Original Article

CT-based deep learning segmentation of ovarian cancer and the stability of the extracted radiomics features

Yiang Wang[1]^, Mandi Wang[2]^, Peng Cao[1]^, Esther M. F. Wong[3], Grace Ho[4], Tina P. W. Lam[4], Lujun Han[5], Elaine Y. P. Lee[1]^

[1]Department of Diagnostic Radiology, The University of Hong Kong, Hong Kong, China; [2]Department of Radiology, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, China; [3]Department of Radiology, Pamela Youde Nethersole Eastern Hospital, Hong Kong, China; [4]Department of Radiology, Queen Mary Hospital, Hong Kong, China; [5]Department of Medical Imaging, State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Sun Yat-sen University Cancer Center, Guangzhou, China

Contributions: (I) Conception and design: Y Wang, P Cao, EYP Lee; (II) Administrative support: EYP Lee; (III) Provision of study materials or patients: M Wang, L Han, EMF Wong, G Ho, TPW Lam, EYP Lee; (IV) Collection and assembly of data: Y Wang, M Wang; (V) Data analysis and interpretation: Y Wang, M Wang, EYP Lee; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Dr. Lujun Han. Department of Medical Imaging, State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Sun Yat-sen University Cancer Center, Guangzhou 510060, China. Email: hanlj@sysucc.org.cn; Dr. Elaine Y. P. Lee. Department of Diagnostic Radiology, Room 406, Block K, Queen Mary Hospital, University of Hong Kong, 102 Pokfulam Road, Hong Kong, China. Email: eyplee77@hku.hk.

Background: Radiomics analysis could provide complementary tissue characterization in ovarian cancer (OC). However, OC segmentation required in radiomics analysis is time-consuming and labour-intensive. In this study, we aim to evaluate the performance of deep learning-based segmentation of OC on contrast-enhanced CT images and the stability of radiomics features extracted from the automated segmentation.

Methods: Staging abdominopelvic CT images of 367 patients with OC were retrospectively recruited. The training and cross-validation sets came from center A (n=283), and testing set (n=84) came from centers B and C. The tumours were manually delineated by a board-certified radiologist. Four model architectures provided by no-new-Net (nnU-Net) method were tested in this task. The segmentation performance evaluated by Dice score, Jaccard score, sensitivity and precision were compared among 4 architectures. The Pearson correlation coefficient (ρ), concordance correlation coefficient (ρ_c) and Bland-Altman plots were used to evaluate the volumetric assessment of OC between manual and automated segmentations. The stability of extracted radiomics features was evaluated by intraclass correlation coefficient (ICC).

Results: The 3D U-Net cascade architecture achieved highest median Dice score, Jaccard score, sensitivity and precision for OC segmentation in the testing set, 0.941, 0.890, 0.973 and 0.925, respectively. Tumour volumes of manual and automated segmentations were highly correlated (ρ=0.944 and ρ_c =0.933). 85.0% of radiomics features had high correlation with ICC >0.8.

Conclusions: The presented deep-learning segmentation could provide highly accurate automated segmentation of OC on CT images with high stability of the extracted radiomics features, showing the potential as a batch-processing segmentation tool.

Keywords: Ovarian cancer (OC); computed tomography; deep learning; automated segmentation; radiomics

^ ORCID: Yiang Wang, 0000-0003-2656-1511; Mandi Wang, 0000-0002-8467-148X; Peng Cao, 0000-0002-1116-1171; Elaine Y. P. Lee, 0000-0002-0627-5297.

Submitted Oct 23, 2022. Accepted for publication May 22, 2023. Published online Jun 13, 2023.
doi: 10.21037/qims-22-1135
View this article at: https://dx.doi.org/10.21037/qims-22-1135

Introduction

Ovarian cancer (OC) is one of the female malignancies with high cancer mortality (1). Computed tomography (CT) is used to evaluate the disease extent of OC at presentation and at disease recurrence (2). Radiomics that enable the extraction of features from digital medical images have potential in tumour characterization through feature-based model building to improve clinical management (3). Study showed that CT radiomic features were helpful in histological classification and predicting response to chemotherapy in OC (4-6). For radiomics study and statistical model building, large amount of data is required. However, identification of the regions of interest (ROIs) and tumour segmentation, key steps of radiomics, are generally performed in a manual or semi-manual way, which can be extremely time-consuming and labour intensive, especially in large dataset. Rizzo *et al.* listed automated segmentation as one challenge of radiomics analysis in a narrative review because of reproducibility problems (7).

Although many automated segmentation methods have been developed to provide fast and accurate results using different imaging modalities, very few focused on OC and based on CT images (8). Ovarian tumours are usually heterogeneous and complex with both solid and cystic components in an anatomical area where soft tissue resolution is limited on CT. There were few studies in tumour segmentation specific to OC, but these were either based on different imaging modality or on metastatic tumours rather than the primary tumour (9-11). Deep learning-based methods have not been explored in segmenting OC on CT images.

Deep learning-based semantic segmentation methods could make full use of huge number of medical images and provide faster and more accurate segmentations, compared with conventional methods (12). U-Net, one of the most impactful deep learning-based architectures in medical image segmentation, has been successfully applied in hundreds of studies with different imaging modalities and clinical applications (13). Most improvements based on classic U-Net structure focused on network architecture variations, such as the usage of attention gate and residual unit (14,15). Different from other variants of U-Net, nnU-Net focused on the configurations of the whole process from pre-processing to post-processing, enabling reproductivity of high performance in new datasets (16). Therefore, nnU-Net became a popular deep learning-based segmentation tool especially for novel clinical applications, such as segmentation of myocardium and infarct zone on magnetic resonance imaging (MRI) images, and thoracic lymph node station on CT, since no prior expert experience on this task is required (17,18).

