International Journal of Imaging Systems and Technology

WILEY

RESEARCH ARTICLE OPEN ACCESS

Rectal Cancer Segmentation: A Methodical Approach for Generalizable Deep Learning in a Multi-Center Setting

Jovana Panic[1,2] | Arianna Defeudis[3] | Lorenzo Vassallo[4] | Stefano Cirillo[5] | Marco Gatti[6] | Roberto Sghedoni[7] | Michele Avanzo[8] | Angelo Vanzulli[9] | Luca Sorrentino[10] | Luca Boldrini[11] | Huong Elena Tran[11] | Giuditta Chiloiro[11] | Giuseppe Roberto D'Agostino[12] | Enrico Menghi[13] | Roberta Fusco[14] | Antonella Petrillo[14] | Vincenza Granata[14] | Martina Mori[15] | Claudio Fiorino[15] | Barbara Alicja Jereczek-Fossa[16,17] | Marianna Alessandra Gerardi[17] | Serena Dell'Aversana[18] | Antonio Esposito[19] | Daniele Regge[3] | Samanta Rosati[20] | Gabriella Balestra[20] | Valentina Giannini[3,21]

[1]D3 Center, Osaka University, Osaka, Japan | [2]Premium Research Institute for Human Metaverse Medicine (PRIMe), Osaka University, Osaka, Japan | [3]Candiolo Cancer Institute, FPO-IRCCS, Candiolo (TO), Italy | [4]Department of Diagnostic Imaging and Radiotherapy, AOU Città Della Salute e Della Scienza, Turin, Italy | [5]Department of Radiology, A. O. Ordine Mauriziano (Ospedale Umberto I), Turin, Italy | [6]Department of Surgical Science, Radiology Unit, University of Turin, Turin, Italy | [7]Medical Physics Unit, Azienda USL-IRCCS di Reggio Emilia, Reggio Emilia, Italy | [8]Department of Medical Physics, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Aviano (PN), Italy | [9]Niguarda Cancer Center, Grande Ospedale Metropolitano Niguarda, Milan, Italy | [10]Colorectal Surgery Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milano, Italy | [11]Fondazione Policlinico Universitario'A. Gemelli' IRCCS, Roma, Italy | [12]Department of Radiotherapy and Radiosurgery, IRCCS Humanitas Research Hospital, Milan, Italy | [13]Medical Physics Unit, IRCCS Istituto Romagnolo per lo studio dei Tumori (IRST)'Dino Amadori', Meldola, Italy | [14]Division of Radiology, "Istituto Nazionale Tumori IRCCS Fondazione Pascale – IRCCS di Napoli", Naples, Italy | [15]Medical Physics Dept, IRCCS San Raffaele Scientific Institute, Milan, Italy | [16]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy | [17]Department of Radiation Oncology, IEO European Institute of Oncology IRCCS, Milan, Italy | [18]Department of Radiology, Santa Maria delle Grazie Hospital, Pozzuoli, Italy | [19]School of Medicine, Vita-Salute San Raffaele University, Milan, Italy | [20]Department of Electronics and Telecommunications, Polytechnic of Turin, Turin, Italy | [21]Department of Oncology, University of Turin, Turin, Italy

Correspondence: Jovana Panic (panic.jovana.prime@osaka-u.ac.jp)

Received: 9 December 2024 | **Revised:** 20 February 2025 | **Accepted:** 19 March 2025

Funding: This work was supported by Alleanza Contro il Cancro and AIRC under5 per Mille 2018 ID. 21091 program.

Keywords: automatic segmentation | deep learning | normalizations | preprocessing | rectal cancer

ABSTRACT

Noninvasive Artificial Intelligence (AI) techniques have shown great potential in assisting clinicians through the analysis of medical images. However, significant challenges remain in integrating these systems into clinical practice due to the variability of medical data across multi-center databases and the lack of clear implementation guidelines. These issues hinder the ability to achieve robust, reproducible, and statistically significant results. This study thoroughly analyzes several decision-making steps involved in managing a multi-center database and developing AI-based segmentation models, using rectal cancer as a case study. A dataset of 1212 Magnetic Resonance Images (MRIs) from 14 centers was used. The study examined the impact of different image normalization techniques, network hyperparameters, and training set compositions (in terms of size and construction strategies). The findings emphasize the critical role of image normalization in reducing variability and improving performance. Additionally, the study underscores the importance of carefully selecting network structures and loss functions based on the desired outcomes. The potential of clustering approaches to identify representative training subsets, even with limited data sizes, was also evaluated. While no definitive preprocessing pipeline was identified, several networks developed during the study produced promising results on the external validation set. The insights and methodologies presented may help raise awareness and promote more informed decisions when implementing AI systems in medical imaging.

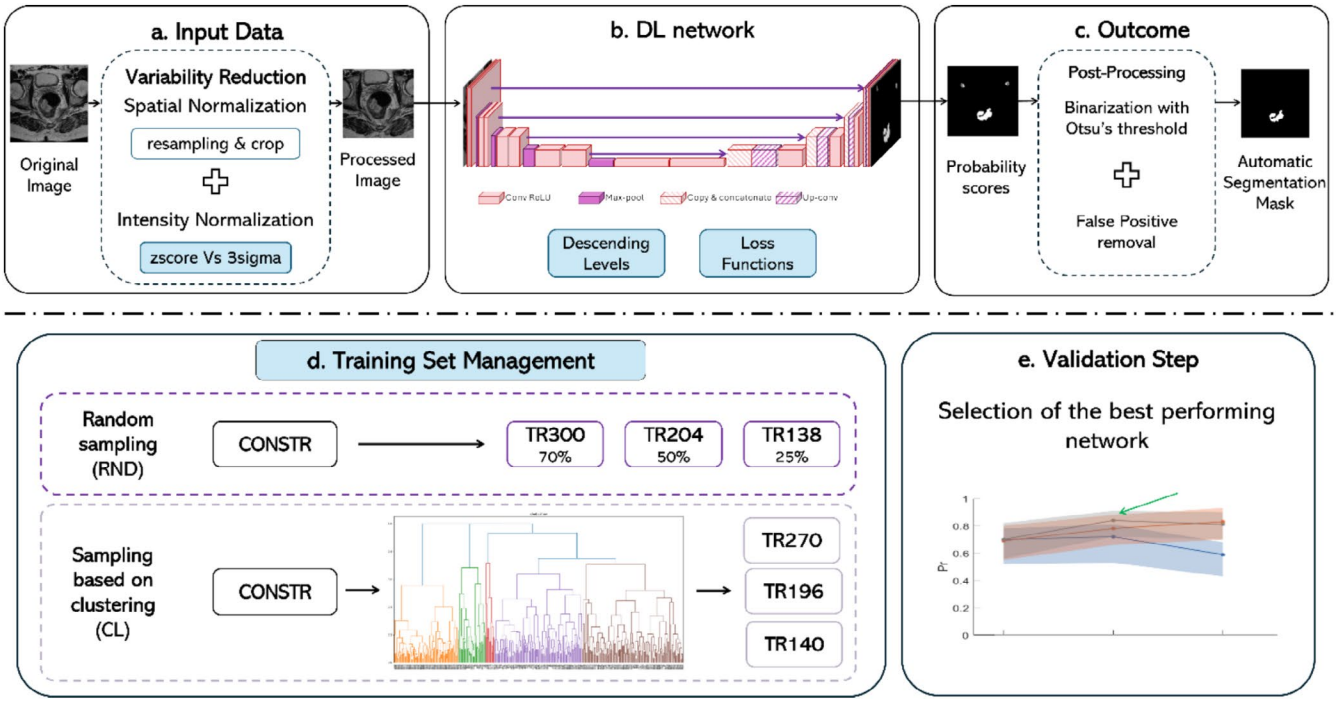

FIGURE 1 | Flowchart of the steps addressed for the development of a DL system: (a) analyses of the input data variability, (b) definition of the DL structure, (c) postprocessing to obtain the automatic segmentation mask, (d) evaluation of the impact of the training set, and finally (e) assessment of the network generalizability validating the results.

