Physics and Imaging in Radiation Oncology 30 (2024) 100578

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Automatic gross tumor volume segmentation with failure detection for safe implementation in locally advanced cervical cancer

Rahimeh Rouhi [a,b], Stéphane Niyoteka [a,b], Alexandre Carré [a,b], Samir Achkar [b], Pierre-Antoine Laurent [a,b], Mouhamadou Bachir Ba [b,c], Cristina Veres [a,b], Théophraste Henry [a,d], Maria Vakalopoulou [e], Roger Sun [a,b], Sophie Espenel [b], Linda Mrissa [b], Adrien Laville [a], Cyrus Chargari [a,b], Eric Deutsch [a,b], Charlotte Robert [a,b,*]

[a] Université Paris-Saclay, Institut Gustave Roussy, Inserm, Radiothérapie Moléculaire et Innovation Thérapeutique, 94800 Villejuif, France
[b] Department of Radiation Oncology, Gustave Roussy Cancer Campus, Villejuif, France
[c] Radiotherapy Department of the University Hospital Center of Dalal Jamm, Guédiawaye, Senegal
[d] Department of Medical Imaging, Gustave Roussy Cancer Campus, Villejuif, France
[e] Laboratoire Mathématiques et Informatique pour la Complexité et les Systèmes, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

ARTICLE INFO

Keywords:
Locally advanced cervical cancer
Adaptive radiotherapy
Deep learning
Automatic segmentation
Failure detection
Quality assurance

ABSTRACT

Background and Purpose: Automatic segmentation methods have greatly changed the RadioTherapy (RT) workflow, but still need to be extended to target volumes. In this paper, Deep Learning (DL) models were compared for Gross Tumor Volume (GTV) segmentation in locally advanced cervical cancer, and a novel investigation into failure detection was introduced by utilizing radiomic features.

Methods and materials: We trained eight DL models (UNet, VNet, SegResNet, SegResNetVAE) for 2D and 3D segmentation. Ensembling individually trained models during cross-validation generated the final segmentation. To detect failures, binary classifiers were trained using radiomic features extracted from segmented GTVs as inputs, aiming to classify contours based on whether their Dice Similarity Coefficient (DSC) $< T$ and DSC$\geqslant T$. Two distinct cohorts of T2-Weighted (T2W) pre-RT MR images captured in 2D sequences were used: one retrospective cohort consisting of 115 LACC patients from 30 scanners, and the other prospective cohort, comprising 51 patients from 7 scanners, used for testing.

Results: Segmentation by 2D-SegResNet achieved the best DSC, Surface DSC ($SDSC_{3mm}$), and 95th Hausdorff Distance (95HD): DSC $= 0.72 \pm 0.16$, $SDSC_{3mm}=0.66 \pm 0.17$, and 95HD $= 14.6 \pm 9.0$ mm without missing segmentation (M=0) on the test cohort. Failure detection could generate precision ($\mathscr{P} = 0.88$), recall ($\mathscr{R} = 0.75$), F1-score ($\mathscr{F} = 0.81$), and accuracy ($\mathscr{A} = 0.86$) using Logistic Regression (LR) classifier on the test cohort with a threshold T = 0.67 on DSC values.

Conclusions: Our study revealed that segmentation accuracy varies slightly among different DL methods, with 2D networks outperforming 3D networks in 2D MRI sequences. Doctors found the time-saving aspect advantageous. The proposed failure detection could guide doctors in sensitive cases.

1. Introduction

Locally Advanced Cervical Cancer (LACC) stands as the fourth most prevalent cancer among women worldwide [1]. Magnetic Resonance Imaging (MRI) is indispensable for cancer lesion management, providing high soft tissue contrast [2]. Automatic segmentation offers precise tumor delineation, easing the manual workload for radiation oncologists [3] and reducing inter-expert variability [4]. Despite its significance, automatic tumor segmentation has seen limited application in the female pelvic region, specifically in Gross Tumor Volume (GTV) segmentation. Several challenges must be addressed in this context. LACC tumor boundaries in MR images may appear blurred due to low tissue contrast. Moreover, the presence of secretions from the uterine cavity, exhibiting a similar intermediate or high-intensity signal as the tumor, can pose challenges in defining the upper limit of the GTV in T2-Weighted (T2W) MR images. Lastly, the scarcity of public datasets may

* Corresponding author at: Department of Radiotherapy, Gustave Roussy Cancer Campus, 114 Rue Edouard Vaillant, 94805 Villejuif Cedex, France.
E-mail address: ch.robert@gustaveroussy.fr (C. Robert).

https://doi.org/10.1016/j.phro.2024.100578
Received 3 October 2023; Received in revised form 8 April 2024; Accepted 8 April 2024
Available online 13 April 2024

explain the infrequent application of automatic Deep Learning (DL)-based segmentation to cervical cancer.

Several studies have developed DL-based models for automatic segmentation in cervical cancers, mainly focusing on Clinical Tumor Volume (CTV) segmentation on Computed Tomography (CT) [3,5–7], and MRI [8]. Regarding GTV, Breto et. al. [9], used a region-based convolutional neural network (R-CNN) for the segmentation of GTV + Cervix in 646 onboard MR images. The Dice Similarity Coefficient (DSC) of 0.84 was reported for GTV + Cervix in their validation. No cohort was used for testing their model. In another work, Breto et. al. [10] used planning and daily treatment fraction setup (RT-Fr) MR images, from 15 LACC patients for GTV + Cervix segmentation. A MASK R-CNN network [11] was trained and tested in three different scenarios. The first scenario involved using planning images of $N-1$ patients for training (Leave-one-out or LOO), the second one tested the network on the RT-Fr MRIs of the left-out patient, and the third scenario involved including the planning MRI of the left-out patient as an additional training sample and testing on RT-Fr MRIs. The best results for GTV + Cervix segmentation concluded from the first scenario, with DSC = 0.67 ± 0.30 and 95th Hausdorff Distance (95HD)=2.7 ± 1.7 mm. Yoganathan et. al. [12], trained 2D (axial only) and 2.5D (axial, sagittal and coronal) models based on ResNet50 [13] and InceptionResNetv2 (InRN) [14] architectures. A total number of 71 T2W MR images from 39 patients for cervix image-based High Dose Rate (HDR) brachytherapy were used. The best results were obtained by 2.5D InRN with an average of DSC = 0.62 ± 0.14 and 95HD = 6.8 ± 2.8 mm for GTV segmentation.