The clinical value of CT radiomics has been reported and discussed in our previous studies (5,6). However, there is a need to develop an automated segmentation tool to avoid labour-intensive manual delineation, and eventually integrate it into the clinical workflow of radiomics analysis. This study aimed to evaluate the segmentation performance of OC on CT images using nnU-Net method and test the stability of radiomics features extracted from the automated segmentation.

Methods

Image database

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (Reference No. UW 20-251), the Hong Kong East Cluster Research Ethics Committee (Reference No. HKECREC-2020-040), and the Institutional Review Board of the Sun Yat-sen University Cancer Center (Approval No. YB2018-52), and individual consent for this retrospective analysis was waived. The whole dataset consisted of anonymized pretreatment contrast-enhanced CT images (portovenous phase) acquired in axial plane collected from three centers: center A (n=283, from February 2012 to May 2019), center B (n=42, from February 2009 to November 2017) and center C (n=42, from November 2008 to August 2019). Training and 5-fold cross-validation were performed on the dataset from center A, and the model performance was tested on the combined external cohorts from centers B and C. Consecutive patients included in this study satisfied these criteria: (I) histologically confirmed OC, (II) available pre-treatment contrast-enhanced CT. The exclusion criteria were: (I) CT images with significant artefacts precluding evaluation of the pelvic region, (II) incomplete coverage of

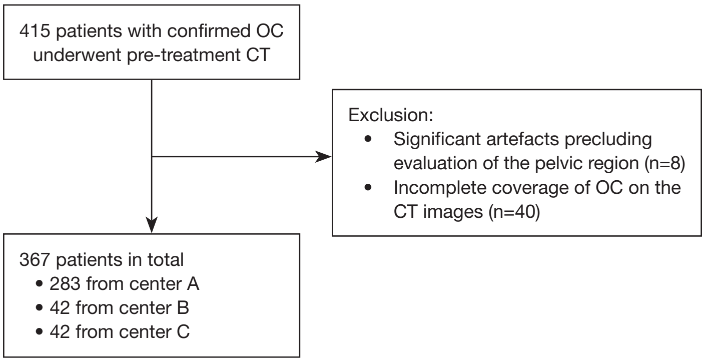
415 patients with confirmed OC underwent pre-treatment CT
Exclusion:
• Significant artefacts precluding evaluation of the pelvic region (n=8)
• Incomplete coverage of OC on the CT images (n=40)
367 patients in total
• 283 from center A
• 42 from center B
• 42 from center C

Figure 1 Flow diagram of patient inclusion and exclusion criteria. OC, ovarian cancer; CT, computed tomography.

Table 1 CT scanning parameters at each center

Dataset	Center A	Center B	Center C
Manufacturers	Toshiba; Philips	Siemens; Toshiba	GE; Toshiba
Tube current (mAs)	200–250	120–180	140–200
Tube voltage (kVp)	120	120	120
Slice thickness (mm)	1.0–5.0	1.25	1.25
Pixel size (mm)	0.578–1.083	0.527–0.703	0.586–0.793
Convolution kernel	FC04	I30f\3; FC10; FC18	Std/FC08
Intravenous contrast material	Ultravist; Iohexol	Iopamiro; Omnipaque	Iohexol

CT, computed tomography

OC on the CT images (*Figure 1*). All patients underwent surgery and the pathological details were recorded. The detailed information on scanner models and scanning parameters are summarized in *Table 1*.

Manual segmentation

All the ROIs (i.e., ground truth masks) were manually delineated on all the axial slices containing the primary ovarian tumours by a board-certified radiologist with more than 15 years of experience in pelvic imaging. Delineation was made around the margin of the primary ovarian tumour to include both cystic and solid components on the axial images of the CT with reference to the reformatted coronal and sagittal planes. Adjacent normal tissues and peritoneal metastases were excluded. The manual delineation was performed using 3D Slicer version 4.11. To test the intra-rater reliability, 50 randomly selected cases from center A were delineated twice at least 4 weeks apart and evaluated by kappa values.

Model architectures

The nnU-Net provides an out-of-the-box pipeline for segmentation tasks with automatically configured image preprocessing and model settings, well-illustrated elsewhere (16). As shown in *Figure 2*, four types of U-Net-based model architectures were used in this work, including two-dimensional (2D) U-Net, three-dimensional (3D) U-Net with low image resolution (3D-lowres), 3D U-Net with full image resolution (3D-fullres), and 3D U-Net cascade (3D-cascade). The architecture configured for our task contained six and five down-sampling operations for 2D and 3D models. Each operation includes two stacked convolutional layers, with each stacked layer consisting of one convolutional layer, followed by one instance normalization layer and the Leaky Rectified Linear Unit activation function. Each convolutional layer had a kernel size of 3×3 pixels for 2D model or 3×3×3 voxels for 3D models. Unlike other three architectures, 3D-cascade includes two training stages, each involving a 3D-lowres

