

Deep learning-assisted target volume segmentation for uterine malignancies on non-contrast CT

zhiwei liu

Ganzhou Cancer Hospital

YuanWu Zhong

Ganzhou Cancer Hospital

lei tan

Ganzhou Cancer Hospital

weiwei wu

Ganzhou Cancer Hospital

xiaoyun lai

441393027@qq.com

Ganzhou Cancer Hospital

Article

Keywords: deep learning, auto-segmentation, non-contrast computed tomography, uterine malignancies, clinical target volume, radiotherapy

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9857280/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Deep learning-assisted target volume segmentation for uterine malignancies on non-contrast CT

Zhi Wei Liu^{1†}, Yuan Wu Zhong^{2†}, Lei Tan¹, Wei Wei Wu^{1*}, XiaoYun Lai^{1*}

1. Center of Radiation Oncology, Ganzhou Cancer Hospital, Ganzhou, Jiangxi, China

2. Department of Bone and Soft Tissue Oncology, Ganzhou Cancer Hospital, Ganzhou, Jiangxi, China

†Zhi Wei Liu and Yuan Wu Zhong contributed equally to this work.

* Wei Wei Wu and XiaoYun Lai contributed equally as co-corresponding authors.

Corresponding Author: Weiwei Wu, 62009626@qq.com

Corresponding Author: XiaoYun Lai, 441393027@qq.com

Abstract

Background: Accurate delineation of clinical target volume (CTV) and planning target volume (PTV) is essential for radiotherapy in uterine malignancies. However, the limited soft-tissue contrast of non-contrast computed tomography (NCCT) makes automated target segmentation challenging.

Methods: This single-center retrospective study included 157 unique patients. Of these, 137 patients were used for model development and internal testing, including a development cohort of 110 patients and an independent test cohort of 27 patients. A separate cohort of 20 patients was used for workflow evaluation. A multi-label nnU-Net model was developed to simultaneously segment the CTV and PTV. Four nnU-Net configurations were compared using five-fold cross-validation. Segmentation performance was evaluated using the Dice similarity coefficient (DSC) and the 95th percentile Hausdorff distance (HD95). Three radiation oncologists independently assessed the clinical acceptability of the auto-contours using a four-point Likert scale. Workflow efficiency was evaluated by comparing manual contouring with AI-assisted contour editing.

Results: The 3D-fullres configuration achieved the numerically highest overall cross-validation performance and was selected for independent internal testing. In the 27-patient independent test cohort, the mean DSC was $90.02\% \pm 2.33\%$ for PTV and $72.16\% \pm 4.54\%$ for CTV. The corresponding median HD95 values were 5.12 mm (interquartile range [IQR], 5.00–7.60 mm) and 5.55 mm (IQR, 5.00–9.50 mm), respectively. Rating-level clinical acceptability was 91.4% for PTV and 75.3% for CTV. Inter-rater agreement was higher for PTV than for CTV. In the separate workflow cohort, AI-assisted contouring reduced the mean total task time from 65.4 to 14.2 minutes per case, corresponding to an observed reduction of 78.3% ($P < 0.001$).

Conclusion: The multi-label 3D-fullres nnU-Net achieved high PTV overlap and moderate CTV overlap on internal NCCT data and reduced contouring time in a separate workflow cohort. The model may provide editable initial contours, but CTV predictions require careful physician review. External validation and dosimetric assessment are required before routine clinical use.

Keywords: deep learning; auto-segmentation; non-contrast computed tomography; uterine malignancies; clinical target volume; radiotherapy.

1. Introduction

Uterine malignancies, primarily cervical and endometrial cancers, account for a substantial proportion of gynecologic cancers worldwide. Radiotherapy is a key component of definitive treatment for cervical cancer, postoperative adjuvant treatment for endometrial cancer, and multidisciplinary management strategies [1–3]. With the increasing use of highly conformal techniques such as intensity-modulated radiotherapy and volumetric modulated arc therapy, treatment quality depends heavily on accurate target volume definition. Standardized delineation of the CTV and PTV is essential for ensuring adequate target coverage, limiting dose to organs at risk, and maintaining quality control in pelvic radiotherapy [4–7].

Pelvic target delineation remains one of the most time-consuming and operator-dependent steps in the radiotherapy workflow. The uterus, cervix, vagina, and adjacent pelvic soft tissues may have poorly defined boundaries on planning CT. Variations in bladder and rectal filling can further alter the position and morphology of the uterus and upper vagina, increasing uncertainty in CTV definition. As a result, manual contouring is susceptible to both inter-observer and intra-observer variability [8–10]. Such variation may increase clinical workload and affect treatment planning consistency, particularly in multi-center settings.

Deep learning-based auto-segmentation has been proposed as a means to improve the efficiency and consistency of radiotherapy contouring. Recent reviews have highlighted its growing role in radiotherapy planning, while also emphasizing the need for careful assessment of generalizability, quality assurance, and physician oversight before clinical deployment [11–14]. From a methodological perspective, U-Net and its three-dimensional extensions have formed the basis of many medical image segmentation models. V-Net introduced Dice-based optimization for volumetric segmentation, and nnU-Net has further improved reproducibility by automatically adapting preprocessing, network configuration, and training strategy to the dataset [15–18]. Nevertheless, segmentation of abdominal and pelvic CT images remains challenging because of complex anatomy, low-contrast boundaries, and limited cross-dataset generalizability [19].

Several studies have investigated automated CTV delineation for cervical and other gynecologic cancers, demonstrating the potential of deep learning models for target contouring using CT or multimodal imaging [22–29]. Nevertheless, automated delineation of gynecologic target volumes remains challenging because of complex pelvic anatomy, variable bladder and rectal filling, and indistinct soft-tissue boundaries. These challenges are particularly relevant on NCCT, where the interfaces among the tumor, normal uterine structures, vaginal fornix, pelvic musculature, and nodal drainage regions may be difficult to identify. Although MRI and MRI–CT fusion can improve soft-tissue visualization, these imaging resources are not uniformly incorporated into routine simulation workflows [30–33]. Therefore, further evaluation of target auto-segmentation under standardized NCCT conditions remains clinically relevant.

To address these challenges, we developed a multi-label nnU-Net model for simultaneous CTV and PTV segmentation on NCCT in patients with uterine malignancies. Rather than relying solely on geometric overlap metrics, we used a multidimensional internal evaluation framework that included model-configuration comparison, quantitative segmentation analysis, 3D surface-error visualization, independent multi-reader clinical assessment, and workflow-time evaluation [20,21,34–40]. Specifically, this study aimed to: (1) compare four nnU-Net configurations using five-fold cross-validation; (2) evaluate the selected model using geometric metrics, spatial error visualization, and multi-reader assessment; and (3) quantify the change in contouring time

associated with AI-assisted editing in a separate workflow cohort. The present study was designed as an internal feasibility evaluation rather than a definitive assessment of external generalizability, dosimetric equivalence, or routine clinical implementation. The overall study workflow is summarized in Figure 1. The remainder of this article is organized as follows. Section 2 describes the study cohorts, image acquisition, reference-contour generation, model development, and evaluation methods. Section 3 presents the geometric, subgroup, multi-reader, and workflow results. Sections 4 and 5 discuss the findings, limitations, and conclusions.

2. Materials and Methods

The Methods section describes three sequential study components: model development and five-fold cross-validation in the development cohort, geometric and multi-reader evaluation in the independent test cohort, and workflow evaluation in a separate cohort. Figure 1 summarizes patient allocation, model selection, independent testing, expert review, and workflow assessment. As shown in Figure 1, 110 patients were used for model development and cross-validation, 27 patients were reserved for independent internal testing, and 20 separate patients were used for workflow evaluation. Four nnU-Net configurations were compared, after which the selected 3D-fullres model underwent geometric, multi-reader, spatial-error, and workflow assessments.

2.1 Patient Cohort and Ethical Considerations

This single-center retrospective study was approved by the Institutional Review Board of Ganzhou Cancer Hospital (No. (2025) Ke-Lun-Shen (89)), and the requirement for informed consent was waived. A total of 157 unique patients with uterine malignancies, including cervical and endometrial cancers, were included. Among them, 137 patients who received definitive or postoperative adjuvant radiotherapy between 2024 and 2025 were used for model development and internal testing.

Stratified random sampling according to pathological type and FIGO stage was used to allocate 110 patients to the development cohort and 27 patients to the independent internal test cohort. The development cohort was used for model training, configuration comparison, and five-fold cross-validation, whereas the independent test cohort was reserved for final model evaluation. A separate cohort of 20 patients was used exclusively for workflow evaluation and secondary segmentation-performance assessment. These patients were not included in model training, cross-validation, hyperparameter selection, configuration selection, or independent test-set evaluation. Therefore, no patient overlap occurred among the development, independent test, and workflow cohorts. Baseline characteristics of the development and independent test cohorts are presented in Table 1.