Even top DL models for segmentation can struggle with challenges such as domain shift, noise, and low image quality, leading oncologists to spend significant time adjusting results. This hampers the practicality of automatic segmentation for real-world use. Clinical implementation of automatic segmentation requires Quality Assurance (QA) tools to maintain model consistency over time (routine QA) and accuracy for each patient (case-specific QA) [15–19]. QA approaches in medical image segmentation can be categorized into two groups. The first group involves estimating and visualizing uncertainties [20–26]. The second group focuses on predicting segmentation quality [16,26–35,28,36,37].

This study aimed to provide significant contributions, including a comprehensive comparison of various deep neural network architectures not previously utilized in GTV segmentation for LACC. Additionally, we evaluated the generalizability of these models on T2W MR images obtained prospectively from diverse sources, encompassing patients with or without vagina/rectum opacification across different centers, specifically in the context of external beam RadioTherapy (RT) applications within a clinical trial. Moreover, we aimed to expedite the integration of segmentation results into clinical practice by identifying and analyzing segmentation failures using radiomics features.

2. Materials and methods

This section covers data, preprocessing steps, and methods for segmentation and failure detection. For more details on data collection, DL architectures, and implementation, see Sections S1, S2, and S3 in the supplementary data.

2.1. Patient cohorts

We collected two cohorts of T2W MR pre-RT images, taken in 2D sequences by 30 and 7 different scanners, respectively, from 115 and 51 patients treated for LACC at different centers. The first cohort (cohort 1 - retrospective) was used for model training, and the second cohort (cohort 2 - prospective - ATEZOLACC clinical trial (NCT03612791)) was used for testing in both segmentation and classification tasks. Tables S3 and S4 in the supplementary data summarize the characteristics of the MRI devices and acquisition parameters for both cohorts. Approximately 33% of the patients (38 out of 115) in the training cohort and 64% of the patients (33 out of 51) in the test cohort underwent rectum/vagina opacification for better tumor and organ at risk visibility.

2.2. DL-based segmentation

We trained 4 deep neural network architectures namely UNet [38,39], VNet [40], SegResNet and SegResNetVAE [41], for both 2D and 3D segmentation, resulting in 8 different models. More information on the applied architectures was provided in Section S2 of the supplementary data. We used 5-fold cross-validation on cohort 1 images for training and validation of all networks. Segmentation results were reported using two strategies: single model and ensemble model. In the single model approach, the best-performing model with the highest average DSC in cross-validation was applied to cohort 2. In the ensemble approach, all five models from cross-validation were used on cohort 2. The ensemble prediction was obtained by averaging the results of all the models [42]. The ensembling strategy decreases the model variance and subsequently improves the segmentation performance [43]. For both 2D and 3D segmentation, we treated the final segmentation as a 3D volume and calculated DSC, Surface DSC ($SDSC_{3mm}$), and 95HD, defined in Section S4 of the supplementary data. Additionally, we presented the segmentation results separately for opacified and non-opacified cases.

2.3. Radiomics-based failure detection

The best ensemble model from 5-fold cross-validation segmented all images in both the training set (cohort 1) and the testing set (cohort 2). Using PyRadiomics (v3.0.1), an open-source Python package [44], 93 radiomic features were extracted from these segmented GTVs, employing a relative discretization method with 32 bins.

We utilized different machine learning models such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbour (KNN), Decision Tree Classifier (DTC), Gaussian Naive Bayes (GNB), and Support Vector Classifier (SVC). We used the open-source library of Scikit-learn [45] in the classification. The classifiers were trained with the goal of classifying images for which segmentation resulted in a $DSC/SDSC_{3mm} < T$ or $DSC/SDSC_{3mm} \geqslant T$, where T is a threshold on the selected metric $DSC/SDSC_{3mm}$. Nested Cross-Validation (NCV) [46] was used to select the best classifier and its hyperparameters. This strategy enhances ML model performance by nested loops: one for model selection, hyperparameter tuning, and feature selection, and another for validating on unseen data. We used 5 outer and 3 inner folds. The final model, from NCV, was applied to all training samples. Recursive feature elimination (Guyon et al. [47]) was used individually on outer folds for feature selection. To address class imbalance, we applied SMOTE (Chawla et al. [48]) on the training cohort. The model was then tested on segmentation failure for detection, evaluated with precision ($\mathscr{P}$), recall ($\mathscr{R}$), F-score ($\mathscr{F}$), accuracy ($\mathscr{A}$), and AUC (Sokolova et al. [49]).

To assess the predictive power of radiomic features in detecting failures, we analyzed how the choice of binarization metric and threshold impacts performances. We considered different scenarios: automatically segmented GTVs with or without post-processing (one step after final segmentation, including morphology operations). This resulted in 4 cases: DSC_W, DSC_WO, $SDSC_{3mm}$_W, and $SDSC_{3mm}$_WO. Additionally, to select the optimal threshold T, we used a range $[b_l, b_h]$, statistically determined as follows:

$$[b_l, b_h] = \bigcap_{i=1}^{n} [\overline{x_i} - s_i, \overline{x_i} + s_i] \quad (1)$$

where b_l and b_h are respectively the lowest and highest values of $DSC/SDSC_{3mm}$, $\overline{x_i}$ and s_i are respectively the mean and standard deviation of $DSC/SDSC_{3mm}$ values in i^{th} case out of the $n = 4$ aforementioned cases. Then, we performed the failure detection $b_h - b_l + 1$ times, based on the best classifiers resulting from the aforementioned setting, by selecting each time a $T \in [b_l, b_h]$. In doing so, the T, which resulted in the best

accuracy in the validation, was considered for the final failure detection model.