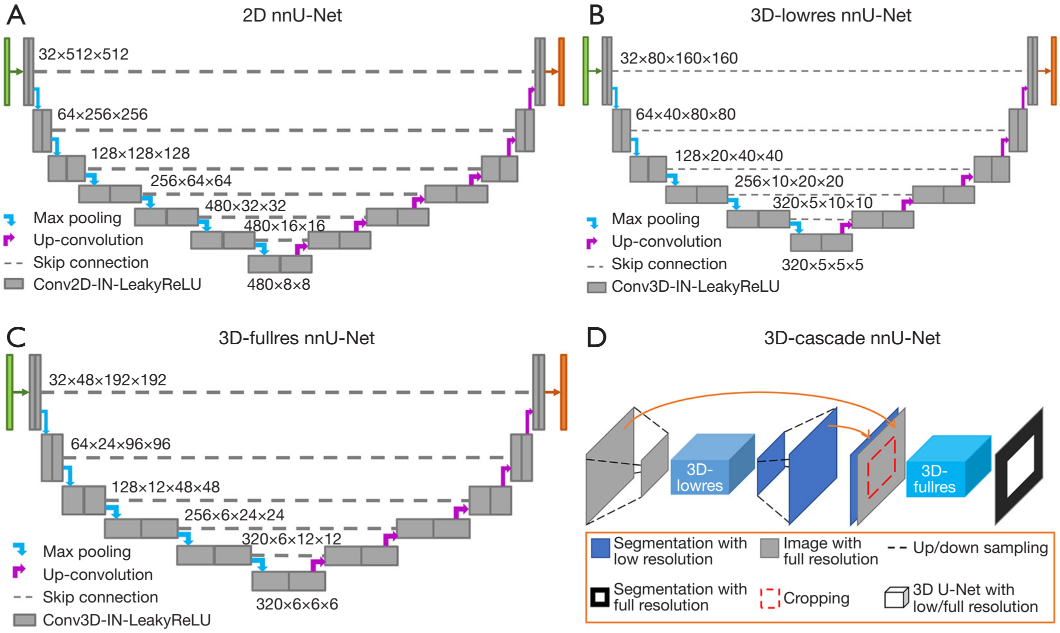

Figure 2 Four model architectures generated by nnU-Net in this study. (A) 2D; (B) 3D-lowres; (C) 3D-fullres; (D) 3D-cascade. Conv2D, 2D convolutional layer; Conv3D, 3D convolutional layer; IN, instance normalization; LeakyReLU, Leaky rectified linear unit.

and 3D-fullres architecture respectively. The second stage can be implemented after the training of all five 3D-lowres models in the first stage have completed. The batch size was 32 for 2D model and 2 for 3D models. The input patch size for 2D, 3D-lowres/3D-cascade, and 3D-fullres architecture was 512×512, 80×160×160, and 48×192×192 respectively.

Image preprocessing

Dataset properties, such as voxel spacings, image shape, modality and intensity, were first automatically collected by the nnU-Net pipeline. These properties were then used to configure resampling and normalization strategy, batch size, patch size etc. As the voxel spacing of CT images from three centers were different and anisotropic (*Table 1*), all CT images and corresponding masks were resampled to the same voxel spacing (1.492×1.492×5 mm for 3D-lowres, and 0.736×0.736×5 mm for 2D and 3D-fullres). Third-order spline interpolation was used for in-plane resampling, and nearest-neighbor interpolation was used for out-of-plane resampling to suppress resampling artefacts. The target sampling pixel spacing automatically selected by the nnU-Net pipeline was the 10th percentile pixel spacing of the training cases for 3D-lowres model, and the median pixel spacing was selected for 2D and 3D-fullres models. The target slice thickness was the largest slice thickness of the training cases. Global intensity normalization was applied to all the CT images before training.

Training details

The loss function was defined as the sum of Dice loss and cross-entropy loss. The optimizer used in this study was stochastic gradient descent with initial learning rate of 0.01 and momentum of 0.99. Data augmentation methods used were rotation, scaling, gamma correction and mirroring. Each model was trained for 1,000 epochs by default to ensure its convergency. For each type of network architectures (i.e., 2D, 3D-lowres, 3D-fullres, and 3D-cascade), the final predicted segmentation was made based on the average result of five models, which were trained using different dataset divisions for cross-validation. One network architecture with the highest average Dice score computed in the cross-validation set was selected as the recommended architecture, and used for volumetric assessment and radiomics stability analysis.

The whole process was performed on an Intel Xeon Gold 5217 central processing unit and a NVIDIA Tesla V100 graphics processing unit card. The program was written in Python 3.7, with PyTorch library version 1.8.1 and nnUNet library version 1.6.6.

Performance evaluation

The evaluation metrics for the segmentation performance includes Dice score, Jaccard score, sensitivity and precision, defined in previous studies (19,20). To further investigate the possible influence of clinicopathological factors on the auto-segmentation performance, the Dice scores in the testing set were compared between two histological subtypes: high grade serous carcinoma (HGSC) and non-HGSC, FIGO (International Federation of Gynecology and Obstetrics) stages and testing centers. For the selected network architecture, we investigated the accuracy of tumour volume assessments in the testing set. The tumour volume was calculated as number of voxels inside the tumour multiplied by the voxel volume based on the resampled images. In total 1,218 radiomics features defined in the Pyradiomics template (https://github.com/AIM-Harvard/pyradiomics/blob/master/examples/exampleSettings/exampleCT.yaml) were extracted with Pyradiomics library version 3.0.1 to keep the repeatability of feature extraction. These radiomics features were extracted from three image types: original, wavelet filter (including 8 combinations of either high or low pass filter in each spatial dimensions) and Laplacian of Gaussian (LoG) filter (sigma =1.0, 2.0, 3.0, 4.0 and 5.0). For the original image type, the radiomics features include First Order, Shape, Gray Level Cooccurence Matrix, Gray Level Run Length Matrix, Gray Level Size Zone Matrix, and Gray Level Dependence Matrix features. For each type of wavelet or LoG filter, all the aforementioned features were included except for Shape-based feature. The intraclass correlation coefficients (ICCs) of radiomics features between the ground truth and predicted segmentation in the testing set were used to evaluate the stability of the extracted radiomics features. An ICC >0.8 was regarded as marker of high stability, as widely defined in other studies (21-23).