The inclusion criteria were as follows: (1) histologically confirmed cervical or endometrial malignancy treated with definitive pelvic external-beam radiotherapy or postoperative pelvic external-beam radiotherapy; (2) age between 18 and 75 years; (3) a radiotherapy plan containing a single CTV structure; (4) availability of complete pretreatment imaging data suitable for target delineation and segmentation analysis; (5) availability of complete clinical, pathological, treatment, and radiotherapy-planning information; and (6) absence of synchronous or previously treated other malignancies. Patients were excluded if they had: (1) pelvic or para-aortic lymph-node metastasis; (2) multiple CTV structures in the radiotherapy plan or target-volume definitions incompatible with the predefined single-CTV segmentation task; (3) previous pelvic radiotherapy; or (4) incomplete clinical, imaging, or radiotherapy-planning data that precluded study analysis. Previous or concurrent chemotherapy was not an exclusion criterion. Receipt of brachytherapy after pelvic external-beam radiotherapy also did not affect eligibility; however, brachytherapy images and target

volumes were not included in the present segmentation analysis. Patients treated with brachytherapy alone without pelvic external-beam radiotherapy were not eligible.

Regarding treatment setting, 5 of 110 patients (4.5%) in the development cohort received definitive pelvic radiotherapy and 105 (95.5%) received postoperative adjuvant radiotherapy. All 27 patients in the independent internal test cohort received postoperative adjuvant radiotherapy.

2.2 Image Acquisition and Reference Delineation

All patients underwent NCCT simulation using a Philips Brilliance CT Big Bore scanner. To reduce geometric uncertainty related to bladder and rectal filling, patients followed a standardized pelvic radiotherapy preparation protocol. They were instructed to empty the bladder and then drink a fixed volume of water 1 hour before simulation to achieve moderate bladder filling, and to empty the rectum as far as possible. Patients were positioned supine and immobilized using a customized vacuum cushion.

CT images were acquired with a tube voltage of 120 kVp, automatic tube current modulation, slice thickness and increment of 3.0 mm, and an in-plane reconstruction resolution of approximately $1.0 \times 1.0 \text{ mm}^2$. The scan range extended from the L4/L5 intervertebral level to approximately 3 cm below the inferior border of the obturator foramen.

Reference CTV contours were initially delineated on the NCCT simulation images in the treatment planning system by a senior radiation oncologist with 10 years of experience in gynecologic radiotherapy and subsequently reviewed by a second senior radiation oncologist with 20 years of experience. During delineation and review, the radiation oncologists had access to the available diagnostic pelvic MRI images, which were used as complementary anatomical information to assist in identifying tumor extent and low-contrast soft-tissue boundaries. The MRI images were used only as an auxiliary clinical reference and were not provided as input to the nnU-Net model. CTV delineation followed the relevant RTOG and ESTRO consensus guidelines [4,5]. When modifications were considered necessary, the second physician revised the initial contours, and the final clinically approved contours were used as the reference standard.

For all included patients, the clinically approved CTV was represented as a single combined structure for model training and evaluation. Separate primary and regional nodal CTV labels were not used. The anatomical composition of the CTV was defined according to the applicable treatment setting and consensus guideline for definitive or postoperative radiotherapy [4,5]; clinically indicated primary/postoperative target and elective regional nodal components were contained within the single CTV label rather than encoded as separate structures.

The reference PTV was generated from the approved CTV using an isotropic 5-mm expansion according to institutional planning practice. The final verified CTV and PTV contours were used for model training and evaluation. Formal inter-observer variability in the reference contours was not quantified because only the final clinically approved contours were retained.

2.3 Model Architecture and Training Protocol

A single multi-label segmentation model was developed using nnU-Net [18]. The CTV and PTV were encoded as separate segmentation labels, allowing the model to simultaneously predict both target volumes. Only the NCCT simulation images were used as model input; the diagnostic MRI images reviewed during reference-contour generation were not provided to the network. The predicted PTV was generated directly by the model and was not derived by applying a geometric expansion to the predicted CTV.

The nnU-Net framework automatically configured image preprocessing, network architecture,

patch size, batch size, and training strategy according to dataset characteristics, including image dimensions, voxel spacing, and class distribution [18]. During preprocessing, foreground Hounsfield Unit values were clipped to the 0.5th–99.5th percentiles to reduce the influence of extreme intensity values, followed by Z-score normalization.

Four nnU-Net configurations—2D, 3D-lowres, 3D-fullres, and 3D-cascade—were compared within the development cohort using identical five-fold cross-validation partitions [15,16,18]. CTV and PTV performance metrics were calculated separately for each validation fold. The model from the fold with the highest combined CTV and PTV validation performance was used for independent testing. A combined Dice and cross-entropy loss was used to optimize volumetric overlap while accounting for the imbalance between target and background voxels [17,18].

Model development was performed using Python version 3.10.18, PyTorch version 2.7.1, nnU-Net version 2.6.4, and CUDA version 11.8 on a workstation equipped with an NVIDIA GeForce RTX 4060 Ti GPU. Model training was performed for 250 epochs, with 250 training iterations and 50 validation iterations per epoch. The initial learning rate was 1×10^{-2} , and the weight decay was 3×10^{-5} . Foreground oversampling was set to 33% to increase the representation of target-containing patches, whereas probabilistic oversampling was disabled. Deep supervision was enabled throughout training. The patch size and batch size were automatically configured by nnU-Net according to the dataset properties and available GPU memory. Other optimization, learning-rate scheduling, and inference settings followed the predefined nnU-Net trainer configuration.

2.4 Geometric Accuracy Evaluation and 3D Error Mapping

Segmentation performance was assessed using DSC and HD95, which provide complementary information on volumetric overlap and surface boundary error [20,21].

The DSC was calculated as:

$$DSC(P, G) = \frac{2|P \cap G|}{|P| + |G|}$$

where $|P \cap G|$ represents the number of overlapping voxels. The DSC ranges from 0 to 1, with higher values indicating superior overlap.

HD95 was used to assess boundary distance while reducing the influence of extreme outliers. It was defined as:

$$h_{95}(P, G) = \text{percentile}_{95} \left\{ \min_{g \in \partial G} d(p, g) \mid p \in \partial P \right\}$$

$$HD_{95}(P, G) = \max(h_{95}(P, G), h_{95}(G, P))$$

where ∂P and ∂G denote the surface point sets of the predicted and reference contours, respectively, and $d(p, g)$ represents the Euclidean distance. A lower HD95 (measured in millimeters) indicates a closer proximity between the AI-generated and reference boundaries.

To further characterize spatial error patterns, 3D surface error maps were generated by projecting local surface distances between the predicted and reference contours onto the target surface. A consistent 0–10 mm color scale was used to visualize local deviations and identify anatomically relevant regions of surface disagreement.

2.5 Multi-Reader Clinical Acceptability Assessment

Clinical acceptability was evaluated in the 27-patient independent test cohort because geometric metrics alone may not fully reflect clinical usability [34–40]. Three radiation oncologists, each with more than 5 years of experience in gynecologic radiation oncology, independently reviewed the AI-generated CTV and PTV contours. Two of the three readers were the radiation

oncologists who had participated in generation and review of the reference contours (10 and 20 years of experience, respectively), whereas the third reader was the attending radiation oncologist who performed the workflow evaluation. Therefore, the readers were not fully independent of the reference-contour generation and workflow components of the study. During the rating process, however, the three readers evaluated the contours independently and were blinded to the nnU-Net configuration, quantitative segmentation metrics, and the ratings assigned by the other readers.

Each target was assessed using a four-point Likert scale:

Scoring	Definition
4 points	Excellent quality; meets clinical standards; usable without modification.
3 points	Minor deviations; usable after minor editing.
2 points	Significant deviations; requires major revision.
1 point	Unacceptable; unsuitable for clinical use.

Scores of 3 or 4 were classified as clinically acceptable. Each reader evaluated 27 CTV and 27 PTV contours, resulting in 81 ratings per target and 162 target-level ratings in total. Rating-level acceptability was defined as the proportion of all ratings with scores of 3 or 4. Case-level acceptability was additionally defined as agreement by at least two of the three readers that the contour was clinically acceptable. When the readers disagreed regarding the clinical acceptability of an auto-contour, the case was reviewed by the most senior radiation oncologist, whose assessment was used as the final clinical decision. The adjudication process did not replace or modify the original independent ratings. All original reader scores were retained for the calculation of rating distributions, rating-level acceptability, intraclass correlation coefficients, and Fleiss' kappa.