2.4. Statistical and qualitative analysis

In the cross-validation for both segmentation and classification, we compared the performance of DL and ML models using the metrics mentioned earlier to identify the best-performing model. We applied the Friedman method [50] to analyze statistically significant differences among them. Additionally, segmentation and failure detection were clinically evaluated. The radiation oncologist (P-A.L), with 3 years of experience, scored the segmentation output into three classes: A (good with no/minor correction), B (satisfactory with major correction), and C (bad with complete correction or delineation from scratch). The time for contour correction, if needed, was measured by the same oncologist. For comparison, we considered the approximate time needed for segmentation from scratch per subject, provided by another radiation oncologist (L.M, with 5 years of expertise), who segmented 10 randomly selected images from the test cohort. The average time for these 10 cases was used for manual segmentation from scratch per subject.

3. Results

3.1. Automatic GTV segmentation

Table S1 presents the average validation results obtained from the 5 trained models in the 5-fold cross-validation for each network in 2D and 3D segmentation separately on cohort 1. The 2D-SegResNetVAE resulted in the best values of DSC $= 0.61 \pm 0.03$, $SDSC_{3mm} = 0.57 \pm 0.04$, 95HD $= 16.1 \pm 2.6$ mm, and M=0, respectively compared with the other models. The results are statistically different with $p-value < 0.05$. Moreover, Table S2 shows the results of segmentation on cohort 2 (test cohort) based on single models, i.e. the best models obtained from 5-fold cross-validation. The results of SegResNet and SegResNetVAE in the 2D case were the best when considering the output of the single models with/without failure i.e., M=1/0. Furthermore, Table 1 presents the results of ensembling all the predictions from the 5 models on the test cohort. Accordingly, 2D-SegResNet achieved the best performance compared to all the other models. Additionally, we calculated different metrics, including DSC, $SDSC_{3mm}$, and 95HD, for the corrected auto-segmentation volumes compared to their ground truths in cohort 2. We obtained DSC $= 0.73 \pm 0.26$, $SDSC_{3mm}=0.71 \pm 0.28$, and 95HD $= 13.7 \pm 14.4$. In comparison to the segmentation without correction (DSC $= 0.72 \pm 0.16$, $SDSC_{3mm}=0.66 \pm 0.17$, and 95HD $= 14.6 \pm 9.0$ mm), we observed fairly comparable values, although higher for $SDSC_{3mm}$, illustrating performance for the network of the same order of magnitude as the inter-expert variability.

Fig. 1 in sub-figures (a), (b), and (c) displays box-plots of ensemble model segmentation results on cohort 2, showcasing DSC, $SDSC_{3mm}$, and 95HD. Additionally, Figure S1 illustrates patient-wise segmentation outcomes using 2D-SegResNet for the test cohort, presenting both DSC and 95HD. The model achieved DSC$\geqslant$0.70 in 67% of total images (34 out of 51 patients), with corresponding values of $SDSC_{3mm}$ and 95HD in sub-figures (b) and (c).

Fig. 2 shows visual segmentation results for two test cohort patients using 2D-SegResNet. The model achieved the best DSC $=$ 0.89,$SDSC_{3mm}$=0.91, and 95HD $= 4.0$ mm, and the worst DSC $= 0.18$, $SDSC_{3mm} = 0.19$, and 95HD $= 40.9$ mm. Additionally, we obtained DSC $= 0.70 \pm 0.17$, $SDSC_{3mm} = 0.66 \pm 0.18$ and 95HD $= 15.3 \pm 9.9$ mm for opacified cases, and DSC $= 0.74 \pm 0.10$, $SDSC_{3mm} = 0.66 \pm 0.12$, and 95HD $= 13.3 \pm 6.7$ mm for non-opacified cases in the test cohort.

Additionally, based on qualitative results, specifically, 62% (32/51), 27% (14/51), and 5% (3/51) of the images in the test set (cohort 2) were scored as A, B, and C, respectively. The radiation oncologist declined to score 3.9% (2/51) of the images due to their low quality.

3.2. Failure detection

Fig. 3 illustrates the cross-validation results for failure detection. Subfigure (a) depicts average DSC and $SDSC_{3mm}$ values resulting from segmentation with and without post-processing. Notably, DSC_WO achieved superior results compared to other cases, with all results showing statistical significance (p-value $= 0.018$). With DSC_WO generated the best results, we proceeded to classification using different models, as presented in subfigure (b). The results also exhibited statistical significance (p-value $= 0.010$). Figure S2 presents the average accuracy of the LDA classifier as an example in NCV with different thresholds in the range [0.60, 0.80] on the values of DSC_WO, resulting from the 2D SegResNet were considered. As observed, the LDA achieved the highest mean accuracy values, specifically up to the threshold $T =$ 0.65 on the values of DSC_WO.

Table 2 shows classifiers' results in failure detection on the test cohort. LR achieved $\mathscr{P} = 0.88$, $\mathscr{R} = 0.75$, $\mathscr{F} = 0.81$, and $\mathscr{A} = 0.86$ with $T = 0.67$ on DSC_WO for cohort 2. Figure S3 illustrates LR classifier performance with prediction probability. Additionally, Table S5 in the supplementary data lists features selected for LR classification, with the top three being original_firstorder_InterquartileRange, original_firstorder_Variance, and original_glcm_ClusterTendency.