Statistical analysis

Evaluation metrics for segmentation performance, age and tumour volume were represented as median (range). The difference was compared using Mann-Whitney U test for two categories, and Kruskal–Wallis test for multiple categories, considering that some subsets did not follow the normal distribution. Pearson correlation coefficient (ρ), concordance correlation coefficient (ρ_c) and Bland-Altman plots were used to evaluate the correlation or agreement of volumes manually delineated by the radiologist as the ground truth and predicted segmentation provided by selected nnU-Net model. All tests with a P value <0.05 were regarded as statistically significant. All the hypothesis tests were performed using Scipy (version 1.6.3), a Python library. The developed codes of OC segmentation can be found in the online repository (https://github.com/HKUCaoLab/segment_OC).

Results

Patient characteristics

The patient's median age was 51 years (range, 18–90 years), and median tumour volume of manual delineations was 375 cm^3 (range, 6–32,435 cm^3) based on the resampled images. The dataset consisted of patients with all stages of OC: FIGO stage I (n=76, 21.1%), stage II (n=43, 11.9%), stage III (n=180, 50.0%) and stage IV (n=61, 16.9%). The majority of cases were HGSC (n=240, 65.4%). The pathological characteristics are summarized in *Table 2*. The mean Kappa value for intra-rater reliability was 0.886±0.104, demonstrating high consistency between the two delineations.

Segmentation performance

The results of OC segmentation performance in the testing set are summarized in *Table 3*. Representative segmentations in the testing set predicted by four models were shown in *Figure 3*. The highest median Dice score, Jaccard score and sensitivity among the testing set were 0.941, 0.890 and 0.973 achieved by 3D-cascade. The highest median precision among the testing set was 0.938 achieved by both 2D and 3D-fullres models. Dice score, Jaccard score, and sensitivity among four nnU-Net architectures were significantly different (P=0.028, P=0.028, and P<0.001, respectively), while precision was not (P=0.800). The Dice score, Jaccard score, sensitivity and precision of 3D-cascade were significantly higher than 2D (P=0.013, P=0.013, P<0.001 and P<0.001 respectively).

There were 4 types of relationship between the ground truth and automated segmentation: good agreement (*Figure 4A*), area underestimated by automated segmentation (*Figure 4B*), area overestimated by automated segmentation (*Figure 4C*) and mixture of both underestimated and overestimated area (*Figure 4D*). It was also noted that

Table 2 Pathological characteristics of patients

Dataset	Center A	Center B	Center C	Overall
Age (years)	52 (18–80)	54 (22–90)	48.5 (36–78)	51 (18–90)
Tumour volume* (cm^3)	360 (6–32,435)	343 (7–3,389)	652 (27–3,660)	375 (6–32,435)
Histological types (n)	283	42	42	367
HGSC	214 (75.6%)	10 (23.8%)	16 (38.1%)	240 (65.4%)
Non-HGSC	69 (24.4%)	32 (76.2%)	26 (61.9%)	127 (34.6%)
FIGO stages (n)	283	41[†]	36[†]	360[†]
I	37 (13.1%)	26 (63.4%)	13 (36.1%)	76 (21.1%)
II	37 (13.1%)	2 (4.9%)	4 (11.1%)	43 (11.9%)
III	155 (54.8%)	11 (26.8%)	14 (38.9%)	180 (50.0%)
IV	54 (19.1%)	2 (4.9%)	5 (13.9%)	61 (16.9%)

Age and tumour volume were presented as median (range), while each histological type or FIGO stage was presented as value (percentage). *, tumour volume was estimated on resampled images; [†], one patient from Center B and 6 patients from Center C had ovarian cystectomy for histological diagnosis but did not proceed to surgical staging. FIGO, International Federation of Gynecology and Obstetrics; HGSC, high-grade serous carcinoma; Non-HGSC, including low-grade serous carcinoma, clear cell carcinoma, endometrioid carcinoma, and mucinous carcinoma.

Table 3 Segmentation performance on external testing set (n=84)

nnU-Net Architectures	Dice score	Jaccard score	Sensitivity	Precision
2D	0.918 (0[†]–0.986)	0.848 (0[†]–0.973)	0.946 (0[†]–0.993)	0.938 (0.012–0.990)[†]
3D-fullres	0.936 (0[†]–0.986)	0.881 (0[†]–0.973)	0.967 (0[†]–0.995)	0.938 (0.263–0.984)[†]
3D-lowres	0.941 (0.352–0.985)	0.888 (0.214–0.971)	0.971 (0.806–0.992)	0.926 (0.216–0.984)
3D-cascade	0.941 (0.371–0.985)	0.890 (0.228–0.971)	0.973 (0.844–0.995)	0.925 (0.231–0.986)
P value	0.028	0.028	<0.001	0.800[†]

Segmentation performance for each architecture was represented as median (range). [†], the 2D and 3D-fullres model failed to segment two cases and one case respectively (i.e., Dice=0). In this case, the precision scores were NaN, and they were excluded from these calculations. 3D-lowres, 3D U-Net with low image resolution; 3D-fullres, 3D U-Net with full image resolution; 3D-cascade, 3D U-Net cascade.