2.6 Clinical Workflow Efficiency Evaluation

Workflow efficiency was evaluated in the separate 20-patient cohort. The same attending radiation oncologist completed both AI-assisted and manual contouring tasks. For all cases, AI-assisted contouring was performed first, followed by manual contouring; therefore, the order of the two conditions was fixed and was not randomized. No formal washout period was implemented between the AI-assisted and manual contouring conditions.

In the AI-assisted condition, nnU-Net-generated contours were imported into the treatment planning system, reviewed, edited where necessary, and finalized by the physician. In the manual condition, the physician delineated the CTV and PTV de novo.

Total task time was measured from the beginning to the completion of each contouring workflow and included image loading, AI-contour import when applicable, preliminary image review, contour creation or editing, final review, and saving. AI-assisted contouring time was recorded by a study author, whereas manual contouring time was recorded by the physician.

Because the task order was fixed and no formal washout period was implemented, potential order, recall, and learning effects were considered when interpreting the workflow results [34–40].

2.7 Statistical Analysis

Continuous variables were assessed for normality using the Shapiro–Wilk test. Normally distributed variables are presented as mean $\pm$ standard deviation, whereas skewed variables are presented as median and interquartile range. Categorical variables are reported as counts and percentages. Baseline characteristics were compared using the Mann–Whitney U test, chi-square test, or Fisher's exact test, as appropriate.

DSC was summarized using mean $\pm$ standard deviation, with case-level nonparametric bootstrap 95% confidence intervals. Because HD95 showed a markedly skewed distribution, it was

primarily summarized using median and interquartile range, with mean $\pm$ standard deviation additionally reported to facilitate comparison with previous studies.

Inter-rater agreement for the four-point Likert scores was evaluated using a two-way random-effects, absolute-agreement intraclass correlation coefficient for single measurements [ICC(2,1)]. The reliability of the average rating across the three readers was additionally summarized using ICC(2,k). After dichotomizing scores into clinically acceptable (scores 3–4) and unacceptable (scores 1–2), inter-rater agreement was evaluated using Fleiss' kappa. Ninety-five percent confidence intervals for acceptability proportions were calculated using the Wilson method, whereas confidence intervals for agreement coefficients were estimated using case-level bootstrap resampling.

Pearson correlation analysis was used to assess the association between target volume and DSC, with Spearman correlation performed as a sensitivity analysis. Pearson correlation coefficients were reported with 95% confidence intervals. FIGO subgroup analyses were exploratory and interpreted descriptively.

Manual and AI-assisted contouring times were compared using a paired-samples t-test after assessing the distribution of the within-patient time differences. All statistical analyses were performed using SPSS version 27.0 and Python version 3.9. A two-sided $P < 0.05$ was considered statistically significant.

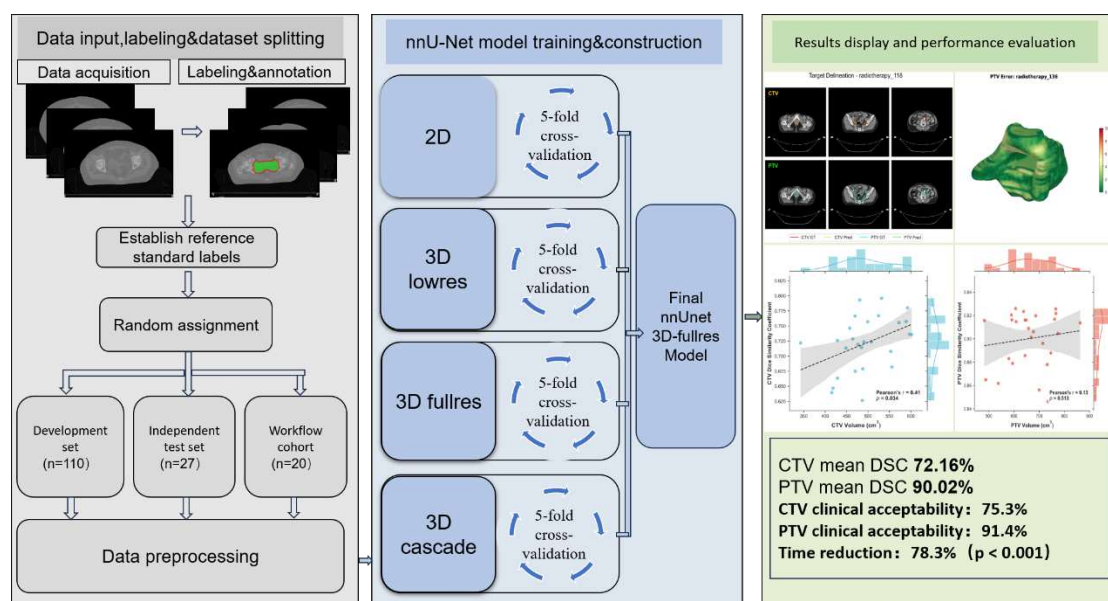


Figure 1. Study workflow and multi-dimensional evaluation framework for NCCT-based target volume auto-segmentation in uterine malignancies.

3 Results

3.1 Patient Baseline Characteristics

A total of 157 unique patients were included. Among them, 110 patients were assigned to the development cohort, 27 to the independent internal test cohort, and 20 to the separate workflow cohort. The development and independent test cohorts were generated using stratified random sampling according to pathological type and FIGO stage. Baseline characteristics of the development and independent test cohorts are summarized in Table 1.

No statistically significant differences were detected in the reported baseline characteristics between the development and independent test cohorts. The median age was 55 years (range, 27–85 years)

in the development cohort and 59 years (range, 47–73 years) in the independent test cohort ($p=0.056$). Squamous cell carcinoma was the most common pathological type, accounting for 61.8% and 48.1% of patients in the development and independent test cohorts, respectively. No statistically significant difference was detected in pathological-type distribution between the two cohorts ($p=0.571$). The distribution of FIGO stage was also not significantly different between the development and independent test cohorts ($p=0.308$). Regarding treatment setting, 5 of 110 patients (4.5%) in the development cohort received definitive radiotherapy, whereas 105 (95.5%) received postoperative adjuvant radiotherapy. All 27 patients in the independent test cohort received postoperative adjuvant radiotherapy. The treatment-setting distribution did not differ significantly between the two cohorts (Fisher’s exact test, $P=0.583$).

Table 1. Baseline characteristics of the development and independent test cohorts.

Characteristic	Development cohort (n=110)	Independent test cohort (n=27)	P value
Age, median (range), years	55 (27-85)	59 (47-73)	0.056
Pathological types, n (%)			
Squamous cell carcinoma	68 (61.8)	13 (48.1)	0.571
Adenocarcinoma	30 (27.3)	9 (33.3)	
Adenosquamous carcinoma	2 (1.8)	1 (3.7)	
Others	10 (9.1)	4 (14.8)	
FIGO stage, n (%)			
I	51 (46.4)	8 (29.6)	0.308
II	27 (24.5)	11 (40.7)	
III	31 (28.2)	8 (29.6)	
IV	1 (0.9)	0 (0.0)	
Treatment setting, n (%)			
Definitive radiotherapy	5 (4.5)	0 (0.0)	0.583
Postoperative adjuvant radiotherapy	105 (95.5)	27 (100.0)	

(Table note: FIGO, International Federation of Gynecology and Obstetrics. Percentages may not total 100% because of rounding.)

3.2 Comparison of nnU-Net Configurations

Four nnU-Net configurations, including 2D, 3D-lowres, 3D-fullres, and 3D-cascade, were compared using identical five-fold cross-validation partitions within the development cohort. CTV and PTV performance metrics were calculated separately for each validation fold. The cross-validation results are presented in Table 2.

The 3D-fullres configuration achieved the numerically highest overall performance across the evaluated CTV and PTV metrics. For CTV segmentation, the mean DSC and HD95 across the five folds were 72.14% and 13.09 mm, respectively. For PTV segmentation, the corresponding mean DSC and HD95 were 88.13% and 15.42 mm. The 2D configuration showed greater between-fold variability, particularly for CTV segmentation, suggesting that slice-based processing may provide insufficient three-dimensional anatomical context for complex pelvic target volumes. Based on its overall cross-validation performance across both target labels, the 3D-fullres configuration was selected for independent internal testing.

Table 2. Five-fold cross-validation performance of the four nnU-Net configurations.