We presented the confusion matrices for binary classification in failure detection in Table 3. These matrices were derived from both qualitative segmentation scoring (A, B, C) and quantitative assessment using DSC_WO values. Specifically, 65.6% (21/32) of cases were correctly classified as A (no/minor correction), and 75.0% (3/4) as C (complete correction), including samples refused during segmentation scoring by the radiation oncologist. The average correction time for test set cases (cohort 2) was 4.26 ± 3.62 min. Breakdown by category revealed: 2.03 ± 1.33 min for A, 7.28 ± 1.43 min for B, and 14.00 ± 2.00 for C, assessed by the same radiation oncologist. Manual segmentation time, reported by another radiation oncologist (L.M), averaged approximately 16.0 ± 3.8 min.

4. Discussion

In this study, 8 DL models were trained for both 2D and 3D segmentation to compare GTV segmentation in LACC using T2W MRI from clinical routine. Results indicated superior performance of 2D SegResNet and SegResNetVAE models, likely due to their robust design with residual connections, asymmetrically larger encoder for feature extraction, and smaller decoder with a variational auto-encoder (VAE) for segmentation mask reconstruction [41]. The differences in data

Table 1
Average results of DSC, $SDSC_{3mm}$, and 95HD(mm) and the total number of segmentation failures M on the testing set (cohort 2), based on ensemble models resulted by model averaging in 5-fold cross validation obtained from different network architectures in 2D and 3D segmentation.

	2D				3D			
Network	DSC ± SD	$SDSC_{3mm}$ ± SD	95HD ± SD	M	DSC ± SD	$SDSC_{3mm}$ ± SD	95HD ± SD	M
UNet	0.69 ± 0.23	0.65 ± 0.23	18.1 ± 24.1	**0**	0.54 ± 0.28	0.44 ± 0.25	32.0 ± 36.1	**0**
VNet	0.71 ± 0.20	0.66 ± 0.19	14.6± 11.7	1	0.50 ± 0.22	0.32 ± 0.19	30.9 ± 24.0	1
SegResNet	**0.72 ± 0.16**	**0.66 ± 0.17**	**14.6 ± 9.0**	**0**	0.57 ± 0.26	0.45 ± 0.23	37.4 ± 53.3	**0**
SegResNetVAE	0.70 ± 0.18	0.63 ± 0.18	15.0 ± 10.8	**0**	0.63 ± 0.21	0.52 ± 0.19	20.5 ± 16.4	**0**

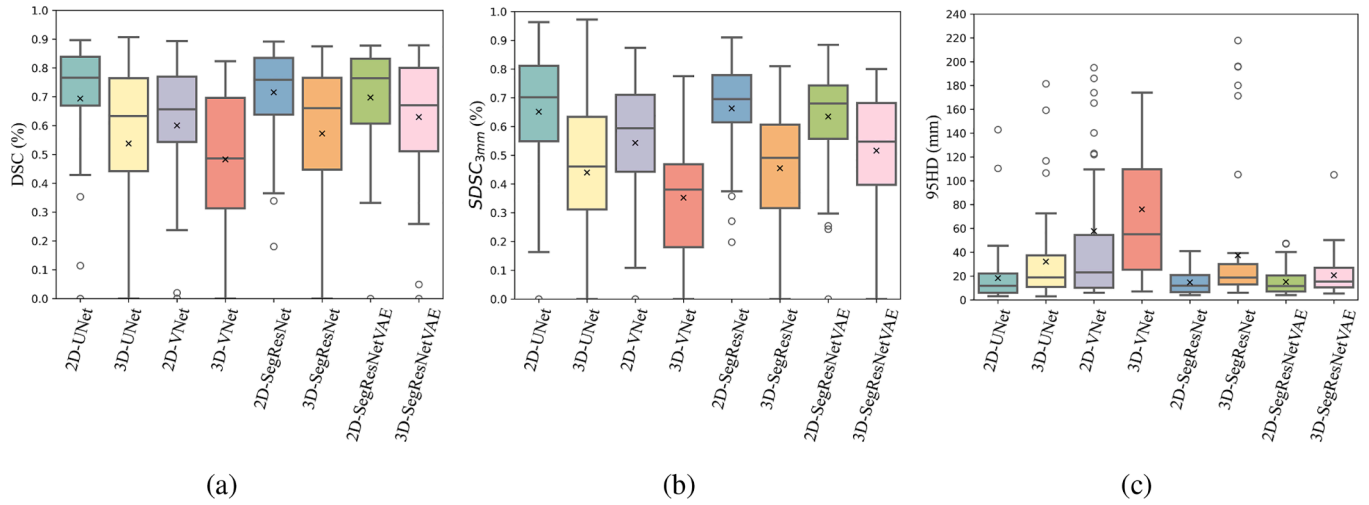


Fig. 1. Box-plots results of the different network architectures on the testing set (cohort 2) for ensemble 2D and 3D segmentation in terms of DSC, $SDSC_{3mm}$, and 95HD metrics, respectively shown in sub-figures (a), (b), and (c).