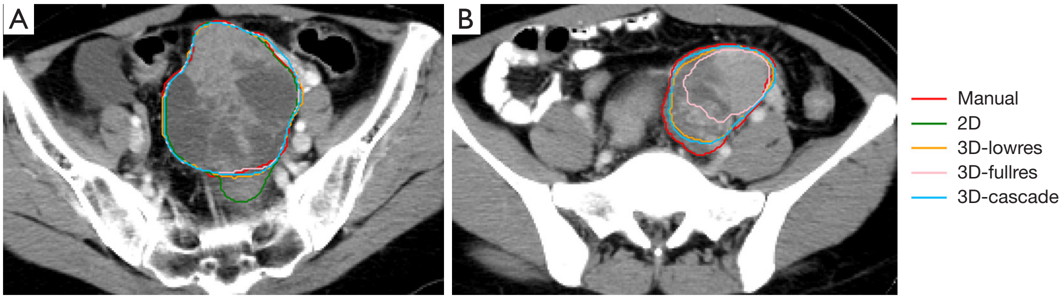

Figure 3 Representative examples of segmentation results predicted by the four models (A,B). The area inside the red ROI represents manually delineated tumour by radiologist as ground truth, and the area inside the green/orange/pink/blue ROI represents the segmentation predicted by 2D/3D-lowres/3D-fullres/3D-cascade model. The 2D model failed to identify the tumour in (B). ROI, region of interest.

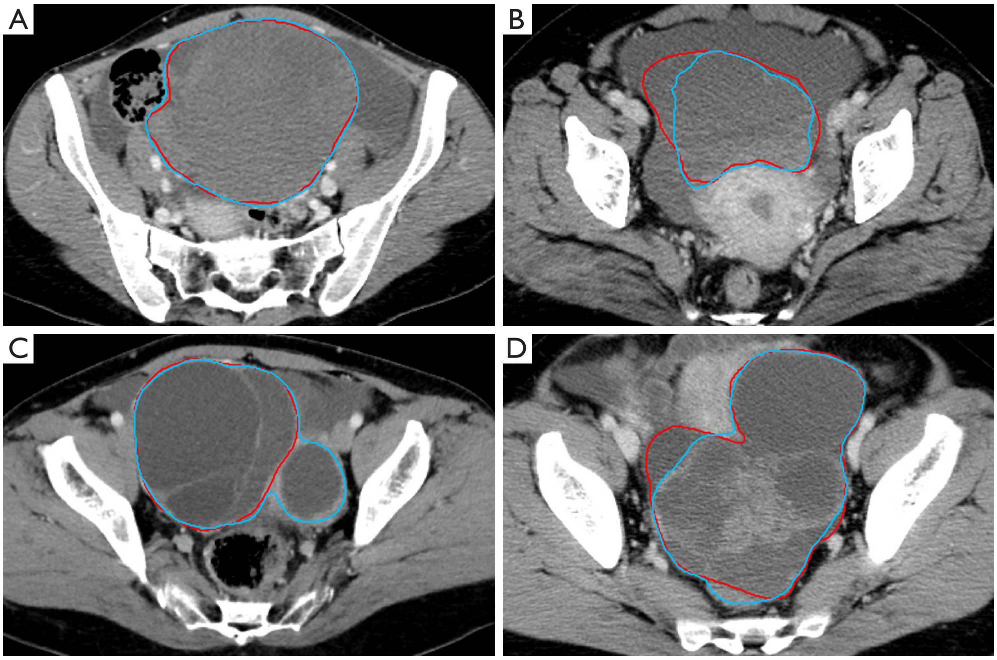

Figure 4 Representative examples of segmentation results predicted by the 3D-cascade model (Dice score for the whole tumour: A. 0.963, B. 0.869, C. 0.930 and D. 0.881). The area inside the red ROI represents manually delineated tumour by radiologist as ground truth, and the area inside the blue ROI represents the predicted segmentation. ROI, region of interest.

physiological corpus luteal cyst, pelvic ascites, pelvis peritoneal and nodal metastases, and the presence of bilateral ovarian masses affected the performance of the automated segmentation resulting in low Dice scores due to incorrect segmentation of these pelvis masses or areas (*Figure 5*).

3D-cascade was the selected network architecture with the highest average Dice score in the cross-validation set among the 4 tested nnU-Net architectures. The differences of Dice scores between histological types, FIGO stages, validation and testing sets, as well as two external testing sets for the 3D-cascade model are presented in *Table 4*. The Dice score of HGSC was significantly lower than non-HGSC (HGSC: Dice =0.919; non-HGSC: Dice =0.958; P<0.001), and the Dice score of cross-validation set was significantly lower than testing set (validation: Dice =0.892; testing: Dice =0.941; P<0.001).

Volumetric assessment

The 3D-cascade nnU-Net model underestimated the tumour volume by 75.29 cm3 on average (*Figure 6A*). High correlation and concordance were found between manually segmented volumes and predicted tumour volumes (ρ=0.944 and ρ_c =0.933, *Figure 6B*). *Figure 5B* shows a representative slice of the outlier identified in *Figure 6B*.

Radiomics features stability

Most of the radiomics features extracted from segmentation predicted by the 3D-cascade model were stable compared with radiomics features extracted from manual delineation, in that 85.0% of the radiomics features had high correlation with ICCs >0.8 (*Figure 7*).

Discussion

In this study, deep learning-based models (i.e., nnU-Net) were used to segment primary OC on contrast-enhanced CT images. Among the tested nnU-Net architectures, the 3D-cascade performed best in the testing set with the highest median Dice score of 0.941. In addition, the automated segmented volume was highly concordant to the manually segmented volume by radiologist with high stability of the extracted radiomics features.