Method	Target	Metric	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Average
2D	CTV	DSC (%)	72.30±5.89	68.50±9.05	65.78±17.82	71.40±5.82	69.48±7.84	69.49
		HD95 (mm)	10.02±13.48	36.71±18.90	21.57±27.74	11.32±16.78	12.60±16.75	18.44
	PTV	DSC (%)	89.83±3.12	77.28±20.66	84.95±17.12	89.97±2.77	88.44±4.38	86.09
		HD95 (mm)	12.65±14.14	33.70±30.03	25.50±33.44	11.44±17.49	17.11±22.38	20.08
3D-lowres	CTV	DSC (%)	72.43±6.46	60.12±25.66	69.51±16.86	73.04±3.89	71.17±7.32	69.25
		HD95 (mm)	10.34±15.88	27.20±38.23	18.88±30.63	12.06±21.04	13.33±19.56	16.36
	PTV	DSC (%)	89.95±3.64	81.36±25.15	86.34±16.74	89.92±2.97	88.66±5.55	87.25
		HD95 (mm)	10.68±15.01	22.99±35.11	20.85±35.21	12.93±21.00	14.48±21.37	16.39
3D-fullres	CTV	DSC (%)	73.37±6.21	70.43±8.80	70.44±16.92	74.27±4.15	72.19±7.63	72.14
		HD95 (mm)	9.08±14.18	14.67±21.66	19.10±31.81	10.41±17.76	12.17±17.57	13.09
	PTV	DSC (%)	90.28±3.35	84.39±20.86	86.82±16.77	90.29±3.03	88.86±5.62	88.13
		HD95 (mm)	9.42±13.21	20.11±30.43	19.69±33.44	13.58±22.06	14.32±21.25	15.42
3D-cascade	CTV	DSC (%)	72.17±5.81	66.88±18.20	69.55±17.12	73.67±4.28	71.23±7.73	70.70
		HD95 (mm)	8.40±7.81	18.60±28.56	20.59±34.10	19.62±20.41	13.59±20.63	16.16
	PTV	DSC (%)	89.34±3.32	84.89±20.36	86.81±16.78	86.15±2.81	88.42±5.40	87.12
		HD95 (mm)	16.04±17.77	18.75±30.67	20.75±35.32	19.90±19.83	19.09±24.26	18.91

3.3 Segmentation Performance in the Independent Test and Workflow Cohorts

The segmentation performance of the final 3D-fullres model in the independent test and workflow cohorts is summarized in Table 3. Because the HD95 values were markedly right-skewed, median and interquartile range (IQR) were used as the primary summary measures, while mean $\pm$ standard deviation was additionally reported to facilitate comparison with previous studies.

In the 27-patient independent test cohort, the model achieved a mean DSC of 90.02% $\pm$ 2.33% for PTV segmentation, with a bootstrap 95% confidence interval (CI) of 89.12%–90.84%. The median HD95 was 5.12 mm (IQR, 5.00–7.60 mm), whereas the mean HD95 was 8.75 $\pm$ 9.85 mm. For CTV segmentation, the mean DSC was 72.16% $\pm$ 4.54% (bootstrap 95% CI, 70.44%–73.80%). The corresponding median HD95 was 5.55 mm (IQR, 5.00–9.50 mm), whereas the mean HD95 was 9.04 $\pm$ 10.77 mm.

In the separate 20-patient workflow cohort, the model achieved a mean DSC of 80.55% $\pm$ 5.65% for PTV segmentation, with a bootstrap 95% CI of 78.05%–82.85%. The median HD95 was 6.14 mm (IQR, 1.76–14.33 mm), whereas the mean HD95 was 10.25 $\pm$ 12.34 mm. For CTV segmentation, the mean DSC was 68.54% $\pm$ 6.72% (bootstrap 95% CI, 65.79%–71.50%). The median HD95 was 3.22 mm (IQR, 1.67–9.33 mm), whereas the mean HD95 was 9.63 $\pm$ 12.82 mm.

Compared with the independent test cohort, the workflow cohort showed numerically lower DSC values for both targets, with a more pronounced reduction for PTV. The HD95 values in the workflow cohort also showed substantial variability, as reflected by the wide IQRs and the large differences between the median and mean values. These findings indicate that model performance varied across subsequent clinical cases, particularly in a limited number of cases with relatively large boundary deviations.

Table 3. Segmentation performance in the independent test and workflow cohorts.

cohort	Target	DSC, mean $\pm$ SD, %	Bootstrap 95% CI	HD ₉₅ , median (IQR), mm	HD95, mean $\pm$ SD, mm
Independent test cohort	PTV	90.02 $\pm$ 2.33	89.12–90.84	5.12 (5.00–7.60)	8.75 $\pm$ 9.85

Internal workflow cohort	PTV	80.55 ± 5.65	78.05–82.85	6.14(1.76–14.33)	10.25 ± 12.34
Independent test cohort	CTV	72.16 ± 4.54	70.44–73.80	5.55 (5.00–9.50)	9.04 ± 10.77
Internal workflow cohort	CTV	68.54 ± 6.72	65.79–71.50	3.22 (1.67–9.33)	9.63 ± 12.82

3.4 Exploratory Subgroup Analysis and Association Between Target Volume and DSC

Exploratory subgroup analyses were performed according to FIGO stage. The independent test cohort included 8 patients with FIGO stage I disease, 11 with stage II disease, and 8 with stage III disease. Stage-specific segmentation performance is summarized in Table 4.

For CTV segmentation, the mean DSC values were 71.18% ± 5.08%, 72.69% ± 4.15%, and 72.41% ± 4.95% for FIGO stages I, II, and III, respectively. The corresponding mean HD95 values were 14.01 ± 19.18 mm, 6.39 ± 2.14 mm, and 7.72 ± 3.91 mm.

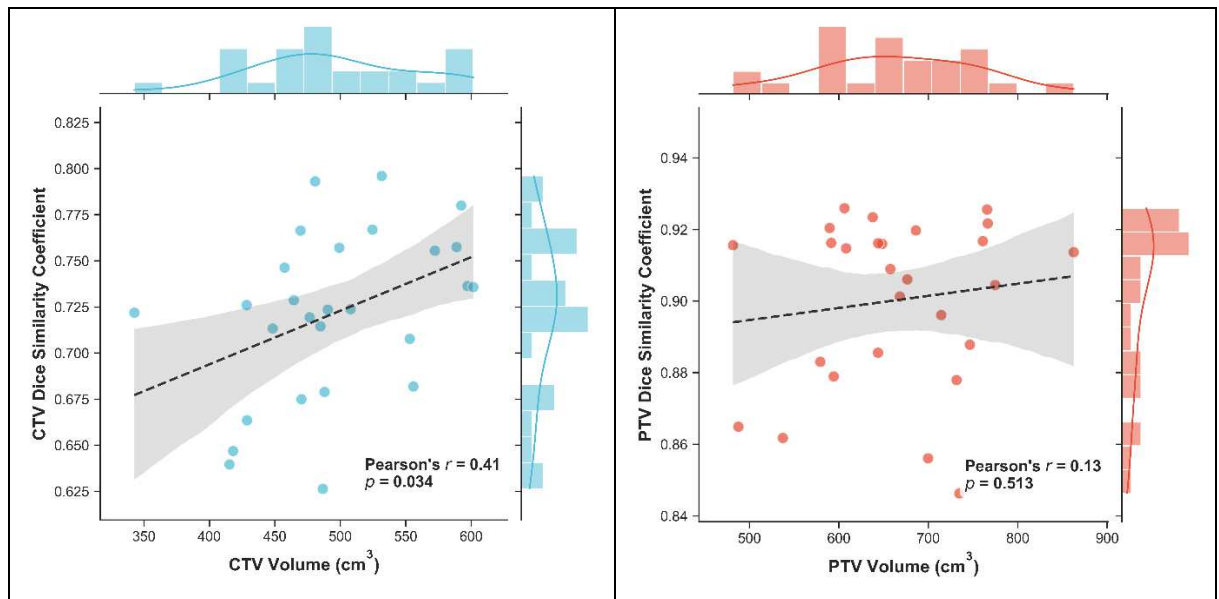
For PTV segmentation, the mean DSC values were 89.27% ± 3.17%, 90.41% ± 1.86%, and 90.24% ± 2.04% for FIGO stages I, II, and III, respectively. The corresponding mean HD95 values were 14.19 ± 17.17 mm, 6.41 ± 3.09 mm, and 6.52 ± 1.87 mm. Because of the limited number of patients in each subgroup, these analyses were considered exploratory and descriptive.

The relationship between target volume and DSC is shown in Figure 2. CTV volume was positively correlated with DSC (Pearson $r=0.410$, 95% CI, 0.035–0.683; $P=0.034$). Spearman correlation analysis yielded a similar result ($\rho=0.444$, $P=0.020$). In contrast, no statistically significant association was observed between PTV volume and DSC (Pearson $r=0.132$, 95% CI, –0.261 to 0.487; $P=0.513$; Spearman $\rho=0.073$, $P=0.719$).

Table 4. Exploratory segmentation performance according to FIGO stage.