1 | Introduction

In recent years, more and more efforts have been made to develop Artificial Intelligence (AI) based systems to support clinicians in noninvasively detecting and characterizing tumors, using medical images. Despite the wide application and encouraging results of recent studies [1–4], there is still a long way to go before these systems can be commonly used in clinical practice either for automatic segmentation or for detection of tumors [5]. Many challenges and limitations still need to be overcome. Among these, the lack of multicenter studies, which are essential for system development and validation, represents a significant hurdle that has proven difficult to overcome [6, 7]. The difficulties involved in conducting large-scale clinical trials are also represented by several technical challenges, as well as legal and administrative issues, which jointly make it difficult to collect images from different institutions [8]. Regarding the technical challenges, the inevitable high image variability between patients' images, related to both biological and nonbiological factors strongly constrain the reproducibility, repeatability, and generalizability of the results [9, 10]. Indeed, it's of fundamental importance to overcome the variability issues to propose a commonly agreed pipeline for Magnetic Resonance Imaging (MRI), a standardization guideline to be followed by each center to achieve compliance between images from different centers. Currently, various guidelines have been proposed to address the aforementioned problem, but only for computed tomography (CT) and positron emission tomography (PET) imaging [9–11].

In the context of AI in medicine, deep learning (DL) algorithms are becoming increasingly prevalent in the development of segmentation tools for organs and tumors. Among various DL systems, the U-Net network [12] has become the backbone of the most widely used architectures due to its distinctive structure decoder–encoder and its promising results [1, 13–15]. Indeed, one of the main advantages of the U-Net is to automatically obtain a probability score map with the same size as the input data, classifying each pixel instead of the whole image in seconds. Even if different structures have been proposed, it is still complex to precisely define the most suitable hyperparameters and training approaches providing robust and generalizable networks to solve the clinical task on real-world data.

Despite several studies having proposed a U-net structure for the segmentation of tumoral areas, few of them have used a multi-center dataset, and even fewer have specified and/or compared different development choices made, crucial for the reproducibility of the methods and subsequent evaluations [16–19]. Therefore, the aim of our study is to assess and evaluate the impact of multiple technical decisions related to the management of a multicenter medical image database, variability reduction, and the development of DL-based systems in the oncological domain. In this way, we want to contribute to filling the gap related to the lack of commonly agreed guidelines providing suggestions and insights for AI applications in medical imaging.

2 | Experimental Setup

In this study, we have addressed some technical aspects related to the different steps of the development pipeline of a DL system (Figure 1). More in detail, we focused on the analyses of the input data variability (Figure 1a), the impact of different U-Net architectures (Figure 1b), and the impact of the size and composition of the training set (Figure 1d).

a)

Scanner	Pixel resolution (M-IQR) (mm)	Slice thickness (M-IQR) (mm)	FOV (M-IQR) (mm)
GE Medical System	0.70 (0.39-1.09)	3.00 (3.00-4.00)	180 (180-220)
Philips Healthcare	0.49 (0.47-0.67)	3.00 (3.00-4.00)	250 (240-280)
Siemens	0.78 (0.59-0.94)	3.00 (3.00-3.00)	250 (200-300)
NA info	0.70 (0.59-0.74)	4.00 (3.60-4.40)	315 (200-405)

b)

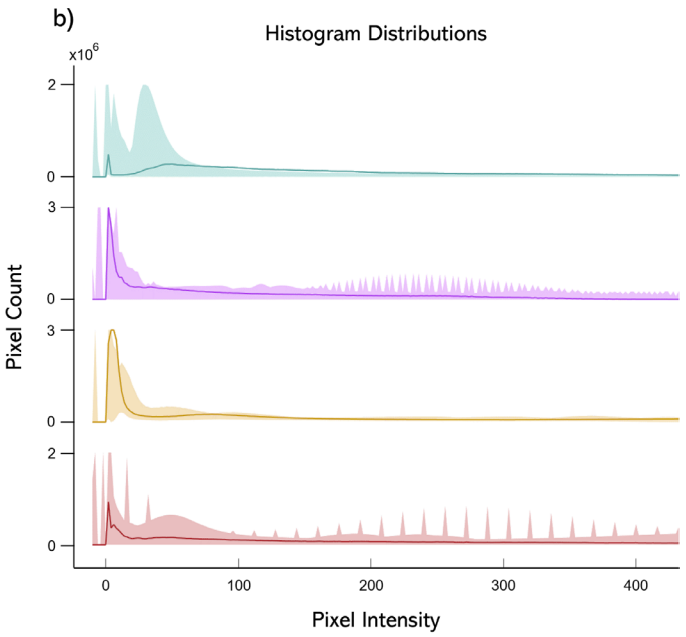

FIGURE 2 | (a) on the left are shown the median (M) and the interquartile range (IQR) of the pixel resolution, slice thickness, and field of view (FOV) of the different vendors; (b) on the right, the histogram distributions of the sequences according to the different vendors. The solid line represents the M intensity histogram distribution, while the colored areas represent the IQR.

3 | Materials and Methods

3.1 | Dataset

For this study, 1212 fast spin-echo axial T2-weighted (T2w) MRI sequences of patients with pathologically proven rectal cancer (RC), acquired before neoadjuvant chemo-radiotherapy after October 2000, were retrospectively collected from 14 different Italian institutions, 10 from the "Alleanza Contro il Cancro (ACC)"—Record project and four from other multicenter collaborations. To ensure the development of a system able to handle heterogeneity and variability of real-world images due to scanners and protocol differences, we decided to follow a vendor-agnostic strategy by stratifying patients according to the manufacturer of the MRI scanner, as follows:

- the construction (CONSTR) and the Internal Validation (IntVAL) sets were composed of sequences acquired with GE and Philips scanners;
- the External Validation (ExtVAL) set included sequences acquired with Siemens or sequences for which we did not have information about the manufacturer (N.A.).

The CONSTR was used to define both a preprocessing step, which is useful for variability reduction and developing the DL network. In contrast, IntVAL and ExtVAL were used to internally and externally validate the systems' performances.

This multicenter retrospective project was approved by the institutional review boards (IRBs) in each institution, with a waiver for the requirement of signed informed consent, as de-identified data were used. All exams were acquired according to MRI guidelines [20] for reporting RC staging.

3.2 | Reference Standard

All tumor volumes were manually segmented on the T2w sequences by different radiologists, one per center, with high experience in reporting MRIs, and then revised by a centralized expert radiation therapist. These segmentation masks were used as ground truth for the development and validation of the automatic segmentation algorithms.

3.3 | Assessment and Reduction of Data Variability

As expected, the dataset was characterized by high variability in terms of both spatial and intensity characteristics, which are both highly dependent on the scanner characteristics and acquisition parameters [21], as shown in Figure 2 and details in Table S1. This can result in very different images in terms of field of view (FOV), i.e., inclusion and exclusion of different anatomical structures, and signal intensities. To address these issues, we defined a preprocessing step that included both spatial and intensity normalization approaches.

3.3.1 | Spatial Normalization

Since the pixel resolution ranges widely among sequences and the DL model requires an input with a fixed dimension, we first applied spatial normalization by resampling all images to the same in-plane resolution, defined as the median resolution of all sequences in the CONSTR (0.47 mm). Then, we centrally cropped the sequences to the same 2D FOV, i.e., $180 \times 180\,mm^2$, as it was the lowest FOV among all vendors and institutions (Table S1). In this way, we obtained images with fixed dimensions (384×384 pixels) and the same FOV, thus reducing the variability of anatomical structures included within the