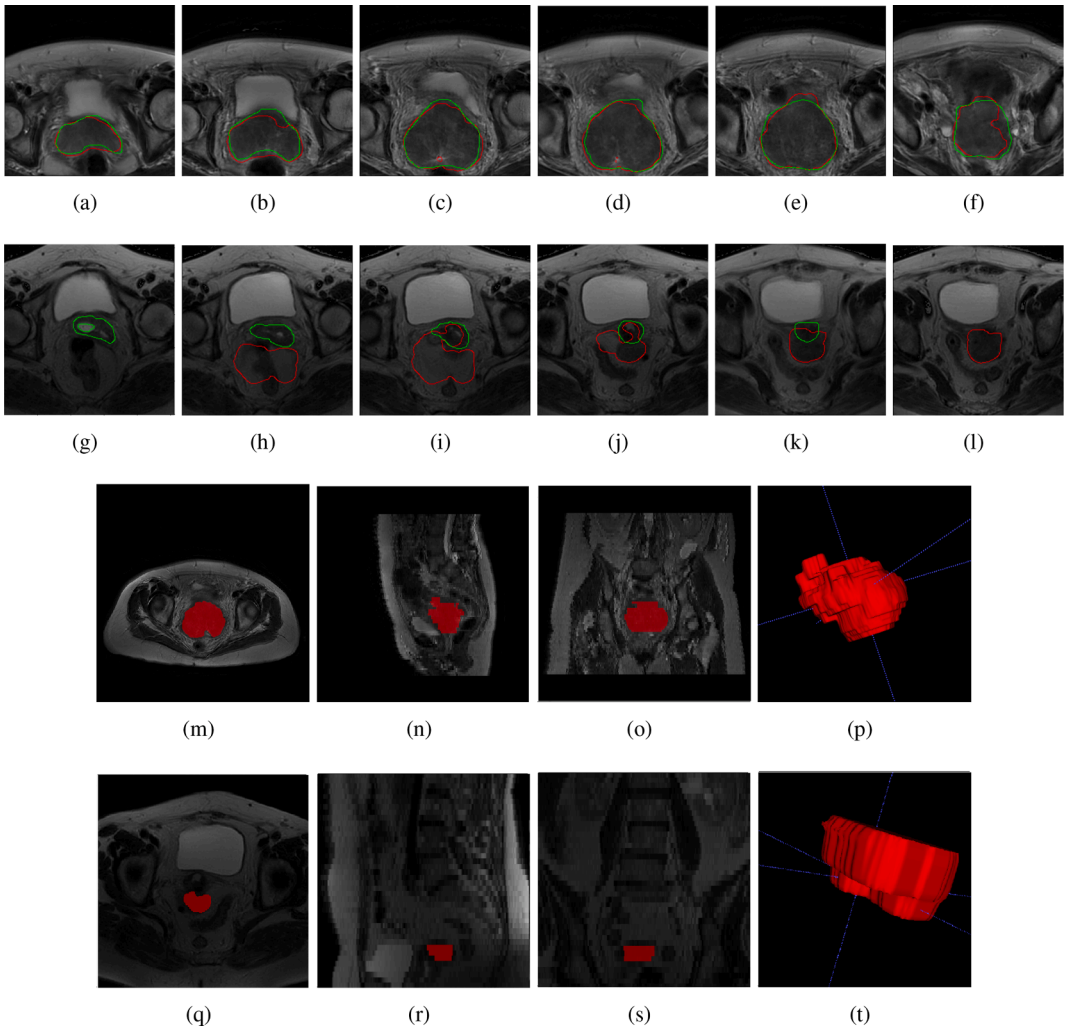
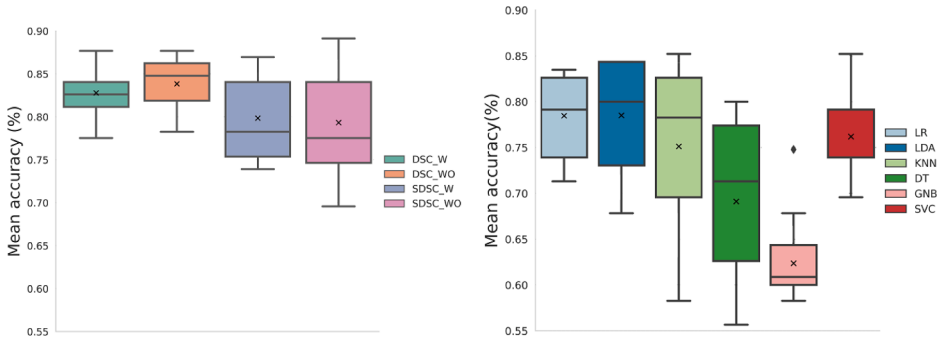


Fig. 2. The best (a)-(f) and the worst (g)-(l) performing samples in the test cohort for GTV segmentation, by 2D-SegResNet as our proposed ensemble model, shown in the first and the second rows, respectively, with DSC = 0.89, $SDSC_{3mm}$=0.91 and 95HD = 4.0 mm, and DSC = 0.18, $SDSC_{3mm}$=0.19 and 95HD = 40.9 mm. The images (m)-(p) and (q)-(f) correspond to the images in the first and second rows with ground truth visualization in axial, sagittal, coronal, and 3D visualization. Green and red colors correspond to ground truth and segmentation output.



(a) (b)

Fig. 3. (a) Box-plot results of average accuracy on validation folds resulted from all the applied models by considering different thresholds of DSC and $SDSC_{3mm}$ ranging in [0.60,0.80] and application or not of the post-processing step and (b) box-plots results of average accuracy resulted from each ML model by considering different thresholds of DSC ranging in [0.60,0.80]. The results are statistically different with $p-value < 0.05$.

Table 2
Results of failure detection on test cohort, in terms of precision ($\mathscr{P}$), recall ($\mathscr{R}$), F1-score ($\mathscr{F}$), accuracy ($\mathscr{A}$), and AUC based on different classifiers.

Classifier	T	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{A}$	AUC
LR	0.67	**0.88**	**0.75**	**0.81**	**0.86**	**0.91**
LDA	0.65	0.70	0.73	0.71	0.78	0.90
KNN	0.67	0.64	0.44	0.52	0.75	0.76
DT	0.64	0.57	0.22	0.32	0.66	0.63
GNB	0.62	0.62	0.50	0.55	0.75	0.74
SVC	0.62	0.72	0.50	0.60	0.78	0.80

acquisition between the retrospective and prospective cohorts for training and testing respectively, introduce complexity to the segmentation process. To visualize these disparities, we extracted quality metrics from images in both cohorts, following the methodology of Sadri et al. [51]. The probability distribution of these metrics, depicted in Figure S4, offers a visual representation of the variations and patterns across the two cohorts. Image characteristics play a crucial role in the quality of tumor volume segmentation, with higher-quality images yielding better segmentation outputs [52]. The worst results of 2D-SegResNet, shown in Figs. 2 (g)-(l), correspond to the first outlier in Fig. 1 (a), due to poor tumor-to-non-tumor contrast. Our model is effective on both opacified and non-opacified MR images, an aspect overlooked in prior LACC GTV segmentation studies.