Dataset	Target	DSC, mean ± SD, %	HD95, mean ± SD, mm
FIGO stage I	CTV	71.18±5.08	14.01 ± 19.18
(n=8)	PTV	89.27±3.17	14.19±17.17
FIGO stage II	CTV	72.69 ± 4.15	6.39 ± 2.14
(n=11)	PTV	90.41 ± 1.86	6.41 ± 3.09
FIGO stage III	CTV	72.41±4.95	7.72 ± 3.91
(n=8)	PTV	90.24±2.04	6.52±1.87

Figure 2. Association between target volume and Dice similarity coefficient in the independent test cohort.

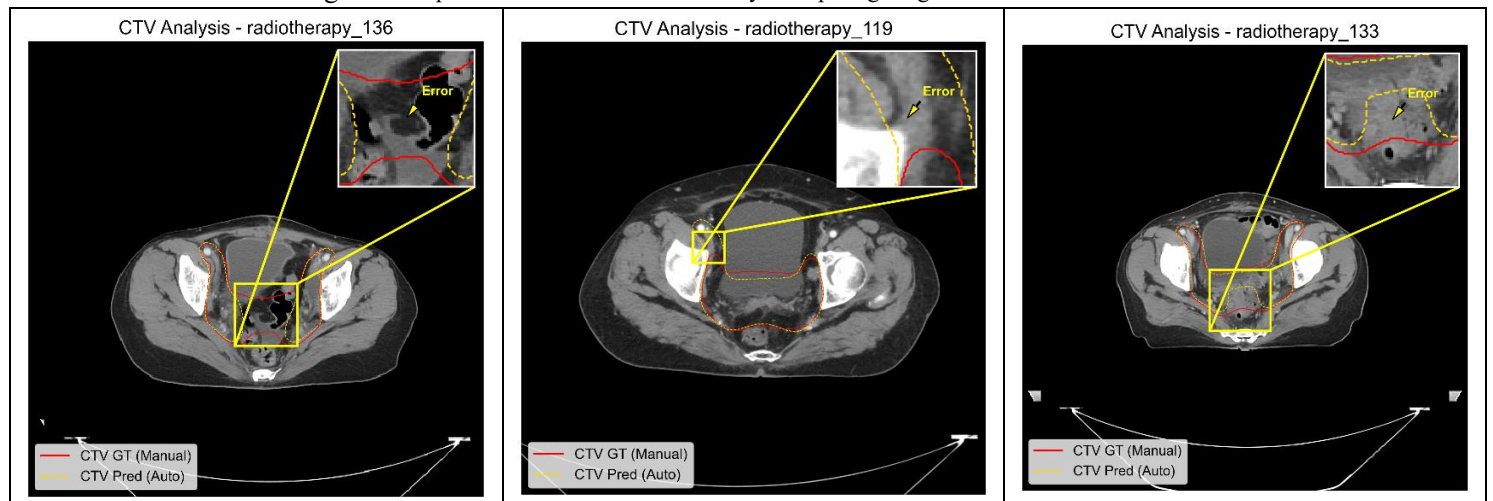


3.5 Spatial Distribution of Segmentation Errors

Representative axial contour overlays comparing the AI-generated and reference contours are shown in Figure 3. In the displayed cases, the predicted CTV and PTV contours generally followed the corresponding reference contours across most axial slices. Local discrepancies were visually observed near the cranio-caudal target boundaries and at several low-contrast anatomical interfaces, including the uterine fundus, superior extent of the nodal target, vaginal fornix, and rectal interface.

Representative 3D surface-error maps are shown in Figure 4. Most displayed target surfaces showed relatively small local deviations between the predicted and reference contours, whereas larger surface distances were observed in limited anatomical regions. The maps were used to illustrate the spatial distribution of local contour differences in selected cases and were not intended as a formal quantitative regional error analysis.

Figure 3. Representative axial contour overlays comparing AI-generated and reference contours.



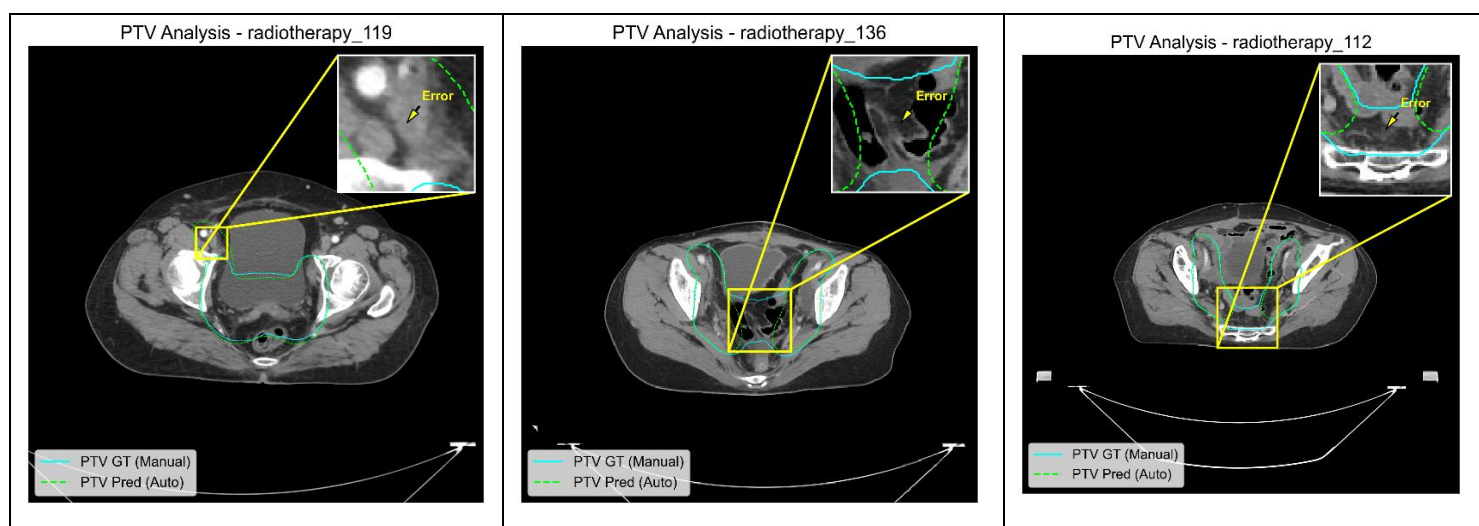
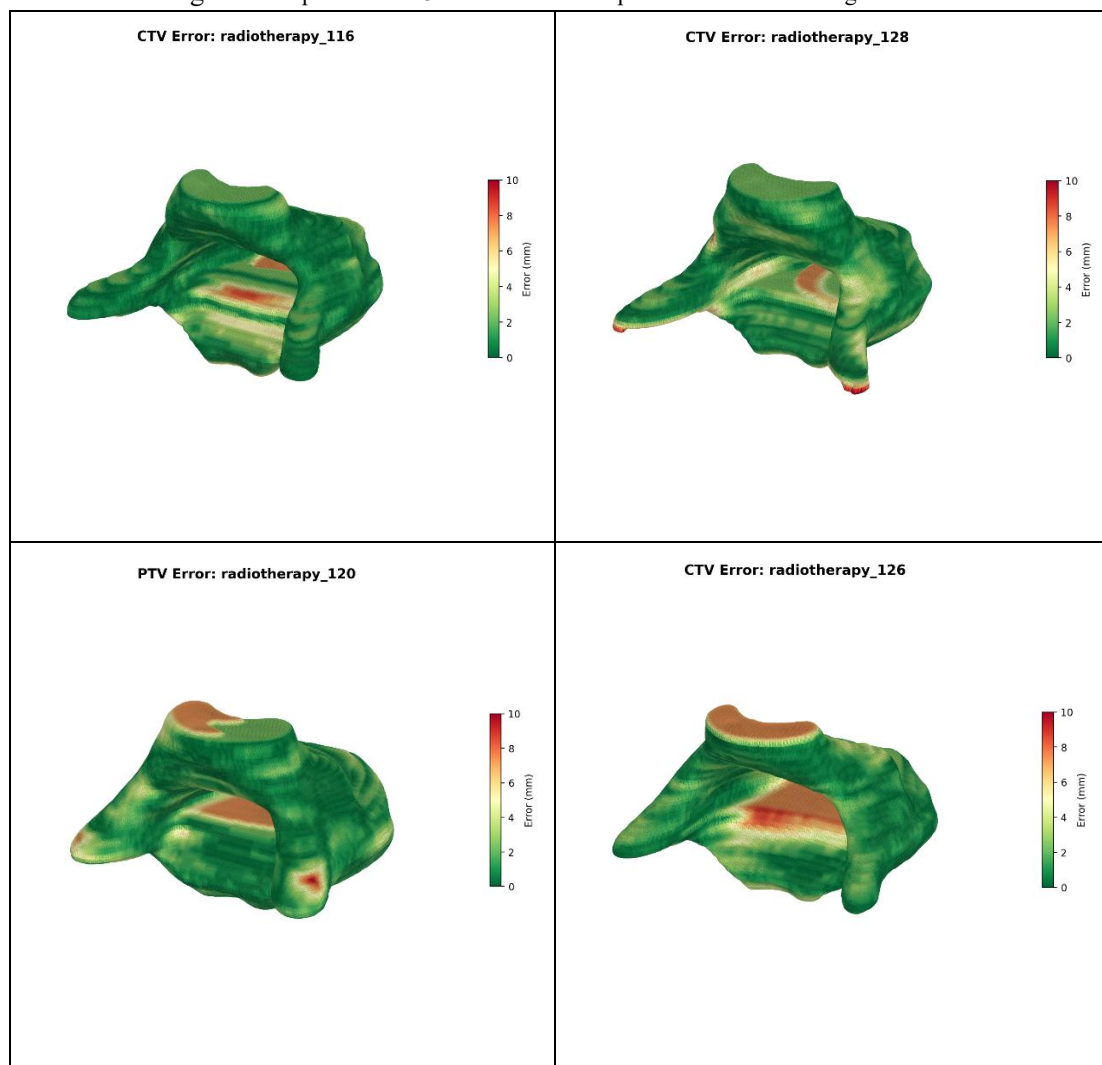