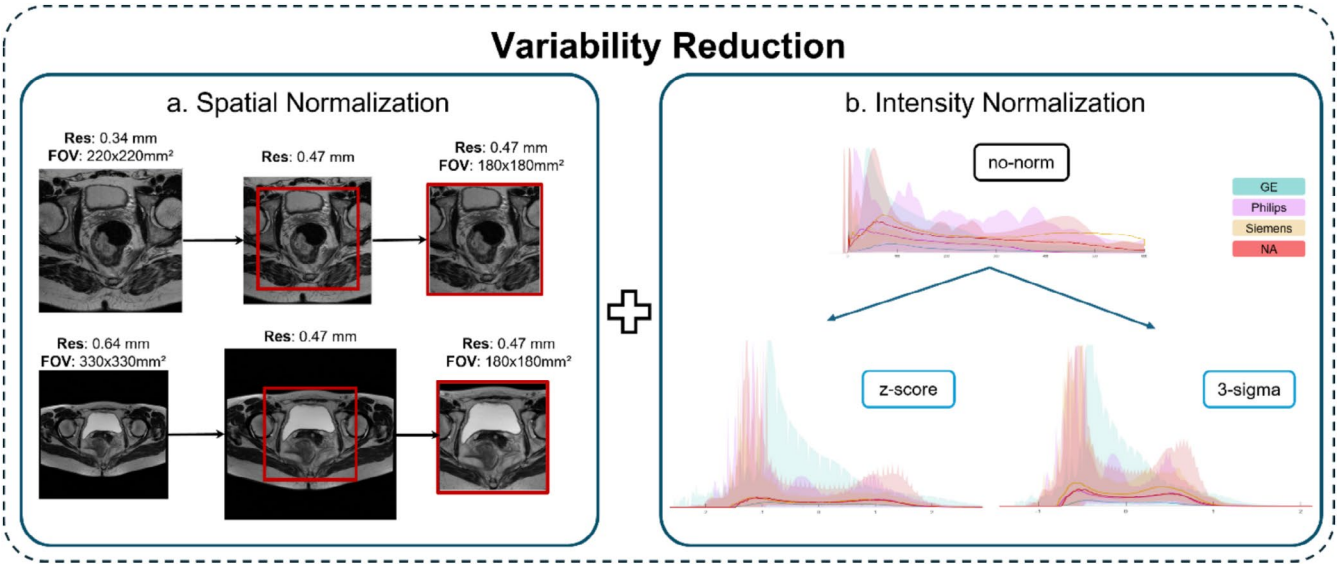

FIGURE 3 | Variability reduction approaches: (a) description of the spatial normalization approach, i.e., resampling and cropping; (b) histogram distributions per each vendor with and without intensity normalization.

sequences. The same spatial normalization was applied to the masks as well (Figure 3a) [22].

3.3.2 | Intensity Normalization

Since it has been demonstrated that among different intensity normalization methods, *z-score* and *3-sigma* were the most suitable to properly reduce the differences in terms of histogram shape and ranges [9, 10], we decided to evaluate their impact on the performances of a DL model. In particular, we compared the *z-score* and *3-sigma* methods to the *no-norm* case (Figure 3b). In addition, we applied the min–max normalization after each normalization method, rescaling the input data intensities between [0,1]. This approach helps reduce computational demands and prevents issues like the vanishing gradient problem [23, 24].

3.4 | Impact of Architectures and Parameters of the DL Network

3.4.1 | Network's Architecture

Among the different parameters useful for defining the U-Net, there is the setting of the proper number of descending levels (or convolutional blocks), according to the difficulty of the task, the input data size, and the target object dimensions. In this case, considering the average dimension of the RC, we compared the performance of U-Nets with 3 and 4 descending levels (Figure 1b). In the proposed U-Nets, all convolutional layers were characterized by a 3×3 kernel and the Rectified Linear Unit (*ReLU*) activation function, except the output layer, which is defined by a 1×1 kernel and the sigmoid activation function [25].

3.4.2 | Loss Functions

Due to the high imbalance between pixels related to the background and pathological ones, we evaluated the impact of three different loss functions, addressing this issue:

- The Binary Focal Loss (BFL):

$$\mathrm{BFL}(g_t, p_r) = -g_t\alpha(1-p_r)^{\gamma}\log(p_r) - (1-g_t)\alpha p_r^{\gamma}\log(1-p_r) \quad (1)$$

where g_t is the ground truth, p_r is the prediction, and α and γ are weightings and modulating factors, respectively.

- The Dice Loss (DL):

$$\mathrm{DL}(g_t, p_r) = 1 - \frac{2g_t p_r + 1}{g_t + p_r + 1} \quad (2)$$

- The Combo Loss (CoL), is obtained by summing (1) and (2):

$$\mathrm{CoL}(g_t, p_r) = \mathrm{BFL}(g_t, p_r) + \mathrm{DL}(g_t, p_r) \quad (3)$$

3.5 | Training Set Management

3.5.1 | Composition of the Training Set

Another fundamental step for the development of robust systems is the training one, in which the network learns directly from the data provided; therefore, it is of key importance to provide datasets that are as representative as possible of the whole target population. For this reason, we analyzed the impact of two training set composition approaches.

- *Sampling-based on random approach*

As the first approach, we assessed the impact of the most used training set (TR) composition procedure, *random sampling*. This method consists of a random selection of a subgroup of elements for the TR, while the remaining cases are included in the test (TS) set [1, 26–28]. We chose to split the CONSTR to have 70% of samples in the TR (300 sequences), and 30% (111) in TS. This set will be called RND.

- *Sampling-based on clustering-based approach*

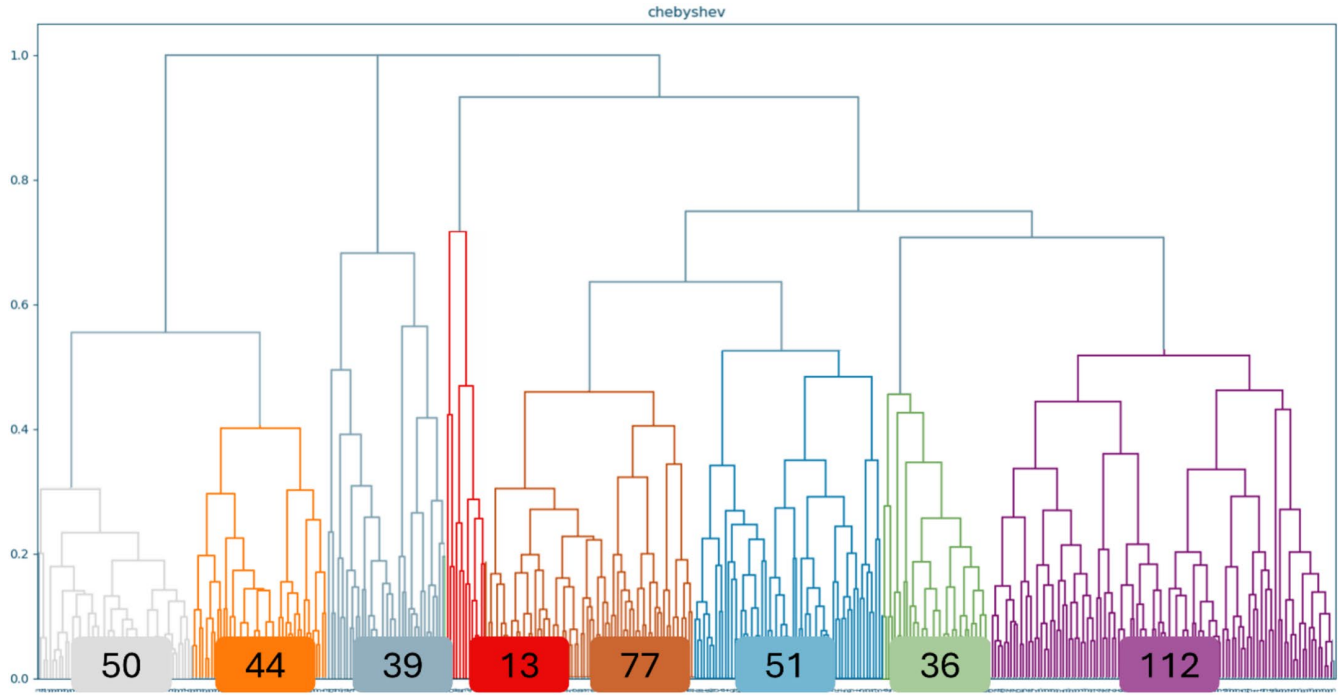

FIGURE 4 | Dendrogram of patients, highlighting the number of samples for each cluster.

This approach is based on an agglomerative hierarchical clustering method that organizes data in a hierarchical tree (dendrogram) based on proximity measures. Then, the final clusters are obtained by cutting the tree at a certain level. To apply this method, we first extracted the following 10 variables: mean, standard deviation, median, 25, and 75 percentiles of both tumoral and whole patient volumes [15]. The distances were evaluated using the Chebyshev metric. We then applied the hierarchical clustering on the patients and cut the tree to have eight clusters (Figure 4).

The cluster with only 13 samples was discarded for the TR construction, and those patients were included in the TS. The TR sets were then created by extracting the same number of patients from each cluster, when feasible, while the discarded ones were included in the TS set. To create a TR that had nearly the same dimension as that obtained with the RND approach, we decided to include 39 patients from each cluster except for the smallest ones ($n = 39$ and $n = 36$), which were entirely included in the TR. This clustering method to construct the TR has been called CL.

