Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

journal homepage: www.sciencedirect.com/journal/clinical-and-translational-radiation-oncology

Auto-detection and segmentation of involved lymph nodes in HPV-associated oropharyngeal cancer using a convolutional deep learning neural network

Nicolette Taku[a], Kareem A. Wahid[a], Lisanne V. van Dijk[a,b], Jaakko Sahlsten[c], Joel Jaskari[c], Kimmo Kaski[b], Clifton D. Fuller[a], Mohamed A. Naser[a,*]

[a] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[b] Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
[c] Aalto University School of Science, Aalto, Finland

ABSTRACT

Purpose: Segmentation of involved lymph nodes on head and neck computed tomography (HN-CT) scans is necessary for the radiotherapy planning of early-stage human papilloma virus (HPV) associated oropharynx cancers (OPC). We aimed to train a deep learning convolutional neural network (DL-CNN) to segment involved lymph nodes on HN-CT scans.
Methods: Ground-truth segmentation of involved nodes was performed on pre-surgical HN-CT scans for 90 patients who underwent levels II-IV neck dissection for node-positive HPV-OPC (training/validation [n = 70] and testing [n = 20]). A 5-fold cross validation approach was used to train 5 DL-CNN sub-models based on a residual U-net architecture. Validation and testing segmentation masks were compared to ground-truth masks using predetermined metrics. A lymph auto-detection model to discriminate between “node-positive” and “node-negative” HN-CT scans was developed by thresholding segmentation model outputs and evaluated using the area under the receiver operating characteristic curve (AUC).
Results: In the DL-CNN validation phase, all sub-models yielded segmentation masks with median Dice $\geq$ 0.90 and median volume similarity score of $\geq$ 0.95. In the testing phase, the DL-CNN produced consensus segmentation masks with median Dice of 0.92 (IQR, 0.89–0.95), median volume similarity of 0.97 (IQR, 0.94–0.99), and median Hausdorff distance of 4.52 mm (IQR, 1.22–8.38). The detection model achieved an AUC of 0.98.
Conclusion: The results from this single-institution study demonstrate the successful automation of lymph node segmentation for patients with HPV-OPC using a DL-CNN. Future studies, including validation with an external dataset, are necessary to clarify its role in the larger radiation oncology treatment planning workflow.

Introduction

Approximately 66,000 cases of head and neck cancers will be diagnosed in the United States in 2022, including 30% of which will be oropharynx cancers (OPC) associated with human papilloma virus (HPV) [1,2]. Treatment options for patients with early-stage HPV-associated OPC include transoral robotic surgery (TORS) with ipsilateral neck dissection or primary radiotherapy, with or without chemotherapy [3]. As evidenced by prospective studies of TORS and neck dissection for resectable HPV-associated OPC, 70–90% of patients managed surgically will require adjuvant radiotherapy–suggesting a greater role for definitive radiotherapy as the upfront treatment strategy for this population [4,5].

Among patients with early-stage HPV-associated OPC dispositioned to definitive radiotherapy, accurate identification of involved lymph nodes is paramount to ensuring adequate dose delivery to all sites of regional disease. Although lymph node segmentation has traditionally been performed manually by a clinician, there is an evolving role for deep learning algorithms in the automation of target volume segmentation for cancers of the head and neck [6,7]. Within the context of OPC, deep learning algorithms have been used to auto-segment clinical target volumes (CTVs) inclusive of areas at risk for clinical and subclinical disease [8]. However, no deep learning algorithm has focused on the identification and segmentation of lymph node gross tumor volumes (GTVs) for HPV-associated OPC. The purpose of this study was to develop a deep learning convolutional neural network (DL-CNN) capable of identifying and segmenting radiographically and pathologically involved lymph nodes for HPV-associated OPC on contrast-enhanced head and neck computed tomography (HN-CT) scans. Furthermore, we aimed to use the DL-CNN to discriminate between