Figure 4. Representative 3D surface-error maps for CTV and PTV segmentation.



The color scale represents local surface distances between the AI-generated and reference contours. Smaller distance values indicate closer surface agreement, whereas larger values indicate greater local deviation. All panels used the same 0–10 mm color scale. The maps represent selected cases and were used for qualitative visualization

rather than formal regional quantitative analysis.

3.6 Multi-Reader Clinical Acceptability Assessment

Three radiation oncologists independently evaluated 27 CTV and 27 PTV auto-contours, resulting in 81 ratings per target and 162 target-level ratings in total. The distributions of the Likert scores, rating-level clinical acceptability, inter-rater agreement, and case-level acceptability are summarized in Table 5.

For CTV, 12 ratings (14.8%) received a score of 4, indicating clinical acceptability without modification; 49 ratings (60.5%) received a score of 3, indicating acceptability after minor editing; and 20 ratings (24.7%) received a score of 2, indicating that major revision was required. No CTV rating received a score of 1. The mean CTV score was 2.90 ± 0.62 . Overall, 61 of 81 CTV ratings were classified as clinically acceptable, corresponding to a rating-level acceptability rate of 75.3% (95% CI, 64.9%–83.4%).

For PTV, 23 ratings (28.4%) received a score of 4, 51 ratings (63.0%) received a score of 3, and 7 ratings (8.6%) received a score of 2. No PTV rating received a score of 1. The mean PTV score was 3.20 ± 0.58 . Overall, 74 of 81 PTV ratings were classified as clinically acceptable, corresponding to a rating-level acceptability rate of 91.4% (95% CI, 83.2%–95.8%).

For the four-point ordinal ratings, the single-reader absolute-agreement ICC was 0.565 for CTV (95% CI, 0.385–0.697) and 0.749 for PTV (95% CI, 0.543–0.892). The reliability of the mean rating across the three readers, expressed as ICC(2,k), was 0.796 for CTV and 0.899 for PTV.

After dichotomizing the ratings into clinically acceptable and unacceptable categories, Fleiss' kappa was 0.535 for CTV (95% CI, 0.221–0.786) and 0.687 for PTV (95% CI, 0.445–1.000). At the case level, at least two of the three readers considered the contour clinically acceptable in 20 of 27 CTV cases (74.1%; 95% CI, 55.3%–86.8%) and 24 of 27 PTV cases (88.9%; 95% CI, 71.9%–96.1%). Cases with discordant assessments of clinical acceptability were subsequently reviewed by the most senior radiation oncologist for final clinical adjudication. The adjudicated decision was used for clinical interpretation only, whereas all original independent reader scores were retained for the rating-level and inter-rater agreement analyses.

Table 5. Multi-reader clinical acceptability and inter-rater agreement for AI-generated contours.

Panel A. Rating distribution and rating-level clinical acceptability

Target	Score 4, n (%)	Score 3, n (%)	Score 2, n (%)	Score 1, n (%)	Overall score, mean $\pm$ SD	Acceptable ratings, n/N (%)	95% CI
CTV	12 (14.8)	49 (60.5)	20 (24.7)	0(0.0)	2.90 ± 0.62	61/81 (75.3)	64.9–83.4
PTV	23 (28.4)	51 (63.0)	7 (8.6)	0(0.0)	3.20 ± 0.58	74/81 (91.4)	83.2–95.8

Panel B. Inter-rater agreement and case-level clinical acceptability

Target	ICC(2,1)	95% CI	ICC(2,k)	Fleiss' κ	95% CI	Majority-acceptable cases, n/N (%)	95% CI
CTV	0.565	0.385–0.697	0.796	0.535	0.221–0.786	20/27 (74.1)	55.3–86.8
PTV	0.749	0.543–0.892	0.899	0.687	0.445–1.000	24/27 (88.9)	71.9–96.1

(Table note: Scores of 3 or 4 were considered clinically acceptable. Rating-level acceptability was calculated from 81 ratings per target. Case-level clinical acceptability was defined as agreement by at least two of the three readers that the contour was clinically acceptable.)

3.7 Workflow Efficiency

Workflow efficiency was evaluated in the separate cohort of 20 patients. Manual contouring required a mean total task time of 65.4 ± 12.3 minutes per case, whereas AI-assisted contouring required 14.2 ± 4.5 minutes per case. The observed mean task time was therefore 51.2 minutes shorter with AI assistance, corresponding to a relative reduction of 78.3% ($P < 0.001$).

The recorded total task time included image loading, AI-contour import when applicable, preliminary image review, contour creation or editing, final contour review, and saving. Both manual and AI-assisted contouring were completed by the same radiation oncologist. For all patients, AI-assisted contouring was completed before manual contouring, and no formal washout period was implemented between the two conditions.

Table 6. Paired comparison of manual and AI-assisted contouring time.

Contouring method	Patients, n	Total task time, mean $\pm$ SD, min	Observed mean reduction, min	Relative reduction, %	P-value
Manual contouring	20	65.4 ± 12.3	Reference	Reference	-
AI-assisted contouring	20	14.2 ± 4.5	51.2	78.3	< 0.001

(**Table note:** Total task time included image loading, AI-contour import when applicable, preliminary image review, contour creation or editing, final review, and saving. Both conditions were completed by the same physician. AI-assisted contouring was performed before manual contouring for all patients. The comparison was performed using a paired-samples t-test. SD, standard deviation.)

4. Discussion

In this single-center retrospective study, we developed a multi-label 3D-fullres nnU-Net model for the simultaneous segmentation of CTV and PTV on NCCT in patients with uterine malignancies. In the 27-patient independent internal test cohort, the model achieved mean DSC values of 90.02% for PTV and 72.16% for CTV. Rating-level clinical acceptability was 91.4% for PTV and 75.3% for CTV. In the separate 20-patient workflow cohort, the mean DSC values were 80.55% for PTV and 68.54% for CTV, while AI-assisted contouring reduced the mean total task time from 65.4 to 14.2 minutes per case. These findings indicate promising internal performance and substantial workflow-time savings, while also demonstrating that CTV auto-contours still require careful physician review.

Previous studies have shown the feasibility of deep learning-based target segmentation for cervical and other gynecologic malignancies [22–29]. The present study extends this work by combining simultaneous CTV/PTV prediction on NCCT with independent internal testing, spatial error visualization, multi-reader clinical assessment, and workflow evaluation. This multidimensional evaluation is important because geometric overlap alone does not fully reflect clinical usability or implementation risk [20,21,34–40]. Nevertheless, the current findings should be interpreted as evidence of internal feasibility rather than external generalizability or readiness for routine clinical use.

The difference between CTV and PTV performance requires careful interpretation. CTV delineation depends on clinical judgment regarding disease extent, nodal drainage regions, postoperative anatomical changes, and low-contrast soft-tissue boundaries. On NCCT, the interfaces among the uterus, cervix, vaginal fornix, parametrial tissues, rectum, and adjacent structures may be difficult to distinguish. Previous studies have similarly reported that automated CTV segmentation in gynecologic radiotherapy remains challenging [22–25]. An additional consideration is that the reference CTV contours in the present study were delineated on NCCT with access to diagnostic pelvic MRI, whereas the segmentation model received NCCT alone as input. The MRI-

informed reference contours therefore incorporated anatomical information that was not directly available to the network. This information asymmetry may partly explain the moderate CTV overlap and highlights the intrinsic difficulty of approximating clinically defined, MRI-informed boundaries using NCCT alone. Accordingly, the mean CTV DSC of 72.16% should be interpreted as moderate agreement with an MRI-informed clinical reference standard and does not support autonomous contouring.