In this study, we evaluated the impact of all the above-mentioned parameters, considering both composition approaches. To evaluate the impact of all the mentioned parameters, considering the normalization methods, loss functions, number of descending levels, and training set combinations, a total number of 36 networks were trained. Once the best architecture and parameters were selected, we evaluated the impact of different training set dimensions on the IntVAL.

3.5.2 | Training Set Size

The size of the training itself is another limiting aspect in the development of robust models. In this study, we evaluated the impact of three different dimensionalities, paying close attention to obtaining similar dimensions between the two approaches, i.e., CL and RND. In particular, we first defined three different sizes; then we applied both training composition methods per each, as follows:

1. *Big*: for the RND, we randomly selected 70% ($n = 300$) of the CONSTR dataset for the TR set. For the CL, we randomly selected 39 samples from each cluster except for the two smallest ones that were entirely included, finally including 270 sequences.
2. *Medium*: for the RND, we randomly selected 50% ($n = 204$) of the samples for the TR set. For the CL, we randomly selected 28 samples from each cluster, finally including 196 sequences.
3. *Small*: for the RND, we randomly selected 25% ($n = 138$) of the samples for the TR set. For the CL, we randomly selected 20 samples from each cluster, finally including 140 sequences.

For each training set, we developed three models (called net1, net2, and net3), characterized by different starting random seeds, for a total of 18 trained networks. We decided to start with different starting seeds to assess the impact of the initializations on the training. All networks were trained with 100 epochs of training and batch size 10 (due to the GPU memory available). Additionally, we tried to avoid the overfitting of the networks by stopping the training using the callback function EarlyStopping, monitoring the "val_loss," and maintaining the default value of *min_delta* (absolute change [24]. All analyses were implemented in Python (v. 3.7.4), using the Tensorflow (v. 2.2.0) library, with the Adam optimizer [18] and a starting learning rate value of 0.001, $\beta 1$ of 0.9 and $\beta 2$ of 0.999. The GPU used was NVIDIA Tesla T4 with 16GB of memory.

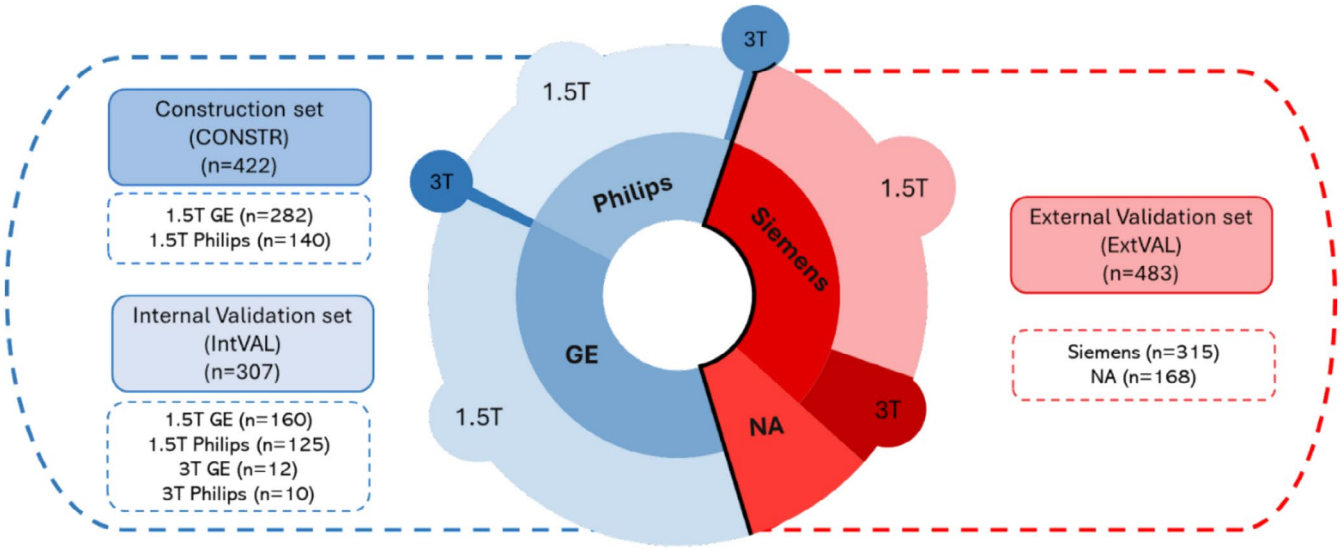

FIGURE 5 | Dataset division into CONSTR, IntVAL, and ExtVAL.

3.6 | Statistical Analysis

To evaluate the performances of the DL models, all the output masks underwent post-processing steps. First, output masks were binarized using Otsu's threshold [29] evaluation, which considered all the patient's slides. Then, volumes spatially connected on <3 slices and having the maximum predicted value lower than the median of the related patient were discarded (Figure 1c). The parameters evaluated were:

- Dice Similarity Coefficient

$$\mathrm{DSC} = \frac{2\mathrm{TP_v}}{2\mathrm{TP_v} + \mathrm{FP_v} + \mathrm{FN_v}} \tag{4}$$

- Precision

$$\mathrm{Pr} = \frac{\mathrm{TP_v}}{\mathrm{TP_v} + \mathrm{FP_v}} \tag{5}$$

- Recall

$$\mathrm{Re} = \frac{\mathrm{TP_v}}{\mathrm{TP_v} + \mathrm{FN_v}} \tag{6}$$

where $\mathrm{TP_v}$ are all true positive voxels within the 3D masks, $\mathrm{FP_v}$ and $\mathrm{FN_v}$ are false positive and false negative (FN) voxels, respectively. The detection rate was computed as the number of correctly detected tumors over the total number of tumors. As commonly used [30], a tumor was considered detected if the DSC between the manual and automatic masks was ≥ 0.10; otherwise, it was defined as a FN. Only detected tumors were included in the evaluation of median DSC, Pr, and Re. Once all models' combinations (data pre-processing, network structure, and training size) have been trained and optimized on the CONSTR and then internally validated on the IntVAL, the best model that achieved the highest DSC and lower percentage of FN tumors was selected. Subsequently, its performances were validated on ExtVAL. To statistically compare the difference between FN we used the Pearson's Chi-Square Test, while for DSC, Pr, and Re distributions, we performed the Mann–Whitney *U*-Test. For *p-values* <0.05, the difference is considered statistically significant. All analyses were performed using Python 3.7, Matlab (R2023a) and MedCalc Software Ltd. 2024.

4 | Results

4.1 | Dataset

The CONSTR set comprised 422 patients, all of whom were acquired using 1.5T scanners. Of these, 282 were acquired with a GE scanner and 140 with a Philips scanner. The IntVAL set included 307 patients, 285 of whom were acquired with a 1.5T scanner and 22 with a 3T one. The ExtVAL set included 483 examinations, 315 of which were acquired with both a 1.5T and 3T Siemens scanner and 168 for which the scanner information was missing (Figure 5).

4.2 | Parameters' Impact Analysis

Considering all the parameters' combinations, we trained 36 networks. Figure 6 shows the impact of the different normalizations' strategies, loss functions, and descending levels on the segmentation performances (DSC, Pr, and Re) obtained on the IntVAL with different training compositions, cluster, and random. The details of the performance of all combinations and networks are presented in Table S2.