In clinical practice, GTV delineation in LACC typically relies on transverse views, with corrections made on coronal or sagittal planes, taking adjacent slices into account. Thus, 3D segmentation is favored for preserving spatial context. However, our dataset comprises 2D imaging sequences, explaining why 2D segmentation outperformed 3D in our study. As shown in Fig. 2, subfigures (n), (o), (p), (r), (s), and (t), axial image acquisition can cause misalignment between slices in sagittal and coronal directions. Moreover, our dataset had varying slice numbers per patient (12 to 90), with ground truth labels interpolated rather than manually created for each slice, potentially compromising segmentation accuracy.

Rodríguez et al. [53] utilized a 3D nnU-Net architecture to automate GTV delineation from axial T2w MR images in cervical cancers for brachytherapy. This single-center study involved 195 patients treated from August 2012 to December 2021 with varied 1.5 T and 3T MRI settings. Evaluation metrics (DSC, 95HD, MSD) were computed on 39 test patients, yielding a median DSC of 0.73 (IQR = 0.50–0.80), median 95HD of 6.8 mm (IQR = 4.2–12.5 mm), and median MSD of 1.4 mm (IQR = 0.9–2.8 mm). Notably, significant DSC differences were seen with stratification by GTV volumes; lower volumes (0.26–2.83 cc) showed the lowest DSC. In our study, 2D SegResNet achieved the best median DSC = 0.76 (IQR = 0.64–0.73), median $SDSC_{3mm}$=0.70 (IQR = 0.61–0.78), and median 95HD = 12.0 mm (IQR = 6.5–20.8 mm) compared to other models on the test cohort (cohort 2). Given our model's multi-center training, these results are promising.

Implemented as a binary classification task, our proposed failure detection approach links radiomic features extracted from automatically segmented tumors to the success or failure of automatic segmentation. During cross-validation, both LR and LDA exhibited nearly identical mean accuracy values of 0.78, as depicted in Fig. 3 (b). However, LDA had a higher median than LR (0.80 versus 0.79), while LR showed lower variance (0.003 versus 0.005). Notably, LR achieved a higher average accuracy of $\mathscr{A} = 0.86$ on the test cohort (cohort 2) compared to LDA, potentially due to its lower variance. The negative class, encompassing samples with DSC≥0.67, demonstrated a higher accuracy of $\mathscr{A} = 0.88$, compared to the positive class with DSC < 0.67, where $\mathscr{A} = 0.75$. However, examining the classification confusion results in Table 3 reveals that the classification is not perfect; out of 32 cases assessed as A by doctors, 21 (65.6%) were classified as 0, indicating no segmentation failure by the classifier. Regarding the green values in Table 3, 3 out of 4 cases (75.0%) categorized as C were correctly identified as 1. Notably, training in binary classification utilized the DSC metric rather than doctors' scores to prevent bias caused by subjective scoring. Additionally, an analysis explored whether radiomic assessment indirectly reflected image quality indices associated with the machine. Results showed that 17 out of 51 cases categorized as 1 (poorly delineated cases) originated from different machines/centers. However, 3 out of 4 cases scored as C originated from the same center, indicating a potential connection.

Future research should address the limitations of this study. In clinical practice, physicians often incorporate additional MR sequences

Table 3
Confusion matrix for failure detection. Cases A are shown in red, cases B in blue, and cases C in green.

	Predicted label=0	Predicted label=1
True label = 0	21/7/1	2/0/0
True label= 1	2/2/0	7/5/3

like functional Diffusion-weighted imaging (DWI) and clinical data for challenging tumor delineation cases [54]. Therefore, combining different MR sequences for automatic segmentation could enhance accuracy, especially for poor-quality MR images. While this study provided a quantitative evaluation of segmentation, it lacked access to MR-LINAC. With the MR-linac's development, efficient automatic segmentation could facilitate swift and easy implementation of adaptive RT during treatment. The evolving landscape of DL networks, such as SwinUNetR, nnUNet, [55–57], for medical image segmentation warrants evaluation. Additionally, exploring diverse ensembling methods, like majority voting on outputs of 2D and 3D segmentation networks, is valuable. Enhancements to our failure detection method for clinical suitability may involve extracting image quality metrics and integrating them with radiomic features from segmented tumors. This approach aims to precisely characterize inter-image differences and enhance failure detection effectiveness [58].

Despite needing improvement, our center implements the model in routine clinical practice. The CTV for LACC in external beam RT encompasses the uterine body, vagina, bilateral parametria, and pelvic nodal regions (common, internal, and external iliacs, obturator, and presacral lymph nodes). Thus, a new model is necessary for this clinical task. Additionally, the model may have clinical relevance in defining the High Risk-CTV for brachytherapy. An ongoing evaluation considers pre-brachytherapy images to assess its generalizability.

5. Compliance with ethical standards

The utilization of the retrospective training cohort was performed under the General Data Protection Regulation (GDPR) and approved by the Institutional Review Board (n° IRB2023-271 Gustave Roussy cancer campus). For the prospective trial data, written informed consent was obtained from all patients prior to the start of the trial.

CRediT authorship contribution statement

Rahimeh Rouhi: Methodology, Conceptualization, Writing-original-draft. **Stéphane Niyoteka:** Methodology, Data-curation, Writing-review-editing. **Alexandre Carré:** Methodology, Writing-review-editing. **Samir Achkar:** Data-curation, Validation. **Pierre-Antoine Laurent:** Data-curation, Validation. **Mouhamadou Bachir Ba:** Data-curation. **Cristina Veres:** Data-curation. **Théophraste Henry:** Methodology. **Maria Vakalopoulou:** Methodology. **Roger Sun:** Validation, Writing-review-editing. **Sophie Espenel:** Resources, Data-curation, Writing-review-editing. **Linda Mrissa:** Validation. **Adrien Laville:** Validation. **Cyrus Chargari:** Funding-acquisition, Resources. **Eric Deutsch:** Supervision, Funding-acquisition, Resources, Writing-review-editing. **Charlotte Robert:** Funding-acquisition, Conceptualization, Methodology, Writing-review-editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gustave Roussy has received financial support from Elekta company for the completion of the study.