[*] Corresponding author at: The University of Texas MD Anderson Cancer Center, Department of Radiation Oncology, 1515 Holcombe Blvd., Houston, TX 77030-4009, United States.
E-mail address: manaser@mdanderson.org (M.A. Naser).

https://doi.org/10.1016/j.ctro.2022.06.007
Received 10 March 2022; Received in revised form 9 June 2022; Accepted 16 June 2022
Available online 18 June 2022
2405-6308/© 2022 The Author(s). Published by Elsevier B.V. on behalf of European Society for Radiotherapy and Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

node-negative and node-positive HN-CT scans.

Methods

After obtaining Institutional Review Board approval, 90 patients who underwent TORS with selective, ipsilateral, levels II-IV neck dissection for newly diagnosed, clinically node-positive, HPV-associated OPC at our institution were retrospectively identified from a departmental clinical database. Additionally, 20 randomly selected patients who underwent TORS with selective, ipsilateral, levels II-IV neck dissection and were found to have clinically and pathologically node-negative, HPV-associated OPC were included in the dataset. The inclusion criterion was age of at least 18 years at the time of diagnosis while the exclusion criteria were a history of radiotherapy treatment to the head and neck region and/or a history of neck dissection. The goal of this study was to train a DL-CNN to segment involved lymph nodes on the pre-surgical HN-CT scans of these patients.

Data preparation and preprocessing

Pre-surgical, contrast-enhanced, diagnostic HN-CT scans were identified for all patients. Expert, ground-truth segmentation of all radiographically involved lymph node GTVs was manually performed on node-positive HN-CT scans using RayStation Research (RaySearch Laboratories, Stockholm, Sweden) by one radiation oncologist with 5 years of clinical experience (NT) and reviewed by a second radiation oncologist with more than 15 years of clinical experience (CDF) [9]. Surgical pathology reports from selective neck dissections, including annotation of involved lymph nodes by nodal level, were reviewed. Histopathology findings were correlated with neuroradiology annotations to ensure that all segmented lymph nodes corresponded to pathologically involved lymph nodes. The ground-truth lymph node segmentations for each patient were then combined into a solitary “ground-truth mask”.

Pre-processing was performed on HN-CT scans to mitigate the variabilities in image size and resolution. The images and structure files were converted from Digital Imaging and Communications in Medicine (DICOM) format to Neuroimaging Informatics Technology Initiative (NIfTI) format and stripped of patient identifiers using the Advanced Medical Imaging Registration Engine (ADMIRE, Elekta AB, Stockholm, Sweden). The images were cropped to a specific sub-volume, with the auto-segmented cephalad border of the mandible, the manually segmented cephalad border of the sternum, and the auto-segmented external patient contour serving as the superior, inferior, and circumferential boundaries, respectively (Fig. 1). Image intensities were then truncated to the range of [−100, 300] Hounsfield units and rescaled to the range of [-1, 1] to increase soft tissue contrast [10]. The images and their respective ground-truth masks were resampled to 1.0 mm isotropic resolution using a trilinear interpolator in ADMIRE.

Model development

A DL-CNN was developed based on a 3-dimensional (3D) residual U-Net architecture included in the Medical Open Network for Artificial Intelligence (MONAI) software package [11]. This architecture has been utilized successfully in previous OPC tumor auto-segmentation studies [12,13]. The network consisted of 4 convolution blocks in the encoding and decoding branches with a bottleneck convolution block separating these two branches (Fig. 2). In the encoding branch, all convolutional layers used a kernel size of 3, with each block consisting of a two-strided convolution layer; the residual connections contained a two-strided and one-strided convolution layer. In the decoding branch, all convolutional layers used a kernel size of 3, with each block consisting of a two-strided transpose convolution layer, a one-strided convolution layer, and a residual connection. In the bottleneck, all convolutional layers used a kernel size of 1 and the residual connection consisted of a two-strided convolution layer. Batch normalization and Parametric Rectified Linear Unit (PReLU) activation functions were utilized throughout the architecture.