Image normalization doesn't have a strong impact on DSC only with the 4-level U-Nets, while it has a strong impact on the 3-level U-Net. In this case, results as high as the 4-level are reached only with either *3-sigma* or *z-score* and RND training set and *no-norm* in combination with the CL approach (Figure 6a). The combination of *no-norm* and RND generally exhibits the lowest results (DSC ranging from 0.48 to 0.68). An important assumption can be made from this analysis regarding the importance of the choice of patients in the training set, especially when the dataset is heterogeneous (*no-norm*). Indeed, in this case, if we

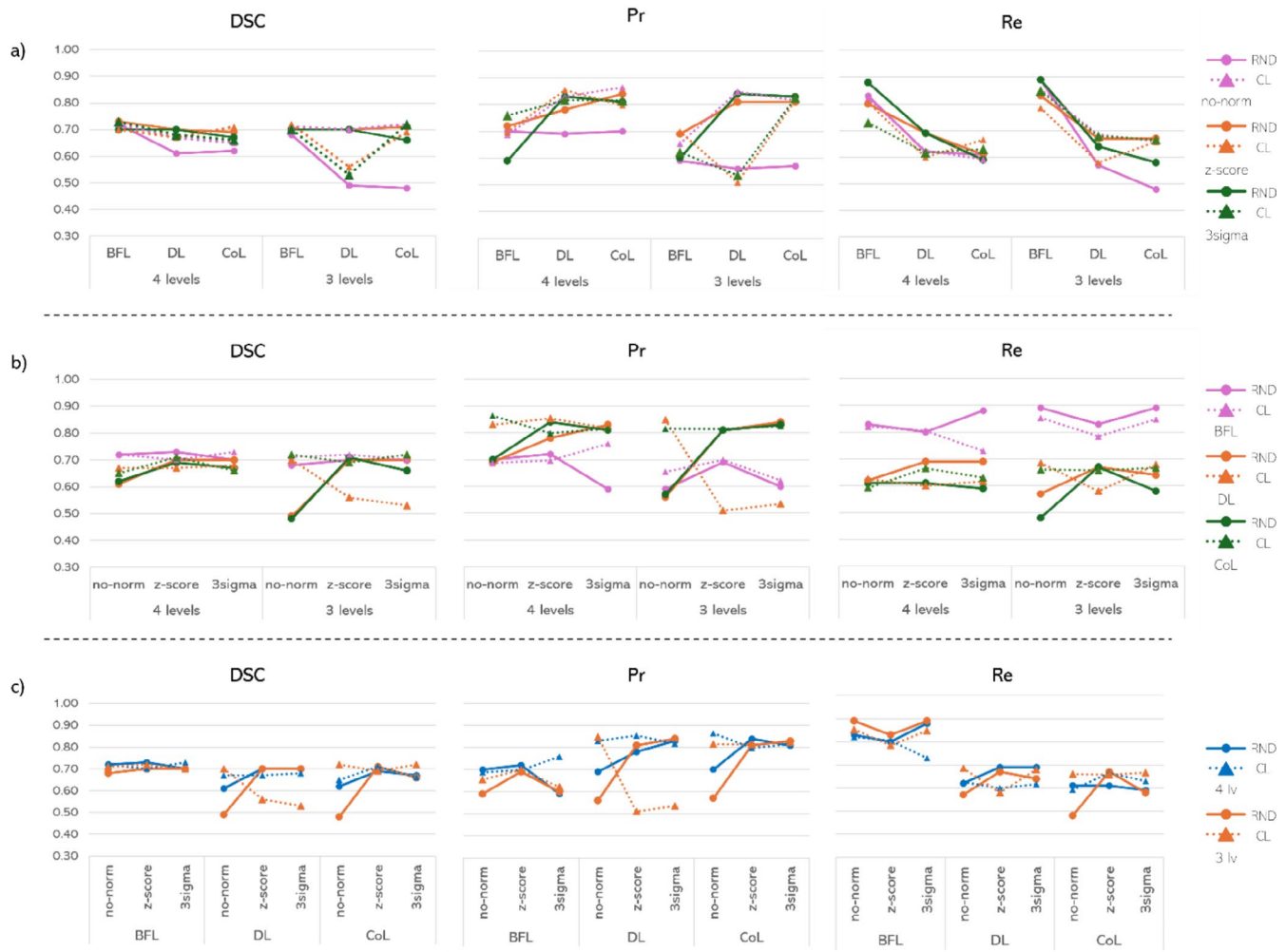

FIGURE 6 | Impact of the normalization approaches (6a), loss functions (6b), and descending levels (6c) on the segmentation performances (DSC, Pr, and Re) obtained on the IntVAL. The solid lines represent the median values across all patients using a random approach (RND) and the dotted line with a cluster one (CL).

apply the CL approach to construct the training set, we can increase DSC up to 20%. The same considerations are true also considering Pr and Re; however, these metrics are largely affected by other parameters and the number of descending levels. In general, the best results in terms of Pr among image normalizations were reached by *z-score* and *3-sigma* (*z-score* ranging from 0.72 to 0.85 and *3-sigma* 0.60 to 0.83).

Focusing on the impact of the loss functions (Figure 6b), it can be observed that they strongly affect Pr and Re. In particular, BFL reached one of the lowest Pr regardless of the number of levels, image normalizations, and training set composition approach (from 0.59 to 0.72 for 4-level and 0.59 and 0.70 for 3-level). Again, the CL method can smooth out differences between images, especially in the *no-norm* dataset for certain configurations, e.g., DL and CL reached Pr > 0.8 both in the 3-level and 4-level networks. In general, except for DL, Pr is similar between different normalization methods for BFL and CoL. However, it can be observed that BFL (both for RND and CL) generally yields the lowest results (0.59–0.72 for 4-level). In contrast, DL and CoL yield similar higher results (0.80–0.85), particularly with the CL approach. Re and Pr exhibit inverse behavior, reflecting the contrasting nature of the concepts they represent. Re imposes a penalty for under-segmentations, whereas precision penalizes over-segmentations.

Lastly, the number of levels' impact is not as evident as for the other parameters (Figure 6c). However, what is clear is that 4-level networks obtained more robust results across loss functions, image normalization, and training set compositions. In this case, BFL combined with *z-score* reached the highest results in terms of DSC with a good compromise between Pr and Re for both RND and CL training, i.e., DSC of 0.73 and 0.70, Pr 0.72 and 0.70, and Re 0.80 and 0.81, respectively. Conversely, the behavior is more variable when considering the 3-level networks, which reached good results only when trained with BLF (Re from 0.83 to 0.89, DSC from 0.68 to 0.73).

Table 1 shows the percentage of FN in IntVAL yielded by each combination of the different parameters. The 3-level shows a statistically significantly higher incidence of FN, with an average of 26.2% ± 12.3% compared to 16.7% ± 6.8% for the 4-level. Focusing solely on the 4-level networks, the mean percentage of FN obtained by RND and CL training

TABLE 1 | Percentages of false negative (FN) patients in the internal validation set (IntVAL) per each combination of loss functions, normalizations, descending levels, and training compositions.

IntVAL	4 levels						3 levels					
	RND			CL			RND			CL		
	no-norm	z-score	3-sigma	no-norm	z-score	3-sigma	no-norm	z-score	3-sigma	no-norm	z-score	3-sigma
BFL	9	20	5	16	8	13	14	15	11	15	27	17
DL	34	15	18	22	17	23	30	35	33	33	40	48
CoL	23	11	22	17	15	12	53	16	15	23	19	27

legend: 0 5 10 15 20 25

Note: The color legend reflects performance in terms of percentage of FN: green denotes better performance, whereas red indicates worse performance.
Abbreviations: BFL: Binary Focal Loss, CL: clustering, CoL: Combo Loss DL: Dice Loss, IntVAL: Internal Validation, RND: random.