Acknowledgement

We would like to acknowledge Elekta company for its financial support. This study was also supported by a grant from the French Ministry of Health and the French National Cancer Institute (MRI-ImmunoRadiomics – PRT-K 2020–040).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.phro.2024.100578.

References

[1] Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Global Health 2020;8(2):e191–203. https://doi.org/10.1016/S2214-109X(19)30482-6.
[2] Ghadimi M, Sapra A. Magnetic resonance imaging contraindications. In: StatPearls. StatPearls Publishing; 2021.
[3] Zhikai L, Guan H. Development And Validation Of A Deep Learning Algorithm For Auto-Delineation Of Clinical Target Volume And Organs At Risk In Cervical Cancer Radiotherapy. Int J Radiation Oncol, Biol, Phys 2020;108(3):e766. https://doi.org/10.1016/j.radonc.2020.09.060.
[4] van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. Radiotherapy Oncol 2019;137:9–15. https://doi.org/10.1016/j.radonc.2019.04.006.
[5] Zhang D, Yang Z, Jiang S, Zhou Z, Meng M, Wang W. Automatic segmentation and applicator reconstruction for CT-based brachytherapy of cervical cancer using 3D convolutional neural networks. J Appl Clinical Med Phys 2020;21(10):158–69. https://doi.org/10.1002/acm2.13024.
[6] Liu Z, Chen W, Guan H, Zhen H, Shen J, Liu X, et al. An Adversarial Deep-Learning-Based Model for Cervical Cancer CTV Segmentation With Multicenter Blinded Randomized Controlled Validation. Front Oncol 2021;3223. https://doi.org/10.3389/fonc.2021.702270.
[7] Yang C, Qin Lh, Ye Xie, Jy Liao. Deep learning in CT image segmentation of cervical cancer: a systematic review and meta-analysis. Rad Oncol 2022;17(1):1–14. https://doi.org/10.1186/s13014-022021486.
[8] Zabihollahy F, Viswanathan AN, Schmidt EJ, Lee J. Fully automated segmentation of clinical target volume in cervical cancer from magnetic resonance imaging with convolutional neural network. J Appl Clinical Medical Phys 2022;23(9):e13725. https://doi.org/10.1002/acm2.13725.
[9] Breto A, Zavala-Romero O, Asher D, Baikovitz J, Ford J, Stoyanova R, et al. A Deep Learning Pipeline for per-Fraction Automatic Segmentation of GTV and OAR in cervical cancer. Int J Radiation Oncol, Biol, Phys 2019;105(1):S202. https://doi.org/10.1016/j.ijrobp.2019.06.267.
[10] Breto AL, Spieler B, Zavala-Romero O, Alhusseini M, Patel NV, Asher DA, et al. Deep Learning for Per-Fraction Automatic Segmentation of Gross Tumor Volume (GTV) and Organs at Risk (OARs) in Adaptive Radiotherapy of Cervical Cancer. Front Oncol 2022;12:854349. https://doi.org/10.3389/fonc.2022.854349.
[11] He K, Gkioxari G. Dollá r P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2961-9. doi: 10.48550/arXiv.1703.06870.
[12] Yoganathan S, Paul SN, Paloor S, Torfeh T, Chandramouli SH, Hammoud R, et al. Automatic segmentation of magnetic resonance images for high-dose-rate cervical cancer brachytherapy using deep learning. Med Phys 2022;49(3):1571–84. https://doi.org/10.1002/mp.15506.
[13] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
[14] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence; 2017. https://doi.org/10.1609/aaai.v31i1.11231.
[15] Patel VM, Gopalan R, Li R, Chellappa R. Visual domain adaptation: A survey of recent advances. IEEE Signal Processing Magazine 2015;32(3):53–69. https://doi.org/10.1109/MSP.2014.2347059.
[16] Zhou L, Deng W, Wu X. Robust image segmentation quality assessment. arXiv preprint arXiv:190308773. 2019.
[17] Claessens M, Vanreusel V, De Kerf G, Mollaert I, Lofman F, Gooding MJ, et al. Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm. Phys Med Biol 2022. doi: 101088/1361-6560/ac6fad.
[18] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiotherapy Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.
[19] Isaksson LJ, Summers P, Bhalerao A, Gandini S, Raimondi S, Pepa M, et al. Quality assurance for automatically generated contours with additional deep learning. Insights into Imaging 2022;13(1):1–10. https://doi.org/10.1186/s13244-022012767.
[20] Jungo A, Balsiger F, Reyes M. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Front Neurosci 2020;282. https://doi.org/10.3389/fnins.2020.00282.
[21] Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing. 2019;338:34–45. https://doi.org/10.48550/arXiv.1807.07356.
[22] Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical Image Anal 2020;59:101557. https://doi.org/10.1016/j.media.2019.101557.
[23] Kwon Y, Won JH, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput Stat Data Anal 2020;142:106816. https://doi.org/10.1016/j.csda.2019.106816.