There was no significant difference in the Dice scores of 3D-cascade model between the two external testing sets and FIGO stages, which implied the generalizability of

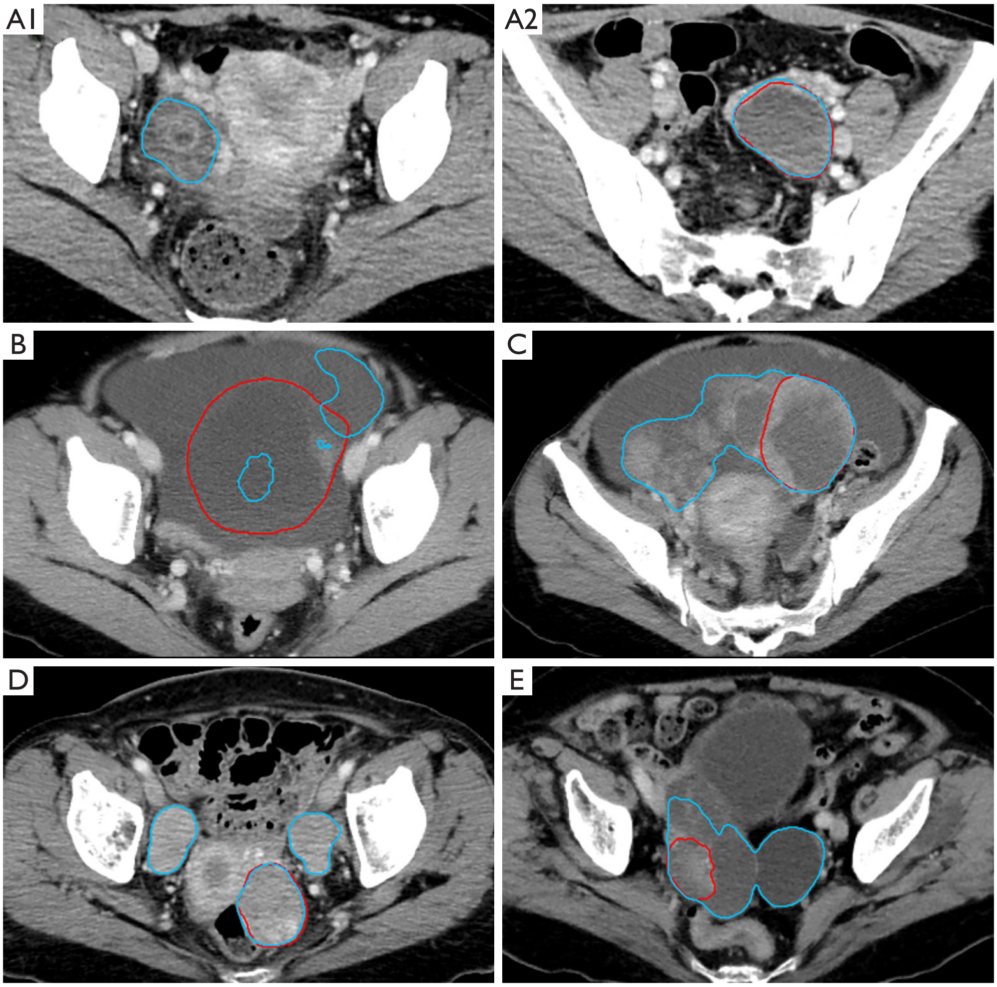

Figure 5 Low Dice scores predicted by the 3D-cascade model due to incorrect segmentation of the right corpus luteal cyst, Dice score 0.760 (A1, A2); ascites, Dice score 0.498 (B); peritoneal metastasis, Dice score 0.501 (C); enlarged pelvic sidewall lymph nodes, Dice score 0.584 (D); contralateral ovarian mass with bilateral involvement, Dice score 0.371 (E). The area inside the red ROI represents manually delineated tumour by radiologist as ground truth, and the area inside the blue ROI represents the predicted segmentation. ROI, region of interest.

our model in different external datasets and FIGO stages. In contrast, the median Dice score of HGSC was found significantly lower than non-HGSC. This might result from the differences in the complexity and heterogeneity of HGSC and non-HGSC tumours, and the different distribution of HGSC and non-HGSC tumours in the training and testing sets. With the latter, the lower Dice score for HGSC could lead to a lower average Dice score for cross-validation results than testing results. Similarly, the imbalanced distribution of each FIGO stage could result in such a significant difference. We speculate that the higher Dice score on testing set might also come from higher proportion of FIGO stage I–II, as higher Dice score was found for FIGO stage I–II (*Table 4*).

Strong correlation and high concordance were found between manually segmented volumes as ground truth and predicted tumour volumes. The change of difference in volume was not observed when the average volume of ground truth and predicted segmentation changed, implying that the performance of volume assessment was independent of tumour volume. However, our model wrongly segmented tissues with similar density to that of OC, for example in physiological corpus luteal cyst that gave rise to more complex appearance with central rim enhancement on CT. The presence of pelvic ascites adjacent to a relatively less complex ovarian tumour could

Table 4 Dice scores of different histological types, FIGO stages, and two testing sets for the 3D-cascade model

	Subgroups	Dice score	P value
Histological types	HGSC	0.919 (0.352–0.985)	<0.001
	Non-HGSC	0.958 (0.812–0.985)	
FIGO stages	I	0.962 (0.352–0.982)	0.277
	II	0.947 (0.897–0.978)	
	III	0.922 (0.496–0.985)	
	IV	0.940 (0.914–0.981)	
Datasets	Validation (from Center A)	0.892 (0–0.982)	<0.001
	Testing (Center B and C)	0.941 (0.371–0.985)	
Testing sets	Center B	0.940 (0.496–0.985)	0.639
	Center C	0.950 (0.352–0.985)	

Dice scores were represented as median (range). 3D-cascade, 3D U-Net cascade; FIGO, International Federation of Gynecology and Obstetrics; HGSC, high-grade serous carcinoma; Non-HGSC, including low-grade serous carcinoma, clear cell carcinoma, endometrioid carcinoma, and mucinous carcinoma.