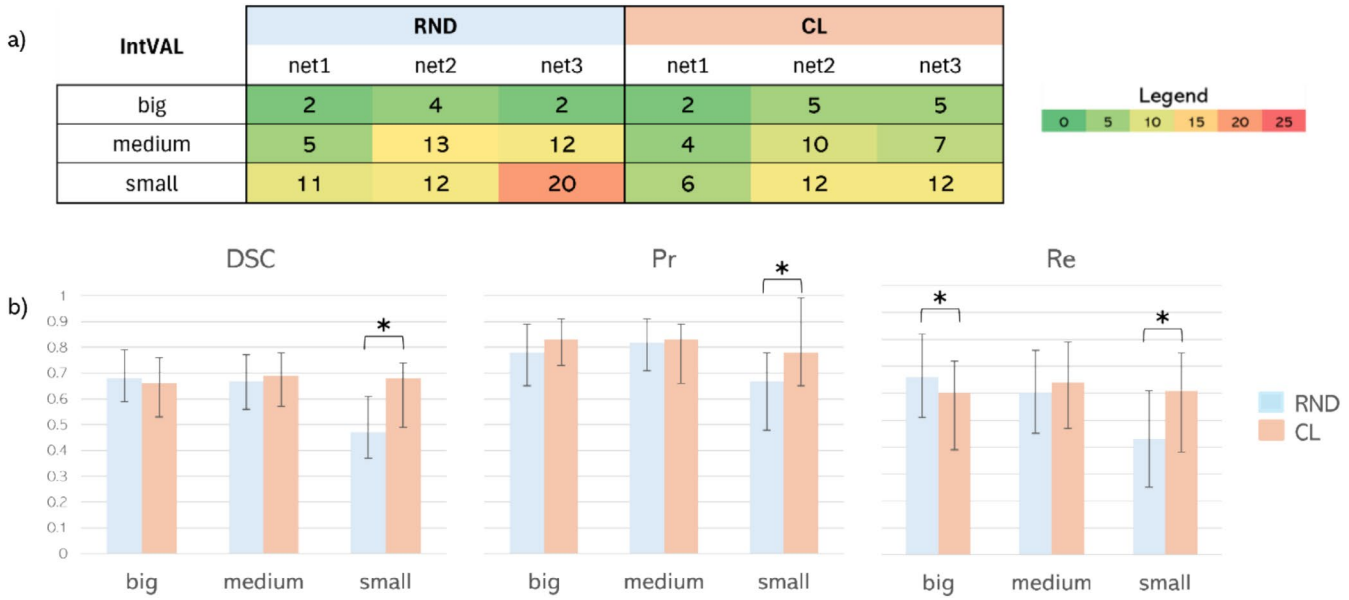

FIGURE 7 | (a) False negative (FN) patients' percentage per combination in the internal validation set. (b) Bar diagrams showing the networks' performance distributions, divided according to the procedure followed for the training construction (CL and RND) and size (big, medium, and small). *Represents statistically significant differences.

compositions is comparable (17.4% vs. 15.9%), while the standard deviation exhibits a notable discrepancy (8.7% vs. 4.7%). However, the overall performances are not statistically significantly different (*p-values* > 0.05). This may indicate that the CL approach allows for more stable performance despite the different parameters. Focusing on the normalization methods, it was found that the *no-norm* exhibited an average FN value of 20.2% ± 8.4%, the *3-sigma* 15.5% ± 6.8%, and the *z-score* 14.3% ± 4.3%, highlighting that the latter allows for a lower percentage of FNs on average as the loss functions vary. BFL, as we expected from high values of Re at the cost of Pr, is prone to over-segmentation, resulting in it having the lowest mean percentages of FNs: 11.8% ± 5.6% vs. 21.5% ± 6.8% and 16.7% ± 5.0% of DL and CoL.

Regarding the sample size, it is evident that an increase in the number of patients in the training set will result in a reduction in the number of FNs (Figure 7a) and this can be attributed to the fact that the samples included in the TR more accurately reflect the heterogeneity that characterizes the RC. This inference is further substantiated by the fact that the approach employed in the construction of the training set (RND vs. CL) dataset exhibited varying effects contingent on the size of the training database. The details of the performance of all combinations and networks are presented in Table S3.

In the big TR size, there are no statistically significant differences between the two construction methods (*p-value* < 0.05); on the contrary, when dealing with smaller dimensions, the CL approach yields better performances in terms of FN percentage (Figure 7a). The percentage for medium TR ranges between 5%–13% and 4%–10% for RND and CL approaches, respectively, while for small TR, it ranges from 11% to 20% and from 6% to 12%, respectively.

Figure 7b shows the bar diagram related to the DSC, Pr and Re metrics. Focusing on the impact of the two composition approaches, no statistically significant differences have been found between big and medium sizes, except for Re considering big size. On the contrary, the results yielded by the CL approach are statistically significantly higher than those yielded by the RND one (*p-value* < 0.05) for the small size. Moreover, they are comparable

TABLE 2 | Best model performances in terms of DSC, Pr and Re for both validation sets.

Dataset	FN %	DSC median (IQR)	PR median (IQR)	RE median (IQR)
IntVAL	2	0.68 (0.20)	0.78 (0.24)	0.66 (0.31)
ExtVAL	4	0.66 (0.24)	0.75 (0.25)	0.61 (0.36)

Abbreviation: IQR: InterQuartile range.

with those yielded by the bigger training set sizes, having always *p-values* > 0.05. In detail, Table S4 shows the *p-values* for each combination of size and construction approach for all metrics. This supports the hypothesis that when dealing with small-sized TR sets, the CL approach may allow the construction of more representative training sets than RND, yielding comparable results in all metrics with respect to big sizes.

In conclusion, from our analysis, we observed, as we expected, that increasing the size of the TR leads to an improvement in the ability networks to detect the tumoral volumes, without strongly affecting its ability to precisely segment it.

4.3 | External Validation

The optimal configuration was identified as one that achieves a satisfactory FN percentage and a high Pr value on the IntVAL. This choice was driven by the consideration that the clinical goal is to precisely and reliably detect the tumoral area, without including too many nonpathological pixels. Therefore, the selected model was the one that used *z-score* as an image normalization technique, whose U-Net was characterized by four descending levels and that used CoL as a loss function during the training, and the big size TR set. This model achieved a DSC of 0.68, Pr of 0.78, and Re of 0.66 for RND and 0.66, 0.83, 0.60 for CL in IntVAL. Since there are no statistically significant differences between the two models for the big training set size, we considered the model trained with the RND set.

The selected network was externally validated on the 483 examinations, and results comparable to the IntVAL (Table 2) proved the ability of the model to generalize its performance also on an external dataset. Figure 8 presents some examples of segmentations provided by the best-performing model on some patients of ExtVAL. The first row shows an example of good segmentation (all metrics), while the second and third rows show two cases with DSC < 0.6 that exhibit divergent patterns regarding Pr and Re. The one in the second row (id. 015) depicts a relatively under-segmented tumor, for which an accuracy of 0.76 is attained as opposed to a Re of 0.18 (Figure 8b). Conversely, the last one exemplifies a case of over-segmentation for whom we obtained a Re of 0.97 and a Pr of 0.32 (Figure 8c).

5 | Discussion and Conclusion

In this work, we proposed a standard pipeline for the development of an automatic DL-based system for RC segmentation on T2w sequences, providing insights related to the impact of multiple decisional parameters, using images from 14 different centers. To our knowledge, no other studies have conducted a comprehensive analysis and comparison of technical parameters such as intensity normalization methods, network architecture, loss functions, and training sets on abdominal MRI.

Focusing on the image preprocessing step, we demonstrated the complexity of defining the most suitable approach. Our previous study [10] revealed that spatial normalization, characterized by cropping and resizing, has emerged as a principal method to reduce spatial variability. Conversely, among studies on automatic RC segmentation, there is still no consensus on the optimal approach for intensity variability reduction [16–19]. For these reasons, we analyzed the effect of the two most frequently used image normalization techniques (*z-score* and *3-sigma*) compared to *no-norm* condition, and we demonstrated that both *z-score* and *3-sigma* normalizations improved the overall performance of the system: the Pr has been increased from 0.58–0.71 for the *no-norm* to 0.72–0.85 and 0.60–0.83 for *z-score* and *3-sigma*, respectively. Although both normalization approaches show potential in reducing intensity variability by realigning the histogram distributions, selecting the optimal method must still be done carefully, tailoring it to the specific dataset and methodology. Additionally, ensuring a consistent rescaling of input intensities to the [0,1] range is highly recommended to mitigate the vanishing gradient problem.

Following the DL development pipeline, we then evaluated how the network architecture and loss function significantly influenced the system's ability to accurately detect tumor volumes. On one hand, we demonstrated that an encoder design consisting of four convolutional blocks yields better results compared to three blocks (DSC ranging from 0.61 to 0.75 vs. 0.49 to 0.73), indicating greater robustness. On the other, we demonstrated that the impact of the overall performance in terms of DSC is not strongly affected by the different loss functions, with values ranging between 0.70–0.73 for BFL, 0.61–0.70 for DL, and 0.62–0.71 for CoL. However, we have proved the potential of combining multiple loss functions to obtain more precise segmentation systems, with an overall improvement of the10% for Pr compared to BFL. No previous studies have investigated the impact of different loss functions on RC segmentation. Instead, most works [17, 19] simply reported the selected loss function without explicitly justifying their choice. This lack of transparency significantly limits the reproducibility of results and a comprehensive understanding of the role played by loss functions in medical image analysis.