Model training & validation

The 90, node-positive HN-CT scans and their respective ground-truth masks served as input data by which the DL-CNN was developed. The node-positive HN-CT scans were randomly divided into 2 datasets–a training/validation dataset (n = 70) and a testing dataset (n = 20). Each of the 70 training/validation HN-CT scans was split into four, random regions (i.e., patches) of 96 × 96 × 96 voxels in dimension. The input tensor consisted of a batch size of 2, a single channel input, and 4 patches per image, yielding a summative input of (8, 1, 96, 96, 96). Each patch was evaluated for the presence of an involved lymph node with

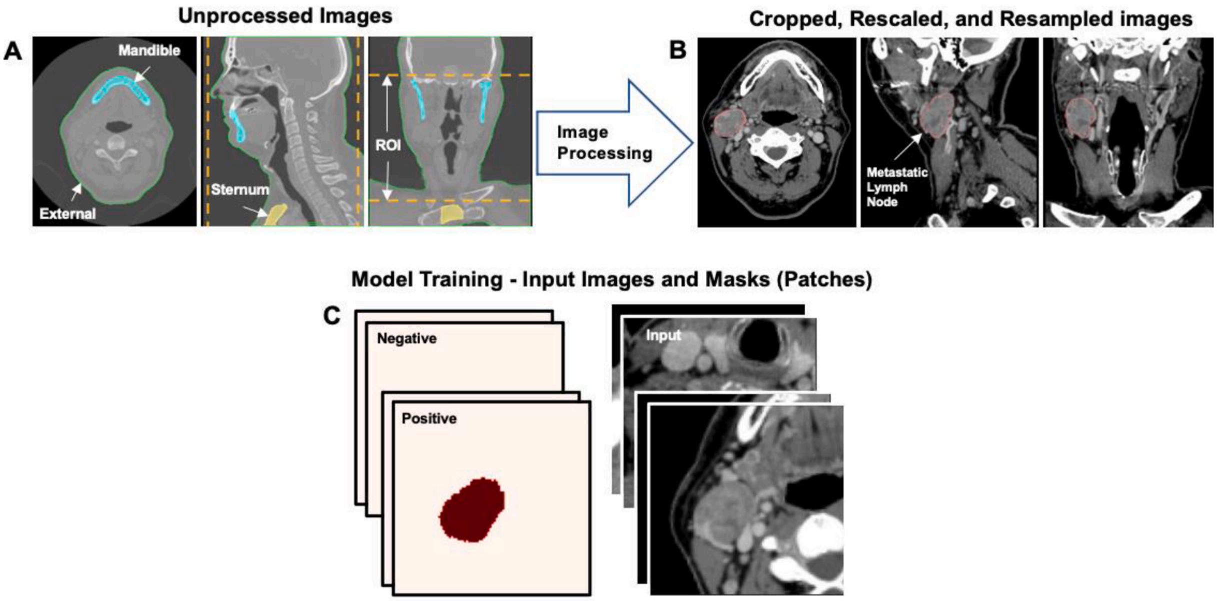

Fig. 1. Schematic representation of the pre-processing workflow. Head and neck computed tomography scans were cropped using the mandible, sternum, and external contours as boundaries (A & B). Scans were divided into 4 patches of 96 × 96 × 96 voxels in dimension (C).

