groups to be used in each of the prior stages: Training, Validation and Testing in respective proportions of 70%, 20% and 10%.

The training phase followed the usual process, which involves forwarding the data through the model, computing the loss function between the prediction and groundtruth and backpropagating the loss to update the weights of the model accordingly. The used loss function and optimizer have been Dice loss[3, 8] and Adam algorithm[3, 8]. Each iteration of the training loop is called epoch. For every pre-established number of epochs, a validation process is performed to ensure net's gradual optimization and correct learning. If the loss results improve with respect to the previous validation process, net's parameters are saved.

When the training process is finished, the model is tested with the images from the testing data set. Once the model delivers the output predicted segmentation, a post-processing clusterization is done to it to remove possible outliers. When this last step is completed, the output results are saved into a NIfTi file. Evaluation metrics are computed before and after the clusterization operation.

The effect of data augmentation[3, 8] (**DA**) on the model was also studied. In DL, the amount of available data to train is of the greatest importance, and the larger it is, the better quality segmentations are obtained. DA is a key tool since its purpose is to increase data set size without the need of acquiring new images. It consists[3, 8] in the application of filters and deformations to the images as a means to obtain new cases. Applied transformations were split into two groups as a way to study the sensitivity of the net towards certain images' properties: image filters (**IF**) and spatial transformations (**ST**). IF included a histogram normalisation and the random application of: gaussian noise, gaussian smooth and random adjustment of contrast. ST included a random application of: zoom, axis flip, translation and image rotation. All the transformations and filters were applied taking into account the observed variability of the available data and the usual characteristics of CT images.

E. Qualitative evaluation of the predicted dosimetric computations

Due to the lack of sufficient patients to do an exhaustive and complete statistical analysis, only a qualitative assessment of the obtained results can be provided. To this end, different outputs from both UNETR and V-Net have been selected, assuring an equitable sample from both of them. These outputs were next taken to the planner to compute the pertinent dose distributions.

Two dosimetric parameters are used to evaluate the administered radiation dose. The first one is $D_{90\%}$[1]. It corresponds to the dose received by the 90% of the target volume closest to the radioactive source. It is a good measure of the effective received dose by the total volume, since the usage of the 100% volume has a greater degree of uncertainty associated to it. The second one is D_{2cc}[1]. It accounts for the dose received by the 2 cm^3 more exposed from the OAR. Particularly, it is the vaginal surface in direct contact with the applicator. D_{2cc} has a good correlation with toxicity.

III. RESULTS AND DISCUSSION

The first arisen technical issue was the inadequacy of the standard python library, called *dcmrtstruct2nii*[2], to correctly convert this type of segmentation from RT-Struct format to NIfTi format. The reason was that it filled the inner part of the cylinder as it is shown in the bottom left image of **FIG.1**. This was solved with its replacement by another library, *DicomRTTool*[2], which delivered successful conversions. Next, a first manual revision was undertaken leading to the rejection of 34 cases that did not fulfil clinical criteria. An example is displayed in the bottom right image of **FIG.1**. After a subsequent visit to the radiotherapy unit, 23 cases could be fixed and, as a result, our final data set consisted of a total number of 208 cases. Following the procedures indicated in the former section, cases were split as follows: a training data set of 146 cases, a validation data set of 41 cases and a testing data set of 21 cases.

Another key aspect was to check the efficacy of the clusterization, which was achieved thanks to HD metric. Given HD's strong sensitivity to outliers[7] it is advisable to not apply it directly in segmentation tasks of medical images, in view of their usual amount of noise and outliers. Nonetheless, in this task it has demonstrated to be an effective way to measure clusterization's effectiveness. Taking as an example the training corresponding to no DA applied and 1200 epochs, it is seen in both V-Net and UNETR that their mean HD±SD, with SD standing for Standard Deviation, significantly diminishes from 136.3±81.4 mm to 8.7±3.3 mm in the case of UNETR, and from 37.22±33.22 mm to 8.18±3.54 mm in the case of V-Net. In contrast, DICE is not effective to detect this kind of error since its value does not significantly vary, in view of outliers' little volume that does not greatly affect the degree of overlapping. In UNETR's case, from 0.74±0.07 to 0.78±0.07, and in V-Net's case from 0.76±0.07 to 0.79±0.07.