Finally, we analyzed the impact of both sample size and the construction method of the TR set on system performance. As expected, smaller TR sets led to lower performance compared to larger ones. However, when data availability was limited, a clustering-based construction approach allowed for DSC values to remain consistent across different sample sizes (0.67, 0.69, and 0.69 for large, medium, and small sets, respectively). Additionally, this method increased Pr by 10% compared to random selection, as it ensured the inclusion of more representative data. This finding is particularly relevant for developers working

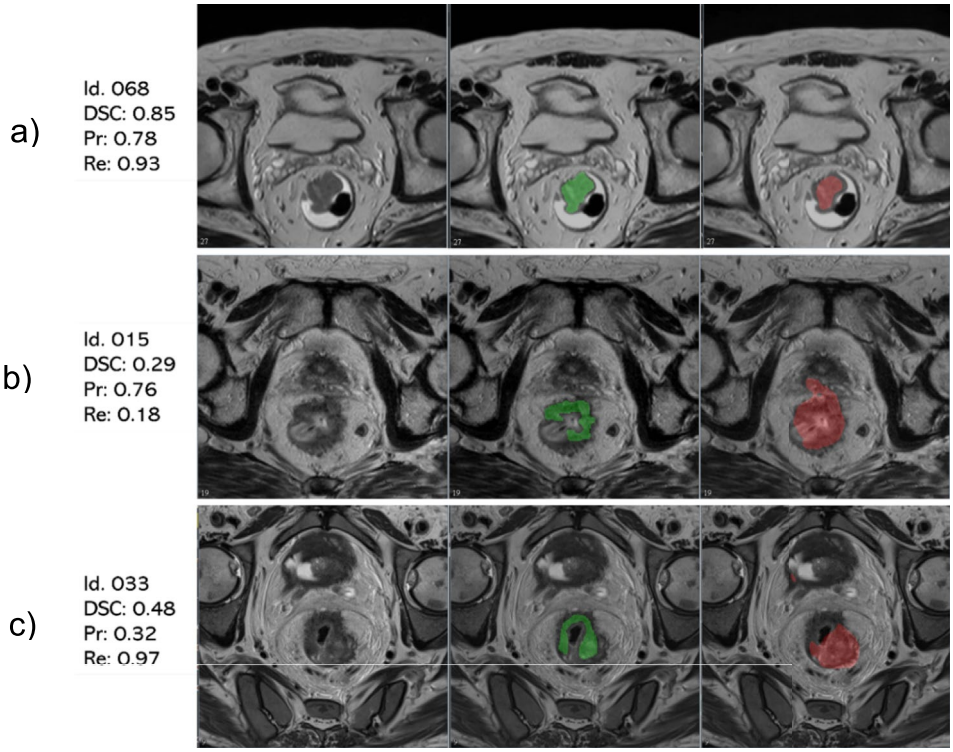

FIGURE 8 | Examples of segmentation masks of rectal cancers belonging to the ExtVAL. In the first column the clear T2w, the second presents the manual segmentation in green, and the third row shows the automatic segmentation in red.

with limited datasets, a common scenario in medical image analysis. Therefore, before applying *Data Augmentation*, which may introduce overfitting, we suggest considering this strategy as an alternative. In the literature, there are only a few papers that have conducted analyses on TR construction, mainly on the brain, and none, as far as we know, on the rectum. On the brain volumes, Narayana et al. [31], Wulms et al. [32], and Fang et al. [33] showed that the relationship between size and DL performance depends on the characteristics of the volume of interest. Despite the different body parts, our study also suggests that an increase in the TR size (at least 300 sequences) could provide better results for the RC segmentation.

Overall, our results demonstrate that following the reported pipeline results in a performance improvement of up to 25% in terms of DSC. Our optimal model shows generalizability across different institutions and MRI scanners (DSC of 0.66 on ExtVAL), yielding similar results as Wang et al. [27] and Pang et al. [28], who performed ExtVAL on three and one different centers and reached DSCs of 0.59 and 0.66, respectively. The best results were reported by Ma et al. [17], who achieved a mean DSC of 0.84. However, their study used data from a single external center with 88 patients, all acquired with a vendor scanner included in the training set, which may have contributed to their higher results compared to ours. It is worth noting that our results were developed and validated on 14 centers, whereas the cited studies included a maximum of 4. To summarize, we are consistent with the recent literature, although the primary objective of this study was not solely focused on the development of automatic RC segmentation.

Our study has some limitations. First, we considered only one architecture, the U-Net, since it has been recognized as the most effective method for biomedical image segmentation [34]; indeed, the most recent papers still chose to use this backbone network [16–19]. However, it could be of interest to compare other more innovative architectures and assess their impact combined with all the other technical parameters. Second, we compared the random sampling approach with one clustering method, obtained by evaluating first-order radiomics variables, and it would be valuable to investigate other clustering techniques using alternative radiomics features. Furthermore, assessing the impact of various transfer learning techniques, which have shown growing success across multiple fields [35, 36], and considering other database divisions into CONSTR, IntVAL, and ExtVAL could offer additional insights and improvements. Future work could be focused also on additional tasks, e.g., characterization and exploring different approaches. Despite these limits, our paper presents some interesting insights from both technical and clinical points of view: on one side, the crucial importance of defining the most suitable DL parameters combination to solve the task; on the other, the importance of normalization and preprocessing in realigning the distributions of images in a multicenter setting.

In conclusion, the study provides a methodological approach for helping in developing robust and reproducible AI-based segmentation tools in medical imaging. Indeed, these guidelines may not only support and guide informed decision-making but also represent a step toward the standardization of the design of such systems for their introduction into clinical practice.

Author Contributions

J.P. and A.D. investigation, data curation, formal analysis, writing, original draft preparation. L.V. data curation, segmentation. S.C., M.G., R.S., M.A., A.V., L.S., L.B., H.E.T., G.C., G.R.D., E.M., R.F., A.P., V.G., M.M., C.F., B.A.J.-F., M.A.G., S.D.A., A.E. clinical project administration and data selection. D.R., S.R., G.B. coordination, supervision. V.G. coordination, supervision, and writing. The authors read and approved the final manuscript.

Acknowledgments

This research was funded by the Italian Ministry of Health through the network project RCR-2019-23669120_001 of the "Alleanza Contro il Cancro (ACC)" network. The authors acknowledge the support from the Radiomics Group of "Alleanza Contro il Cancro" and the Italian Ministry of Health. The research leading to these results has received funding also from AIRC under5 per Mille 2018 -ID. 21091 program—P.I. Bardelli Alberto, G.L. Regge Daniele.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. F. Knuth, I. A. Adde, B. N. Huynh, et al., "MRI-Based Automatic Segmentation of Rectal Cancer Using 2D U-Net on Two Independent Cohorts," *Acta Oncologica* 61, no. 2 (2022): 255–263, https://doi.org/10.1080/0284186X.2021.2013530.

2. A. Defeudis, S. Mazzetti, J. Panic, et al., "MRI-Based Radiomics to Predict Response in Locally Advanced Rectal Cancer: Comparison of Manual and Automatic Segmentation on External Validation in a Multicentre Study," *European Radiology Experimental* 6, no. 1 (2022): 19, https://doi.org/10.1186/s41747-022-00272-2.

3. M. J. Khan, A. K. Singh, R. Sultana, P. P. Singh, A. Khan, and S. Saxena, "Breast Cancer: A Comparative Review for Breast Cancer Detection Using Machine Learning Techniques," *Cell Biochemistry and Function* 41, no. 8 (2023): 996–1007, https://doi.org/10.1002/cbf.3868.

4. A. Defeudis, J. Panic, G. Nicoletti, S. Mazzetti, V. Giannini, and D. Regge, "Virtual Biopsy in Abdominal Pathology: Where do we Stand?," *BJR Open* 5 (2023): 20220055.

5. J. Santinha, D. Dos Pinto Santos, F. Laqua, et al., "ESR Essentials: Radiomics—Practice Recommendations by the European Society of Medical Imaging Informatics," *European Radiology* 35, no. 3 (2025): 1122–1132, https://doi.org/10.1007/s00330-024-11093-9.

6. A. Defeudis, C. de Mattia, F. Rizzetto, et al., "Standardization of CT Radiomics Features for Multi-Center Analysis: Impact of Software Settings and Parameters," *Physics in Medicine and Biology* 65, no. 19 (2020): 195012, https://doi.org/10.1088/1361-6560/ab9f61.

7. P. Lambin, R. T. H. Leijenaar, T. M. Deist, et al., "Radiomics: The Bridge Between Medical Imaging and Personalized Medicine," *Nature Reviews. Clinical Oncology* 14, no. 12 (2017): 749–762, https://doi.org/10.1038/nrclinonc.2017.141.