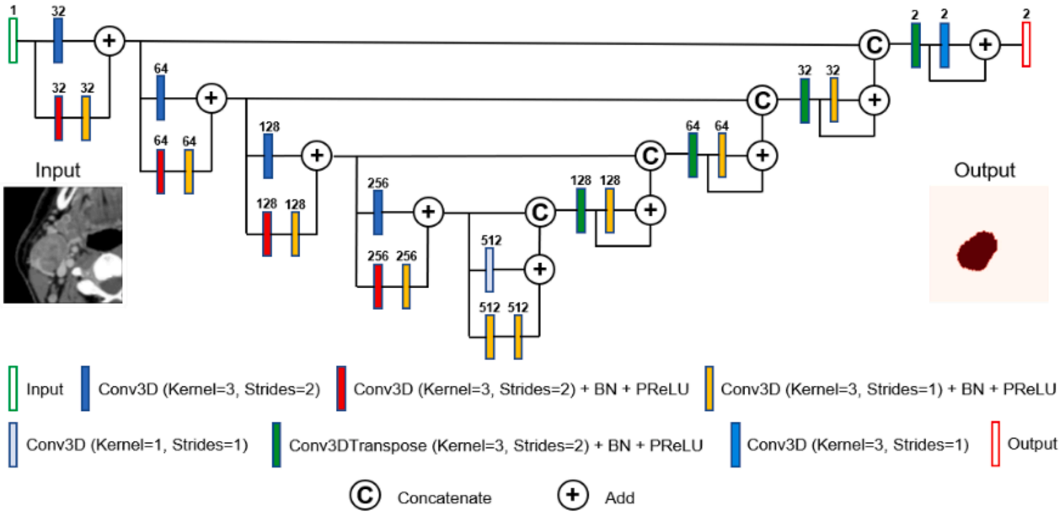
1
32
32 32
64
64 64
128
128 128
256
256 256
512
512 512
128 128
64 64
32 32
2 2
2
Input
Output
Input
Conv3D (Kernel=3, Strides=2)
Conv3D (Kernel=3, Strides=2) + BN + PReLU
Conv3D (Kernel=3, Strides=1) + BN + PReLU
Conv3D (Kernel=1, Strides=1)
Conv3DTranspose (Kernel=3, Strides=2) + BN + PReLU
Conv3D (Kernel=3, Strides=1)
Output
C Concatenate
+ Add

Fig. 2. Schematic representation of the U-Net architecture implemented for the deep learning convolutional neural network with annotations pertaining to the number of channels, batch normalization (BN) layers, and Parametric Rectified Linear Unit (PReLU) layers.

the center as foreground (i.e., involved lymph node present) or background (i.e., involved lymph node absent) and a 50% probability of either condition. Several data augmentation processes were implemented to minimize overfitting. Random spatial cropping was performed to patch the images and ground-truth masks. Random horizontal flips with 50% probability and random affine transformations, with an axial rotation range of 12 degrees and scale range of 10%, were also performed.

A 5-fold cross-validation approach was implemented to train the 5 separate sub-models of the DL-CNN. For each of the 5 sub-models, 80% of the HN-CT scans in the training/validation dataset and their respective ground-truth masks acted as model inputs for training purposes. The remaining 20% of HN-CT scans served for internal validation. One “validation segmentation mask” was generated per training/validation HN-CT scan, for a total of 70 validation segmentation masks. Validation segmentation masks were compared to ground-truth masks using overlap-based (Dice similarity coefficient [DSC]) and volume-based (volume similarity) metrics. The DL-CNN was trained for 700 epochs, with a learning rate of 2 × 10^{-4} for the first 550 epochs and 1 × 10^{-4} for the remaining 150 epochs.