[24] Hu S, Worrall D, Knegt S, Veeling B, Huisman H, Welling M. Supervised uncertainty quantification for segmentation with multiple annotations. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. Springer; 2019. p. 137–45. doi: 10.1007/978-3030322458_16.
[25] Do HP, Guo Y, Yoon AJ, Nayak KS. Accuracy, uncertainty, and adaptability of automatic myocardial ASL segmentation using deep CNN. Magnetic Resonance Med 2020;83(5):1863–74. https://doi.org/10.1002/mrm.28043.
[26] Williams E, Niehaus S, Reinelt J, Merola A, Glad Mihai P, Villringer K, et al. Automatic quality control framework for more reliable integration of machine learning-based image segmentation into medical workflows. arXiv e-prints. 2021: arXiv-2112. doi: 10.48550/arXiv.2112.03277.
[27] Pan H, Feng Y, Chen Q, Meyer C, Feng X. Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE; 2019. p. 468–71. https://doi.org/10.1109/ISBI.2019.8759300.
[28] Liu F, Xia Y, Yang D, Yuille AL, Xu D. An alarm system for segmentation algorithm based on shape model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 10652–61. https://doi.org/10.1109/ICCV.2019.01075.
[29] Brusini I, Padilla DF, Barroso J, Skoog I, Smedby Ö, Westman E, et al. A deep learning-based pipeline for error detection and quality control of brain MRI segmentation results. arXiv preprint arXiv:200513987. 2020 doi: 1048550/arXiv200513987.
[30] Robinson R, Valindria VV, Bai W, Oktay O, Kainz B, Suzuki H, et al. Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. J Cardiovascular Magnetic Reson 2019;21(1):1–14. https://doi.org/10.1186/s12968-0190523x.
[31] Alba X, Lekadir K, Pereanez M, Medrano-Gracia P, Young AA, Frangi AF. Automatic initialization and quality control of large-scale cardiac MRI segmentations. Medical Image Anal 2018;43:129–41. https://doi.org/10.1016/j.media.2017.10.001.
[32] Tarroni G, Oktay O, Bai W, Schuh A, Suzuki H, Passerat-Palmbach J, et al. Learning-based quality control for cardiac MR images. IEEE Trans Medical Imaging 2018;38(5):1127–38. https://doi.org/10.1109/TMI.2018.2878509.
[33] Kohlberger T, Singh V, Alvino C, Bahlmann C, Grady L. Evaluating segmentation error without ground truth. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1–5, 2012, Proceedings, Part I 15. Springer; 2012. p. 528–36. doi: 10.1007/978-3642334153_65.
[34] DeVries T, Taylor GW. Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:180700502. 2018 doi: 1048550/arXiv180700502.
[35] Robinson R, Oktay O, Bai W, Valindria VV, Sanghvi MM, Aung N, et al. Real-time prediction of segmentation quality. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11. Springer; 2018. p. 578–85. doi: 10.48550/arXiv.1806.06244.
[36] Valindria VV, Lavdas I, Bai W, Kamnitsas K, Aboagye EO, Rockall AG, et al. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans Med Imaging 2017;36(8):1597–606. https://doi.org/10.1109/TMI.2017.2665165.
[37] Wang S, Tarroni G, Qin C, Mo Y, Dai C, Chen C, et al. Deep generative model-based quality control for cardiac MRI segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer; 2020. p. 88–97. doi: 10.1007/978-3030597191_9.
[38] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41. https://doi.org/10.1007/978-3319245744_28.
[39] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2016. p. 424–32. https://doi.org/10.1007/978-3319467238_49.
[40] Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV). IEEE; 2016. p. 565–71. https://doi.org/10.1109/3DV.2016.79.
[41] Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. International MICCAI Brainlesion Workshop. Springer; 2018. p. 311–20. https://doi.org/10.1007/9783030117269_28.
[42] Kano Y, Ikushima H, Sasaki M, Haga A. Automatic contour segmentation of cervical cancer using artificial intelligence. J Rad Res 2021;62(5):934–44. https://doi.org/10.1093/jrr/rrab070.
[43] Bishop CM, et al. Neural networks for pattern recognition. Oxford University Press; 1995.
[44] van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-170339.
[45] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30. https://doi.org/10.1016/B978-0120887705/50067-8.
[46] Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079–107.
[47] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422. https://doi.org/10.1023/A:1012487302797.
[48] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[49] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inform Process Manage 2009;45(4):427–37. https://doi.org/10.1016/j.ipm.2009.03.002.
[50] Riffenburgh RH. Chapter Summaries. In: Riffenburgh RH, editor. Statistics in Medicine (Second Edition). second edition ed. Burlington: Academic Press; 2006. p. 533–80. https://doi.org/10.1016/B978-0120887705/50067-8.
[51] Sadri AR, Janowczyk A, Zhou R, Verma R, Beig N, Antunes J, et al. MRQy—An open-source tool for quality control of MR imaging data. Med Phys 2020;47(12):6029–38. https://doi.org/10.1002/mp.14593.
[52] Bottani S, Burgos N, Maire A, Wild A, Ströer S, Dormont D, et al. Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. Med Image Anal 2022;75:102219. https://doi.org/10.1016/j.media.2021.102219.
[53] Rodríguez Outeiral R, González PJ, Schaake EE, van der Heide UA, Simões R. Deep learning for segmentation of the cervical cancer gross tumor volume on magnetic resonance imaging for brachytherapy. Rad Oncol 2023;18(1):91. https://doi.org/10.1186/s13014-023022838.
[54] Tanderup K, Pötter R, Lindegaard J, Kirisits C, Juergenliemk-Schulz I, De Leeuw A, et al. Image guided intensity modulated External beam radiochemotherapy and MRI based adaptive BRAchytherapy in locally advanced CErvical cancer EMBRACE-II. EMBRACE II study protocol. 2015;1. doi:10.1016/j.ctro.2018.01.001.
[55] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. International MICCAI Brainlesion Workshop. Springer; 2021. p. 272–84. https://doi.org/10.1007/978-3031089992_22.
[56] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. arXiv preprint arXiv:230402643. 2023 doi: 1048550/arXiv230402643.
[57] Ma J, He Y, Li F, Han L, You C, Wang B. Segment Anything in Medical Images. arXiv preprint arXiv:230412306. 2023 doi: 101038/s41467-024-44824-z.
[58] Carré A, Klausner G, Edjlali M, Lerousseau M, Briend-Diop J, Sun R, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Sci Rep 2020;10(1):12340. https://doi.org/10.1038/s41598-02069298z.