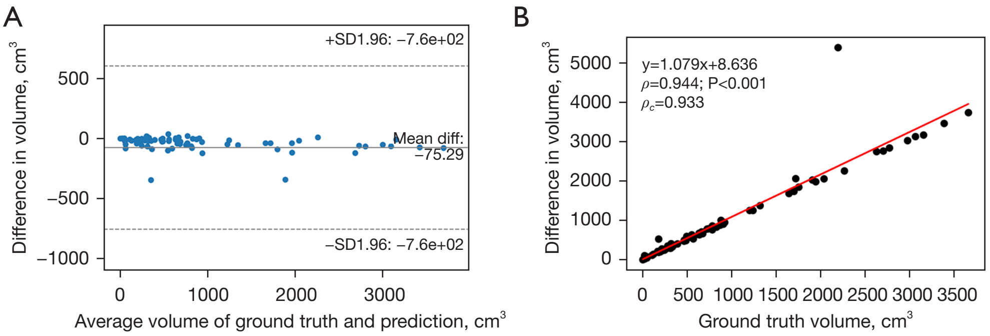

Figure 6 Bland-Altman plot (A), and correlation/concordance analysis (B) for the tumour volumes manually segmented by a radiologist (ground truth), and automatically segmented by the best-performed 3D-cascade model in the testing set. SD, standard deviation; ρ, Pearson correlation coefficient. ρ_c, concordance correlation coefficient.

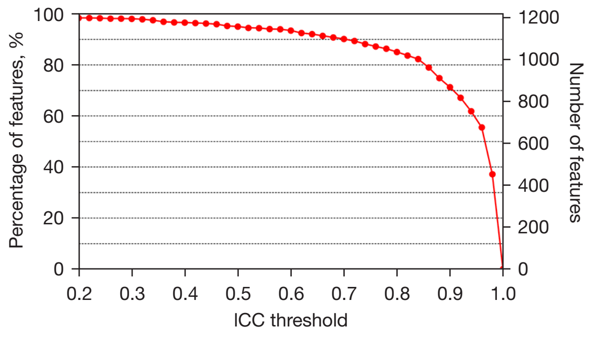

Figure 7 Stability analysis of radiomics features extracted from segmentations predicted by the 3D-cascade model in the testing set. This figure shows the percentage and number of features with ICC larger than the threshold. ICC, intraclass correlation coefficient.

pose challenges to the automated segmentation likely due to the less distinct borders between the two, or being mistaken as the urinary bladder instead of ovarian tumour.

Our results concurred with Liu *et al.* that showed high Dice scores in tumour segmentation in patients with OC, although different types of tumours were segmented; in our study, the primary OC, while Liu *et al.* segmented perihepatic or perisplenic metastases from OC (11). Rundo *et al.* developed an unsupervised fuzzy clustering-based method for the sub-segmentation of specific tissue components inside the ovarian tumour (10). This method retained the quantified radiodensity information for the purpose of results interpretability. Similarly, in our study this quantitative information was also retained by applying global normalization to all images during training and testing.

Different from conventional deep learning-based segmentation methods, nnU-Net focused on building a systemized and generalized processing pipeline for all kinds of biomedical segmentation tasks instead of network architecture (16). The effectiveness of such improvements was confirmed by a series of studies. Zhao *et al.* reported a Dice score of 0.92 for the segmentation of brain hemorrhage on CT images (20). Huo *et al.* reported Dice scores of 0.968 and 0.877 in segmenting whole breast and fibroglandular tissue on dynamic contrast-enhanced magnetic resonance images (24). Our results further confirmed that nnU-Net could adapt to our task and dataset well with high performance, and could become a convenient out-of-the-box tool for further quantitative analysis of OC. To further evaluate its potential clinical or research utility, we investigated the stability of the radiomics features based on tumour segmentation derived from 3D-cascade model and showed that 85.0% of the radiomics features achieved ICC >0.80, similar to Caballo *et al.*, but with more radiomics features tested in our study (25). The performance of our trained segmentation model may have promising role in radiomics analysis with the potential of saving time and labour in tumour segmentation of OC.

There were several limitations in this study. First, histological subtypes and FIGO stages were imbalanced in each center. This may lead to low Dice scores in those infrequent subtypes or FIGO stages. Second, the proportion of histological subtypes and FIGO stages were different in the 3 centers. This might lead to the difference of segmentation performance between validation and testing set, and limit the robustness of these trained models. Third, the training, validation and testing sets were relatively small, which result in a wide range of evaluation metrics on the testing set. The small testing set may limit the verification of model robustness on different types of OC. Continuous effort will be made to increase the sample size and further improve the accuracy and generalizability of the tumour segmentation task. Fourth, the specific task in this study was to segment OC, hence it may not be generalizable to other tumour types or other tasks on abdominopelvic CT images.

Conclusions

The deep learning-based models showed the potential to provide high performance automated segmentation with the highest performance metrics achieved by 3D-cascade model. The 3D-cascade model provided accurate volume assessment of OC on CT images and ensured stability of the extracted radiomic features, showing potential use in quantitative radiomics analysis.