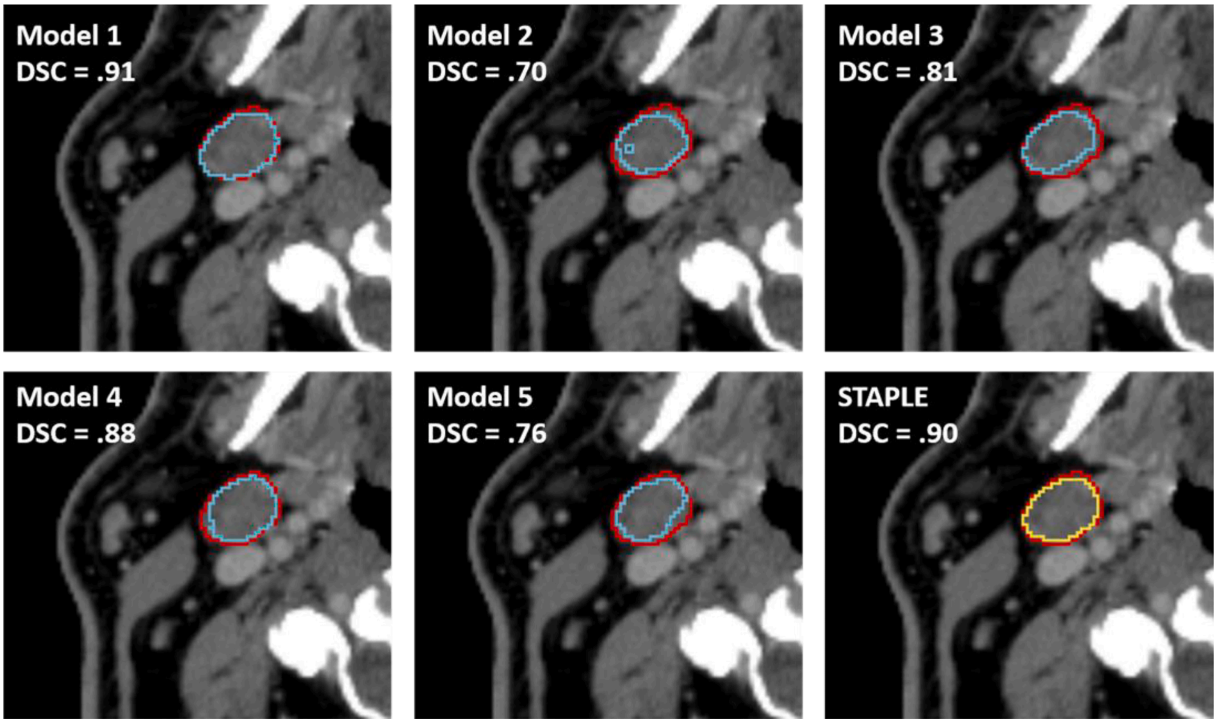

Fig. 3. Five sub-model testing segmentation masks and one consensus segmentation mask were generated for each head and neck computed tomography scan in the testing dataset. The red contour corresponds to the ground-truth masks, the blue contours correspond to the sub-model testing segmentation masks, and the yellow contour corresponds to the consensus segmentation mask generated by combing the 5 sub-model testing segmentation masks using the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Model testing

The performance of the DL-CNN to detect and segment involved lymph nodes was evaluated using the independent testing dataset of 20 HN-CT scans and their respective ground-truth masks. Additionally, 20 randomly selected HN-CT scans pertaining to patients with HPV-associated OPC and no involved lymph nodes were included in the testing dataset to evaluate the ability of the model to discriminate between "node-positive" (i.e., involved lymph node present) and "node-negative" (i.e., involved lymph node absent) HN-CT scans. In total, 5 "testing segmentation masks" were generated per HN-CT scan (1 testing segmentation mask per sub-model). These 5 testing segmentation masks were combined to create a "consensus segmentation mask" using the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm (Fig. 3) [14]. The testing segmentation masks and consensus segmentation masks for the 20 node-positive HN-CT scans in the testing dataset were compared to their respective ground-truth masks using overlap-based (DSC), volume-based (volume similarity), spatial distance-based (Hausdorff distance [HD]), and probabilistic-based (Cohen Kappa Coefficient [CKC]) metrics [15].