A. Nets' results

Metrics' results for both nets are now introduced. DICE and 95 % HD obtained values for each training configuration and net are presented in **FIG.2**. In this section, out-

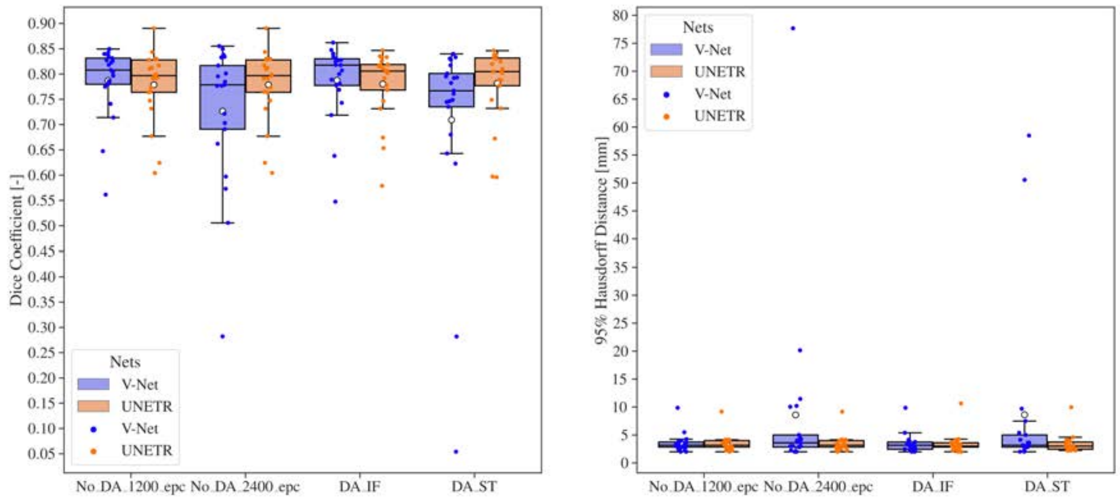

FIG. 2: Figure presents each training's DICE results (left) and 95% HD (right) for both used nets. Boxplots displaying each net's resulting distribution are combined with stripplots indicating each result for the testing data-set. White dots indicate the distribution's mean value. Even though the y-axis' range of the right figure is not the most optimal one to present the data, the way it is displayed serves the purpose of showing the big number of outliers, specifically, in V-Net's results.

liers now refer to the points differing significantly from the rest of the data set. Both nets have shown DICE mean values greater than 0.70 in all training configurations, which is a fair degree of overlapping. In relation to DICE, V-Net exhibits the greater variability among training configurations, delivering the best mean value in the training configuration corresponding to DA with IF. With respect to UNETR, the latter presents a constant value for the mean DICE in each training configuration of 0.78±0.07 and littler variability (narrower distributions) among different training configurations. A striking trend is observed in both nets when more epochs were added to the training process. In the case of V-Net, not only a wider distribution is observed but also a lesser mean value for both DICE and 95% HD, while in contrast, UNETR does not experience any kind of improvement. Great resemblance and low variability among patients' anatomy in the ROI might account for this observed trend, thus making it not advisable to greatly increase the number of training epochs. It is also worth noting that in the case of V-Net, it is clearly visible not solely a greater presence of outliers but also a worst score from both metrics for them. When it comes to the effects of DA on the performances of the nets, V-Net prominently shows a noticeable sensitivity to DA, being the one focused on applying IF, the most beneficial one when compared to DA centred on applying ST, which does not provide an enhancement when compared to no DA with 1200 epochs. This is supported by 95% HD results, as it can be observed by the greater number of outliers. Alternatively, UNETR takes advantage of DA specially in the training configuration focused on ST, as it can be deduced from the narrower distribution achieved in both 95% HD and DICE.

B. Predicted dosimetric computations

The obtained results are presented in **FIG.3** and **TABLE.I**. Despite not having done a sample size computation, a series of results can be inferred. Mean values±SD for $D_{90\%}$ and D_{2cc} are presented in **TABLE.I**. Also, a statistical hypothesis testing has been performed in order to establish dose compatibility between the manual segmentation and the automatic one. The selected test to achieve this task has been Wilcoxon signed-rank test[9] due to the characteristics of our test data, mainly, a sample size with less than 25 samples and the assumption that it is distribution-free (non-parametric). Taking a significance value $\alpha = 0.05$, samples will be compatible if the p-value is greater than 0.05 ($P > 0.05$). The p-value for each parameter is shown in **TABLE.I**. It can be deduced from the obtained results that manual and automatic segmentation are compatible.

	Manual (Mean±SD) [cGy]	**Automatic (Mean±SD) [cGy]**	**p-value**
$D_{90\%}$	807.4±52.3	755.84±109.3	0.095
D_{2cc}	1009.75±63.9	986.3±96.1	0.98

TABLE I: Dosimetric data for the manual and automatic testing data set. P-value calculated with the Wilcoxon signed-rank test has been included.