Acknowledgments

Funding: This work was supported by the Health and Medical Research Fund, Hong Kong (grant number 08192106).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-1135/coif). EL reports that this work was supported by the Health and Medical Research Fund, Hong Kong (grant number 08192106). The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (No. UW 20-251), the Hong Kong East Cluster Research Ethics Committee (No. HKECREC-2020-040), and the Institutional Review Board of the Sun Yat-sen University Cancer Center (No. YB2018-52), and individual consent for this retrospective analysis was waived.

References

1. Stewart C, Ralyea C, Lockwood S. Ovarian Cancer: An

Integrated Review. Semin Oncol Nurs 2019;35:151-6.

2. Javadi S, Ganeshan DM, Qayyum A, Iyer RB, Bhosale P. Ovarian Cancer, the Revised FIGO Staging System, and the Role of Imaging. AJR Am J Roentgenol 2016;206:1351-60.
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563-77.
4. Danala G, Thai T, Gunderson CC, Moxley KM, Moore K, Mannel RS, Liu H, Zheng B, Qiu Y. Applying Quantitative CT Image Feature Analysis to Predict Response of Ovarian Cancer Patients to Chemotherapy. Acad Radiol 2017;24:1233-9.
5. An H, Wang Y, Wong EMF, Lyu S, Han L, Perucho JAU, Cao P, Lee EYP. CT texture analysis in histological classification of epithelial ovarian carcinoma. Eur Radiol 2021;31:5050-8.
6. Wang M, Perucho JAU, Hu Y, Choi MH, Han L, Wong EMF, Ho G, Zhang X, Ip P, Lee EYP. Computed Tomographic Radiomics in Differentiating Histologic Subtypes of Epithelial Ovarian Carcinoma. JAMA Netw Open 2022;5:e2245141.
7. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, Bellomi M. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp 2018;2:36.
8. Jiang H, Diao Z, Yao YD. Deep learning techniques for tumor segmentation: a review. J Supercomput 2022;78:1807-51.
9. Jin J, Zhu H, Zhang J, Ai Y, Zhang J, Teng Y, Xie C, Jin X. Multiple U-Net-Based Automatic Segmentations and Radiomics Feature Stability on Ultrasound Images for Patients With Ovarian Cancer. Front Oncol 2021;10:614201.
10. Rundo L, Beer L, Ursprung S, Martin-Gonzalez P, Markowetz F, Brenton JD, Crispin-Ortuzar M, Sala E, Woitek R. Tissue-specific and interpretable sub-segmentation of whole tumour burden on CT images by unsupervised fuzzy clustering. Comput Biol Med 2020;120:103751.
11. Liu J, Wang S, Linguraru MG, Yao J, Summers RM. Tumor sensitive matching flow: A variational method to detecting and segmenting perihepatic and perisplenic ovarian cancer metastases on contrast-enhanced abdominal CT. Med Image Anal 2014;18:725-39.
12. Wadhwa A, Bhardwaj A, Singh Verma V. A review on brain tumor segmentation of MRI images. Magn Reson Imaging 2019;61:247-59.
13. Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. Ieee Access 2021;9:82031-57.
14. Oktay O, Schlemper J, Folgoc LL, Lee MJ, Heinrich MP, Misawa K, Mori K, McDonagh SG, Hammerla NY, Kainz B, Glocker B, Rueckert D. Attention U-Net: Learning Where to Look for the Pancreas. arXiv preprint 2018; doi: https://arxiv.org/abs/1804.03999.
15. Zhang ZX, Liu QJ, Wang YH. Road Extraction by Deep Residual U-Net. Ieee Geoscience and Remote Sensing Letters 2018;15:749-53.
16. 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;18:203-11.
17. Ye X, Guo D, Ge J, Di X, Lu Z, Xiao J, Yao G, Lu L, Jin D, Yan S. Anatomy Guided Thoracic Lymph Node Station Delineation in CT Using Deep Learning Model. Int J Radiat Oncol Biol Phys 2021;111:e120-e1.
18. Heidenreich JF, Gassenmaier T, Ankenbrand MJ, Bley TA, Wech T. Self-configuring nnU-net pipeline enables fully automatic infarct segmentation in late enhancement MRI after myocardial infarction. Eur J Radiol 2021;141:109817.
19. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29.
20. Zhao X, Chen K, Wu G, Zhang G, Zhou X, Lv C, Wu S, Chen Y, Xie G, Yao Z. Deep learning shows good reliability for automatic segmentation and volume measurement of brain hemorrhage, intraventricular extension, and peripheral edema. Eur Radiol 2021;31:5012-20.
21. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, Hoekstra CJ, Boellaard R, Dekker AL, Gillies RJ, Aerts HJ, Lambin P. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol 2013;52:1391-7.
22. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, Roesch J, Rudofsky L, Friess M, Veit-Haibach P, Huellner M, Opitz I, Weder W, Frauenfelder T, Guckenberger M, Tanadini-Lang S. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol 2018;57:1070-4.
23. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res

2017;77:e104-7.

24. Huo L, Hu X, Xiao Q, Gu Y, Chu X, Jiang L. Segmentation of whole breast and fibroglandular tissue using nnU-Net in dynamic contrast enhanced MR images. Magn Reson Imaging 2021;82:31-41.

25. Caballo M, Pangallo DR, Mann RM, Sechopoulos I. Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence. Comput Biol Med 2020;118:103629.

Cite this article as: Wang Y, Wang M, Cao P, Wong EMF, Ho G, Lam TPW, Han L, Lee EYP. CT-based deep learning segmentation of ovarian cancer and the stability of the extracted radiomics features. Quant Imaging Med Surg 2023;13(8):5218-5229. doi: 10.21037/qims-22-1135