Model discrimination between node-positive and node-negative HN-CT scans was evaluated by scoring each of the 5 testing segmentation masks for a respective HN-CT scan as either "1″ to indicate that an auto-segmented structure was generated or "0" to indicate that no auto-segmented structure was generated. The scores were then summed to yield a "summative score" ranging from 0 (i.e., no auto-segmented structure was generated by any of the 5 sub-models) to 5 (i.e., auto-segmented structures were generated by all 5 sub-models). A HN-CT scan was considered "node-positive" if the summative score was equal to 5, and "node-negative" if the summative score was ≤ 4. This score threshold was chosen empirically from test results to maximize the accuracy, sensitivity, and positive predictive value of the DL-CNN. The model discrimination was evaluated by determining the area under the receiver operating characteristic curve (AUC). Three image resampling resolutions–high (1.0 mm), medium (1.5 mm), and low (2.0 mm)–were used to evaluate the impact of image resolution on the discriminatory ability of the DL-CNN.

Clinical evaluation

Three radiation oncology physician observers (with at least 4 years of clinical experience) were asked to review lymph node segmentations on node-positive HN-CT scans from the independent testing dataset. Each node-positive HN-CT scan appeared twice–once with the manually generated ground-truth mask and once with the DL-CNN generated consensus segmentation mask–for a total of 40 HN-CT scans reviewed per observer. Observers were asked to rate the lymph node segmentations on a Likert scale for clinical acceptability (1 = clinically acceptable, highly accurate; 2 = clinically acceptable, errors not significant; 3 = requires correction, minor errors; 4 = requires correction, large errors). Using a modified Turing test, observers were then asked to determine whether the segmentation was generated by a human or a computer, [12,16]. Lastly, observers were asked to rate their confidence in this determination using a Likert scale (1 = very confident; 2 = somewhat confident; 3 = somewhat unconfident; 4 = very unconfident). The Wilcoxon signed rank test and the McNemar test were used to compare observers' manually generated and DL-CNN generated segmentation evaluations using IBM SPSS Statistics v26, with a *p* value of less than 0.05 indicative of statistical significance.

Results

Patient and tumor characteristics are presented in Table 1. The median age at diagnosis was 60 years and there was a male sex predominance (n = 101, 92%). The majority of the patients had no history of cigarette smoking (n = 72, 66%) and cT1 disease (n = 63, 57%). Among cN1 patients, there was a median of 1 involved lymph node (range, 1–4) in the training/validation dataset and 1 involved lymph node (range, 1–3) in the testing dataset. The median volumes for individual lymph nodes and ground truth segmentation masks were 6.7 cc (interquartile range [IQR], 2.8–10.3 cc) and 8.4 cc (5.6–12.9 cc), respectively.

DL-CNN validation performance

Segmentation mask metrics for model validation are presented in

Table 1
Patient and tumor clinical characteristics for all patients (N = 110), patients in the training/validation dataset (n = 70), and patients in the testing dataset (n = 40).

Characteristic	All n (%)	Training/Validation n (%)	Testing n (%)
Median age [IQR], y	60 [53–65]	60 [54–65]	59 [53–67]
Sex			
Male	101 (91.8)	66 (94.3)	35 (87.5)
Female	9 (8.2)	4 (5.7)	5 (12.5)
Smoking Status			
Never	72 (65.5)	44 (62.9)	28 (70.0)
Former	34 (30.9)	24 (34.3)	10 (25.0)
Current	4 (3.6)	2 (2.9)	2 (5.0)
Oropharynx subsite			
Base of tongue	51 (46.4)	38 (54.3)	13 (32.5)
Tonsil	59 (53.6)	32 (45.7)	27 (67.5)
Clinical tumor classification			
cT1	63 (57.3)	44 (62.9)	19 (47.5)
cT2	47 (42.7)	26 (37.1)	21 (52.5)
Clinical lymph node classification			
cN0	20 (18.2)	0 (0.0)	20 (50.0)
cN1	90 (81.8)	70 (100.0)	20 (50.0)
Median number of removed lymph nodes [IQR]	26 [21–34.8]	26.5 [21–35]	26 [23–29]
Number of involved lymph nodes			
0	20 (18.2)	0 (0.0)	20 (50.0)
1	68 (61.8)	53 (75.7)	16 (40.0)
2	18 (16.4)	16 (22.9)	2 (5.0)
3	3 (2.7)	0 (0)	2 (5.0)
4	1 (0.9)	1 (1.4)	0 (0.0)
Median individual lymph node volume [IQR], cc	6.7 [2.8–10.3]	7.7 [3.0–11.1]	5.3 [2.1–8.5]
Median ground truth segmentation mask volume [IQR], cc	8.4 [5.6–12.9]	9.6 [7.3–14.9]	6.6 [2.2–15.5]