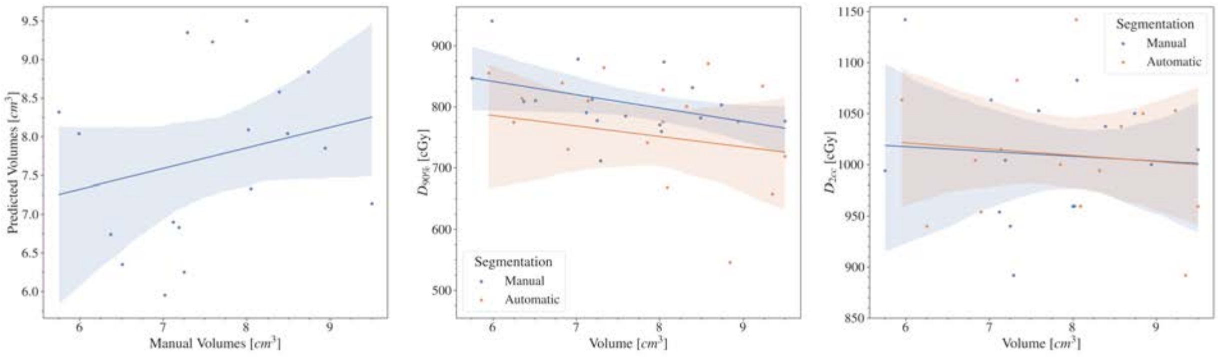

FIG. 3: Figure on the left displays nets' predicted volumes as a function of their corresponding manually traced volume. Centre and right figures display $D_{90\%}$ and D_{2cc} as a function of their corresponding volumes. 95% confidence intervals are represented in all the figures.

IV. CONCLUSIONS

An evaluation of the suitability of DL algorithms for CTV and OAR segmentation in Postoperative endometrial carcinoma treatment has been done. For that purpose, the proper arrangement of a training data set for DL segmentation has been achieved. Also, the training of two different model networks, UNETR and V-Net, has been completed while studying the effects of DA on both nets performances along the process. Dice Coefficient and 95% HD have been the selected metrics to evaluate net's results. Both nets have proven to be capable of accomplishing this task, even though UNETR usage is far more recommended, specially if it is accompanied by DA centred on the use of ST. V-Net can be used if the proper training configuration is guaranteed, i.e, V-Net with a DA centred on IF. It has been also confirmed the adequacy of the clusterization post-process as a means to enhance the quality of the results and remove possible outliers or noise. Besides, the output results have been utilised for dosimetric computations showing a compatibility with the manually computed ones. Finally, I want to mention that this project has offered me a first hand experience of the clinical routine in the radiotherapy unit as well as an opportunity to strengthen my programming skills, specially Python language, and to study in more detail certain topics that were briefly explained to me during the optional courses in Medical Physics and Statistics I attended.

Acknowledgments

First and foremost, I would like to express my utter gratitude to my tutor Aida for her constant support, advice and guidance throughout all the project. I wish to acknowledge Sara, without whose former laborious work this project wouldn't have been possible. Thanks to Adrià, whose expertise, knowledge and aid have given a definite boost to this project. Thanks to Toni and Cristian for all the provided clinical information and specially to Toni for allowing me to visit the radiotherapy unit whenever I was in need. Thanks to Raul, Josep and Roser for all their invaluable counsel. Last but not least, a special thanks to my family and friends whose support has helped me to advance through these arduous years.

[1] Rovirosa. A et al. *Braquiterapia 3D guiada por la imagen* , (EdikaMed S.L., Barcelona 2016, 1st ed.)

[2] Orío, S. (2023). *Building a CT Image Analysis Database for Postoperative Endometrial Carcinoma to Enhance Radiotherapy Treatment Planning* [Master's Thesis, Universitat Politècnica de Catalunya]

[3] Cardoso, M. J. et al. (2022). MONAI: An open-source framework for deep learning in healthcare. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.02701

[4] Milletari, F. et al. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. arXiv (Cornell University).https://doi.org/10.48550/arxiv.1606.04797

[5] Hatamizadeh, A. et al. (2021). UNETR: Transformers for 3D Medical Image Segmentation. arXiv (Cornell University).https://doi.org/10.48550/arxiv.2103.10504

[6] Liu, L. et al. (2020). A survey on U-shaped networks in medical image segmentations. Neurocomputing, **409**, 244–258.https://doi.org/10.1016/j.neucom.2020.05.070

[7] Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging, **15**(1). https://doi.org/10.1186/s12880-015-0068-x

[8] Paszke, A. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv (Cornell University).https://doi.org/10.48550/arxiv.1912.01703

[9] Wang, J. et al. (2023). Evaluation of auto-segmentation for brachytherapy of postoperative cervical cancer using deep learning-based workflow. Physics in Medicine and Biology, **68**(5), 055012. https://doi.org/10.1088/1361-6560/acba76