In contrast, the reference PTV was generated using a uniform 5-mm expansion of the approved CTV and therefore had a larger volume and smoother geometry. Although the model directly predicted PTV as a separate label rather than deriving it from the predicted CTV, these geometric characteristics likely contributed to its higher DSC. In addition, DSC is less sensitive to the same absolute boundary error in larger structures. Thus, high PTV overlap should not be interpreted as evidence that the underlying clinically defined CTV boundaries were equally reliable.

CTV volume was positively associated with DSC, whereas no significant association was observed for PTV. This finding may partly reflect the volume dependence of overlap-based metrics, because the same absolute boundary deviation produces a smaller proportional reduction in DSC for a larger structure [20,21]. Therefore, the observed correlation does not necessarily mean that larger CTVs are anatomically easier to delineate. It instead highlights the need to interpret DSC together with boundary-distance metrics, spatial error maps, and physician assessment.

Segmentation performance was numerically lower in the separate workflow cohort than in the independent test cohort, particularly for PTV. Although no formal statistical comparison between the cohorts was planned, this difference suggests that model performance may vary across subsequent clinical cases because of differences in anatomy, target extent, treatment setting, or other unmeasured factors. The markedly right-skewed HD95 distributions also showed that a limited number of cases had relatively large local boundary errors. This explains why the mean HD95 values were considerably higher than the corresponding medians and supports reporting both median and mean HD95 values.

Representative surface-error maps suggested that larger local deviations occurred near cranio-caudal target boundaries and low-contrast anatomical interfaces, including the uterine fundus, superior nodal extent, vaginal fornix, and rectal interface. However, these maps were based on representative cases and were intended for qualitative visualization. They do not establish a systematic regional error pattern across the entire cohort. Future studies should quantify regional over-contouring, under-contouring, and surface-distance errors in anatomically defined subregions. MRI–CT fusion or multimodal learning may also improve the recognition of complex soft-tissue boundaries [30–33].

The multi-reader assessment provided important clinical information beyond geometric metrics. For CTV, 75.3% of ratings were clinically acceptable, whereas 24.7% indicated that major revision was required. The ICC (2,1) of 0.565 and Fleiss' kappa of 0.535 indicated only moderate reader agreement. Two of the three readers had participated in generation or review of the reference contours, and the most senior adjudicating physician had also reviewed the reference contours. Familiarity with the institutional contouring practice and reference-contour generation process may therefore have influenced the perceived acceptability of the AI-generated contours. Accordingly, the reader-assessment results should be interpreted as an internal clinical evaluation rather than a fully independent external assessment. These findings suggest that CTV predictions may be useful as editable initial contours but should not be considered finalized treatment structures. PTV showed

higher acceptability and agreement, with a rating-level acceptability rate of 91.4%, an ICC (2,1) of 0.749, and a Fleiss' kappa of 0.687. Nevertheless, some PTV ratings still required major revision. The absence of a score of 1 reflects perceived contour acceptability only and should not be interpreted as evidence of clinical or dosimetric safety. Mandatory physician review and structured quality assurance remain necessary [38–40].

AI-assisted contouring reduced the observed total task time by 78.3%. The timing included image loading, contour import, preliminary review, contour creation or editing, final review, and saving, thereby reflecting the overall contouring workflow rather than editing time alone. This result is consistent with previous studies showing that auto-segmentation may reduce radiotherapy contouring time [34–37]. However, the workflow experiment included only one physician and 20 patients. The physician who performed the workflow evaluation also participated as one of the three clinical readers, which may have increased familiarity with the model-generated contours. In addition, all AI-assisted tasks were performed before manual contouring, the task order was not randomized, and no formal washout period was implemented. The two conditions were also timed by different individuals. These factors may have introduced familiarity, order, recall, learning, or recording effects. Therefore, the observed 78.3% reduction should be considered a preliminary single-user estimate rather than a definitive measure of clinical efficiency.

The present study did not evaluate the dosimetric consequences of AI-generated contours. Geometric similarity and physician acceptability cannot establish whether local contour differences affect target coverage, dose conformity, or organ-at-risk exposure [21,34,35,40]. The current findings should therefore not be interpreted as evidence of dosimetric equivalence or treatment-planning safety. Future studies should compare CTV/PTV D98%, D95%, V95%, conformity and homogeneity indices, and relevant organ-at-risk dose parameters using plans based on reference and AI-assisted contours.

Several additional limitations should be acknowledged. First, this was a single-center retrospective study using one CT scanner, one acquisition protocol, and one institutional contouring practice, which limits generalizability. Second, the independent test and workflow cohorts included only 27 and 20 patients, respectively, and the FIGO subgroup analyses were underpowered. In addition, the independent test cohort consisted entirely of patients receiving postoperative adjuvant radiotherapy, whereas only five definitive-radiotherapy cases were included in the development cohort. Consequently, the independent test results primarily reflect model performance in the postoperative setting and should not be extrapolated to intact cervical cancer treated with definitive radiotherapy; dedicated validation in a sufficiently sized definitive-radiotherapy cohort is required. Third, no external validation was performed. Fourth, formal inter-observer variability in the reference contours was not quantified because only the final approved contours were retained. The reference contours were also informed by diagnostic pelvic MRI, whereas the segmentation model was trained and tested using NCCT alone, introducing an information asymmetry that may limit reproducibility in settings where MRI availability or MRI-informed contouring practices differ. Finally, there was personnel overlap across the reference-contour generation, multi-reader assessment, adjudication, and workflow evaluation components. Two readers had participated in reference-contour generation or review, and the workflow physician also served as one of the clinical readers. This overlap may have introduced familiarity or observer-related bias in the acceptability ratings and workflow results. The workflow evaluation additionally involved a single physician, a fixed AI-first task order, no formal washout period, and different time recorders for the

two conditions. Multi-center external validation, dedicated definitive-radiotherapy validation, prospective multi-user workflow studies, reference-contour variability assessment, regional error analysis, and dosimetric evaluation are required before routine clinical use.

Overall, the multi-label 3D-fullres nnU-Net demonstrated high internal PTV overlap, moderate CTV overlap, and substantial observed workflow-time reduction on NCCT. The model may serve as an investigational tool for generating editable initial contours, but it should not be regarded as an autonomous segmentation system. CTV predictions require careful physician review, and external and dosimetric validation remain necessary before broader clinical implementation.

5. Conclusion

The multi-label 3D-fullres nnU-Net achieved mean DSC values of 90.02% for PTV and 72.16% for CTV and reduced contouring time by 78.3%. The model may provide editable initial contours, but CTV predictions require careful physician review. External and dosimetric validation are needed before routine clinical use.

Declarations

Abbreviations: 2D: two-dimensional; 3D: three-dimensional; AI: artificial intelligence; CI: confidence interval; CT: computed tomography; CTV: clinical target volume; DL: deep learning; DSC: Dice similarity coefficient; ESTRO: European Society for Radiotherapy and Oncology; FIGO: International Federation of Gynecology and Obstetrics; HD95: 95th percentile Hausdorff distance; HU: Hounsfield unit; ICC: intraclass correlation coefficient; IMRT: intensity-modulated radiotherapy; IQR: interquartile range; IRB: Institutional Review Board; MRI: magnetic resonance imaging; NCCT: non-contrast computed tomography; OAR: organ at risk; PTV: planning target volume; RTOG: Radiation Therapy Oncology Group; RT: radiotherapy; SD: standard deviation; TPS: treatment planning system; VMAT: volumetric modulated arc therapy.

Ethics approval and consent to participate: This single-center retrospective study was approved by the Institutional Review Board of Ganzhou Cancer Hospital (No. (2025) Ke-Lun-Shen (89)). The requirement for informed consent was waived by the Institutional Review Board because of the retrospective nature of the study. This study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent for publication: Not Applicable.

Code availability: The code used for data preprocessing, nnU-Net training and inference, segmentation-metric calculation, statistical analysis, and figure generation is publicly available in the Zenodo repository at <https://doi.org/10.5281/zenodo.21448582>

Availability of data and materials: The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Competing Interests: The authors declare that they have no competing interests.

Funding: This work was supported by the Ganzhou Science and Technology Planning Project (2025YLCE0174). The funding body had no role in the study design, data collection, data analysis, data interpretation, manuscript preparation, or decision to submit the manuscript for publication.