8. R. Da-ano, I. Masson, F. Lucia, et al., "Performance Comparison of Modified ComBat for Harmonization of Radiomic Features for Multicenter Studies," *Scientific Reports* 10, no. 1 (2020): 1–12, https://doi.org/10.1038/s41598-020-66110-w.

9. V. Giannini, J. Panic, D. Regge, G. Balestra, and S. Rosati, "Could Normalization Improve Robustness of Abdominal MRI Radiomic Features?," *Biomedical Physics & Engineering Express* 9, no. 5 (2023): 055002, https://doi.org/10.1088/2057-1976/ace4ce.

10. J. Panic, A. Defeudis, G. Balestra, V. Giannini, and S. Rosati, "Normalization Strategies in Multi-Center Radiomics Abdominal MRI: Systematic Review and Meta-Analyses," *IEEE Open Journal of Engineering in Medicine and Biology* 4 (2023): 67–76, https://doi.org/10.1109/OJEMB.2023.3271455.

11. R. Da-Ano, D. Visvikis, and M. Hatt, "Harmonization Strategies for Multicenter Radiomics Investigations," *Physics in Medicine and Biology* 65, no. 24 (2020): 24TR02, https://doi.org/10.1088/1361-6560/aba798.

12. O. Ronneberger, P. Fischer, and T. Brox, "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*, eds. N. Navab, J. Hornegger, and F. Lang, vol. 18 (Springer International Publishing, 2015), 234–241.

13. Y.-J. Huang, Q. Dou, Z.-X. Wang, et al., *3D RoI-Aware U-Net for Accurate and Efficient Colorectal Tumor Segmentation* (arXiv, 2018) 1806.10342v3.

14. D. Barra, G. Nicoletti, A. Defeudis, et al., "Deep Learning Model for Automatic Prostate Segmentation on Bicentric T2w Images With and Without Endorectal Coil," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021* (IEEE, 2021), 3370–3373, https://doi.org/10.1109/EMBC46164.2021.9630792.

15. J. Panic, A. Defeudis, S. Mazzetti, et al., "A Fully Automatic Deep Learning Algorithm to Segment Rectal Cancer on MR Images: A Multi-Center Study," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (Institute of Electrical and Electronics Engineers Inc., 2022), 5066–5069, https://doi.org/10.1109/EMBC48229.2022.9871326.

16. C. Tian, X. Ma, H. Lu, et al., "Deep Learning Models for Preoperative T-Stage Assessment in Rectal Cancer Using MRI: Exploring the Impact of Rectal Filling," *Frontiers in Medicine* 10 (2023): 1326324, https://doi.org/10.3389/fmed.2023.1326324.

17. S. Ma, H. Lu, G. Jing, et al., "Deep Learning-Based Clinical-Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Patients With Rectal Cancer: A Two-Center Study," *Frontiers in Medicine* 10 (2023): 1276672, https://doi.org/10.3389/fmed.2023.1276672.

18. L. Li, B. Xu, Z. Zhuang, et al., "Accurate Tumor Segmentation and Treatment Outcome Prediction With DeepTOP," *Radiotherapy and Oncology* 183 (2023): 109550, https://doi.org/10.1016/j.radonc.2023.109550.

19. K. Zhang, X. Yang, Y. Cui, J. Zhao, and D. Li, "Imaging Segmentation Mechanism for Rectal Tumors Using Improved U-Net," *BMC Medical Imaging* 24, no. 1 (2024): 95, https://doi.org/10.1186/s12880-024-01269-6.

20. R. G. H. Beets-Tan, D. M. J. Lambregts, M. Maas, 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," *European Radiology* 28, no. 4 (2018): 1465–1475, https://doi.org/10.1007/s00330-017-5026-2.

21. E. Scalco and G. Rizzo, "Texture Analysis of Medical Images for Radiotherapy Applications," *British Journal of Radiology* 90, no. 1070 (2017): 20160642, https://doi.org/10.1259/bjr.20160642.

22. J. Panic, G. Balestra, A. Defeudis, S. Rosati, D. Regge, and V. Giannini, "Comparison Between Different Approaches for the Creation of the Training Set: How Clustering and Dimensionality Impact the

Performance of a Deep Learning Model," in *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), Dayton, OH, USA* (IEEE, 2023), 393–396, https://doi.org/10.1109/BIBE60311.2023.00070.

23. L. Ven and J. Lederer, *Regularization and Reparameterization Avoid Vanishing Gradients in Sigmoid-Type Networks* (2021), http://arxiv.org/abs/2106.02260.

24. S. Ioffe, and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International Conference on Machine Learning* (PMLR, 2015), 448–456.

25. C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, *Activation Functions: Comparison of Trends in Practice and Research for Deep Learning* (2018), 1–20. arXiv preprint arXiv:1811.03378.

26. S. Trebeschi, J. J. M. van Griethuysen, D. M. J. Lambregts, et al., "Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR," *Scientific Reports* 7, no. 1 (2017): 1, 5301–9, https://doi.org/10.1038/s41598-017-05728-9.

27. M. Wang, P. Xie, Z. Ran, et al., "Full Convolutional Network Based Multiple Side-Output Fusion Architecture for the Segmentation of Rectal Tumors in Magnetic Resonance Images: A Multi-Vendor Study," *Medical Physics* 46, no. 6 (2019): 2659–2668, https://doi.org/10.1002/mp.13541.

28. X. Pang, F. Wang, Q. Zhang, et al., "A Pipeline for Predicting the Treatment Response of Neoadjuvant Chemoradiotherapy for Locally Advanced Rectal Cancer Using Single MRI Modality: Combining Deep Segmentation Network and Radiomics Analysis Based on "Suspicious Region"," *Frontiers in Oncology* 11 (2021): 711747, https://doi.org/10.3389/fonc.2021.711747.

29. M. T. Nyo, F. Mebarek-Oudina, S. S. Hlaing, and N. A. Khan, "Otsu's Thresholding Technique for MRI Image Brain Tumor Segmentation," *Multimedia Tools and Applications* 81, no. 30 (2022): 43837–43849, https://doi.org/10.1007/s11042-022-13215-1.

30. A. Saha, J. S. Bosma, J. J. Twilt, et al., "Artificial Intelligence and Radiologists in Prostate Cancer Detection on MRI (PI-CAI): An International, Paired, Noninferiority, Confirmatory Study," *Lancet Oncology* 25, no. 7 (2024): 879–887, https://doi.org/10.1016/S1470-2045(24)00220-1.

31. P. A. Narayana, I. Coronado, S. J. Sujit, J. S. Wolinsky, F. D. Lublin, and R. E. Gabr, "Deep-Learning-Based Neural Tissue Segmentation of MRI in Multiple Sclerosis: Effect of Training Set Size," *Journal of Magnetic Resonance Imaging* 51, no. 5 (2020): 1487–1496, https://doi.org/10.1002/jmri.26959.

32. N. Wulms, L. Redmann, C. Herpertz, et al., "The Effect of Training Sample Size on the Prediction of White Matter Hyperintensity Volume in a Healthy Population Using BIANCA," *Frontiers in Aging Neuroscience* 13 (2021): 720636, https://doi.org/10.3389/fnagi.2021.720636.

33. Y. Fang, J. Wang, X. Ou, et al., "The Impact of Training Sample Size on Deep Learning-Based Organ Auto-Segmentation for Head-and-Neck Patients," *Physics in Medicine and Biology* 66, no. 18 (2021): 185012, https://doi.org/10.1088/1361-6560/ac2206.

34. S. Pang, A. Du, M. A. Orgun, Y. Wang, and Z. Yu, "Tumor Attention Networks: Better Feature Selection, Better Tumor Segmentation," *Neural Networks* 140 (2021): 203–222, https://doi.org/10.1016/j.neunet.2021.03.006.

35. S. Tammina, "Transfer Learning Using VGG-16 With Deep Convolutional Neural Network for Classifying Images," *International Journal of Scientific and Research Publications* 9, no. 10 (2019): p9420, https://doi.org/10.29322/ijsrp.9.10.2019.p9420.

36. R. Da-Ano, F. Lucia, I. Masson, et al., "A Transfer Learning Approach to Facilitate ComBat-Based Harmonization of Multicentre Radiomic Features in New Datasets," *PLoS One* 16, no. 7 (2021): 1–19, https://doi.org/10.1371/journal.pone.0253653.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.