Authors' contributions: Z.W.L. and Y.W.Z. contributed equally to this work. Z.W.L., W.W.W., and X.Y.L. conceived and designed the study. Z.W.L. and Y.W.Z. collected and organized the clinical imaging data. Z.W.L. performed model development, training, validation, and statistical analysis. Y.W.Z. and L.T. assisted with data preprocessing, contour review, and clinical interpretation. W.W.W. and X.Y.L. supervised target volume delineation, reviewed the reference contours, and provided clinical oversight. Z.W.L. drafted the manuscript. Y.W.Z., L.T., W.W.W., and X.Y.L. critically revised the manuscript for important intellectual content. W.W.W. and X.Y.L. supervised the study. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

References:

- [1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021; 71:209-249.
- [2] Cibula D, Raspollini MR, Planchamp F, Centeno C, Chargari C, Felix A, et al. ESGO/ESTRO/ESP guidelines for the management of patients with cervical cancer - update 2023. *Int J Gynecol Cancer.* 2023; 33:649-666.
- [3] Concin N, Matias-Guiu X, Cibula D, Colombo N, Creutzberg CL, Ledermann J, et al. ESGO-ESTRO-ESP guidelines for the management of patients with endometrial carcinoma: update 2025. *Lancet Oncol.* 2025; 26: e423-e435.
- [4] Small W Jr, Bosch WR, Harkenrider MM, Strauss JB, Abu-Rustum N, Albuquerque KV, et al. NRG Oncology/RTOG consensus guidelines for delineation of clinical target volume for intensity modulated pelvic radiation therapy in postoperative treatment of endometrial and cervical cancer: an update. *Int J Radiat Oncol Biol Phys.* 2021; 109:413-424.
- [5] Fields EC, Bosch WR, Albuquerque KV, Bhatia R, Chino J, Dyer B, et al. Consensus guidelines for delineation of clinical target volumes for intensity modulated radiation therapy for intact cervical cancer: an update. *Pract Radiat Oncol.* 2025; 15:171-179.
- [6] Klopp AH, Yeung AR, Deshmukh S, Gil KM, Wenzel L, Westin SN, et al. Patient-reported toxicity during pelvic intensity-modulated radiation therapy: NRG Oncology-RTOG 1203. *J Clin Oncol.* 2018; 36:2538-2544.
- [7] Yeung AR, Deshmukh S, Klopp AH, Gil KM, Wenzel L, Westin SN, et al. Intensity-modulated radiation therapy reduces patient-reported chronic toxicity compared with conventional pelvic radiation therapy: updated results of a phase III trial. *J Clin Oncol.* 2022; 40:3115-3119.
- [8] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys.* 2014; 41:050902.
- [9] Das IJ, Compton JJ, Bajaj A, Johnstone PA. Intra- and inter-physician variability in target volume delineation in radiation therapy. *J Radiat Res.* 2021; 62:1083-1089.
- [10] Elmali A, Biltekin F, Sari SY, Gultekin M, Yuce D, Yildiz F. Inter-observer variation of target volume delineation for CT-guided cervical cancer brachytherapy. *J Contemp Brachytherapy.* 2023; 15:253-260.
- [11] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018; 18:500-510.
- [12] Samarasinghe G, Jameson M, Vinod S, Field M, Dowling J, Sowmya A, et al. Deep learning for segmentation in radiation therapy planning: a review. *J Med Imaging Radiat Oncol.* 2021; 65:578-595.
- [13] Bibault JE, Giraud P. Deep learning for automated segmentation in radiotherapy: a narrative review. *Br J Radiol.* 2024; 97:13-20.
- [14] Matoska T, Patel M, Liu H, Beriwal S. Review of deep learning based autosegmentation for clinical target volume: current status and future directions. *Adv Radiat Oncol.* 2024; 9:101470.
- [15] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image*

- Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer; 2015. p. 234-241.
- [16]Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer; 2016. p. 424-432.
 - [17]Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision. Stanford: IEEE; 2016. p. 565-571.
 - [18]Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021; 18:203-211.
 - [19]Ma J, Zhang Y, Gu S, Zhu C, Ge C, Zhang Y, et al. AbdomenCT-1K: is abdominal organ segmentation a solved problem? *IEEE Trans Pattern Anal Mach Intell*. 2022; 44:6695-6714.
 - [20]Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015; 15:29.
 - [21]Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. *Nat Methods*. 2024; 21:195-212.
 - [22]Liu Z, Liu X, Guan H, Zhen H, Sun Y, Chen Q, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiother Oncol*. 2020;153:172-179.
 - [23]Shi J, Ding X, Liu X, Li Y, Liang W, Wu J. Automatic clinical target volume delineation for cervical cancer in CT images using deep learning. *Med Phys*. 2021; 48:3968-3981.
 - [24]Ma CY, Zhou JY, Xu XT, Guo J, Han MF, Gao YZ, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. *J Appl Clin Med Phys*. 2022;23: e13470.
 - [25]Ma C, Zhou J, Xu X, Guo J, Han M, Gao Y, et al. Clinical evaluation of deep learning-based clinical target volume three-channel auto-segmentation algorithm for adaptive radiotherapy in cervical cancer. *BMC Med Imaging*. 2022; 22:123.
 - [26]Tian M, Zhang X, Liu X, Wang H, Li X, Li J, et al. Delineation of clinical target volume and organs at risk in cervical cancer radiotherapy using deep learning. *Med Phys*. 2023;50: 6354–6365.
 - [27]Wu Z, Wang D, Xu C, Peng S, Deng L, Liu M, et al. Clinical target volume automatic delineation using deep learning network for cervical cancer radiotherapy: a study with external validation. *J Appl Clin Med Phys*. 2025;26: e14553.
 - [28]Peng H, Liu T, Li P, Yang F, Luo X, Sun X, et al. Automatic delineation of cervical cancer target volumes in small samples based on multi-decoder and semi-supervised learning and clinical application. *Sci Rep*. 2024; 14:26937.
 - [29]Yan J, Qin X, Qiao C, Zhu J, Song L, Yang M, et al. Auto-segmentation of the clinical target volume using a domain-adversarial neural network in patients with gynaecological cancer undergoing postoperative vaginal brachytherapy. *Precis Radiat Oncol*. 2023;7:189-196.
 - [30]Liugang G, Kai X, Chunying L, Zhengda L, Jianfeng S, Tao L, et al. Generation of virtual non-contrast CT from intravenous enhanced CT in radiotherapy using convolutional neural networks. *Front Oncol*. 2020; 10:1715.

- [31]Dimopoulos JC, Petrow P, Tanderup K, Petric P, Berger D, Kirisits C, et al. Recommendations from Gynaecological GEC-ESTRO Working Group IV: basic principles and parameters for MR imaging within the frame of image based adaptive cervix cancer brachytherapy. *Radiother Oncol.* 2012; 103:113-122.
- [32]Haie-Meder C, Pötter R, Van Limbergen E, Briot E, De Brabandere M, Dimopoulos J, et al. Recommendations from Gynaecological GEC-ESTRO Working Group I: concepts and terms in 3D image based 3D treatment planning in cervix cancer brachytherapy. *Radiother Oncol.* 2005; 74:235-245.
- [33]Viswanathan AN, Erickson BA. Three-dimensional imaging in gynecologic brachytherapy: a survey of the American Brachytherapy Society. *Int J Radiat Oncol Biol Phys.* 2010; 76:104-109.
- [34]Wong J, Huang V, Wells D, Kolbeck C, Giambattista J, Otto K, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiat Oncol.* 2021; 16:101.
- [35]Doolan PJ, Charalambous S, Roussakis Y, Leczynski A, Peratikou M, Benjamin M, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Front Oncol.* 2023; 13:1213068.
- [36]Heilemann G, Buschmann M, Lechner W, Dick V, Eckert F, Heilmann M, et al. Clinical implementation and evaluation of auto-segmentation tools for multi-site contouring in radiotherapy. *Phys Imaging Radiat Oncol.* 2023; 28:100515.
- [37]Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol.* 2020; 13:1-6.
- [38]Rusanov B, Ebert MA, Sabet M, Rowshanfarzad P, Barry N, Kendrick J, et al. Guidance on selecting and evaluating AI auto-segmentation systems in clinical radiotherapy: insights from a six-vendor analysis. *Phys Eng Sci Med.* 2025; 48:301-316.
- [39]De Biase A, Sijtsema NM, Janssen T, Hurkmans C, Brouwer C, van Ooijen P. Clinical adoption of deep learning target auto-segmentation for radiotherapy: challenges, clinical risks, and mitigation strategies. *BJR Artif Intell.* 2024;1(1): ubae015.
- [40]Mahmood U, Shukla-Dave A, Chan HP, Drukker K, Samala RK, Chen Q, et al. Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. *BJR Artif Intell.* 2024;1(1): ubae003.