Abbreviations: IQR, interquartile range; cc, cubic centimeters; y, years.

Table 2
Minimum, maximum, median, interquartile range values for the overlap-based (Dice similarity coefficient) and volume-based (volume similarity) metrics for the sub-model validation segmentation masks when compared to the ground-truth masks.

	Sub-Model #1				Sub-Model #2				Sub-Model #3				Sub-Model #4				Sub-Model #5			
	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR
DSC	0.83	0.94	0.91	0.88–0.92	0.81	0.96	0.92	0.90–0.94	0.83	0.94	0.91	0.88–0.93	0.80	0.95	0.91	0.88–0.94	0.67	0.97	0.90	0.85–0.91
VS	0.84	0.99	0.96	0.95–0.98	0.86	1.00	0.96	0.94–0.98	0.85	0.99	0.97	0.93–0.98	0.80	0.99	0.95	0.93–0.97	0.70	1.00	0.95	0.92–0.97

Abbreviations: DSC, Dice similarity coefficient; IQR, interquartile range; Max., maximum; Min., minimum; VS, volume similarity.

Table 3
Minimum, maximum, median, interquartile range values for the overlap-based (Dice similarity coefficient), volume-based (volume similarity), spatial distance-based (Hausdorff distance), and probabilistic-based (Cohen Kappa Coefficient) metrics for the sub-model testing segmentation masks and consensus segmentation masks when compared ground-truth masks.

	Sub-Model #1				Sub-Model #2				Sub-Model #3				Sub-Model #4				Sub-Model #5				Consensus (STAPLE)			
	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR	Min.	Max.	Med.	IQR
DSC	0.55	0.95	0.92	0.89–0.94	0.66	0.95	0.92	0.88–0.94	0.58	0.95	0.92	0.87–0.94	0.62	0.96	0.91	0.87–0.94	0.69	0.96	0.92	0.88–0.94	0.61	0.96	0.92	0.89–0.95
VS	0.64	1.00	0.97	0.95–0.98	0.70	1.00	0.96	0.92–0.99	0.59	1.00	0.97	0.93–0.99	0.73	1.00	0.97	0.91–0.99	0.72	1.00	0.97	0.92–0.99	0.68	1.00	0.97	0.94–0.99
HD	1.11	92.0	4.92	1.11–16.0	1.22	90.0	5.78	1.22–17.4	1.65	86.9	5.08	1.64–18.5	1.22	90.7	4.15	1.22–9.04	1.22	91.4	5.56	1.22–11.7	1.22	90.9	4.52	1.22–8.38
CKC	0.55	0.95	0.92	0.89–0.94	0.66	0.95	0.92	0.88–0.94	0.58	0.95	0.92	0.87–0.95	0.62	0.96	0.91	0.87–0.94	0.69	0.96	0.92	0.88–0.94	0.61	0.96	0.92	0.89–0.95

Abbreviations: CKC, Cohen Kappa Coefficient; DSC, Dice similarity coefficient; HD, Hausdorff distance (in mm); IQR, interquartile range; Max., maximum; Min., minimum; STAPLE, Simultaneous Truth and Performance Level Estimation; VS, volume similarity.

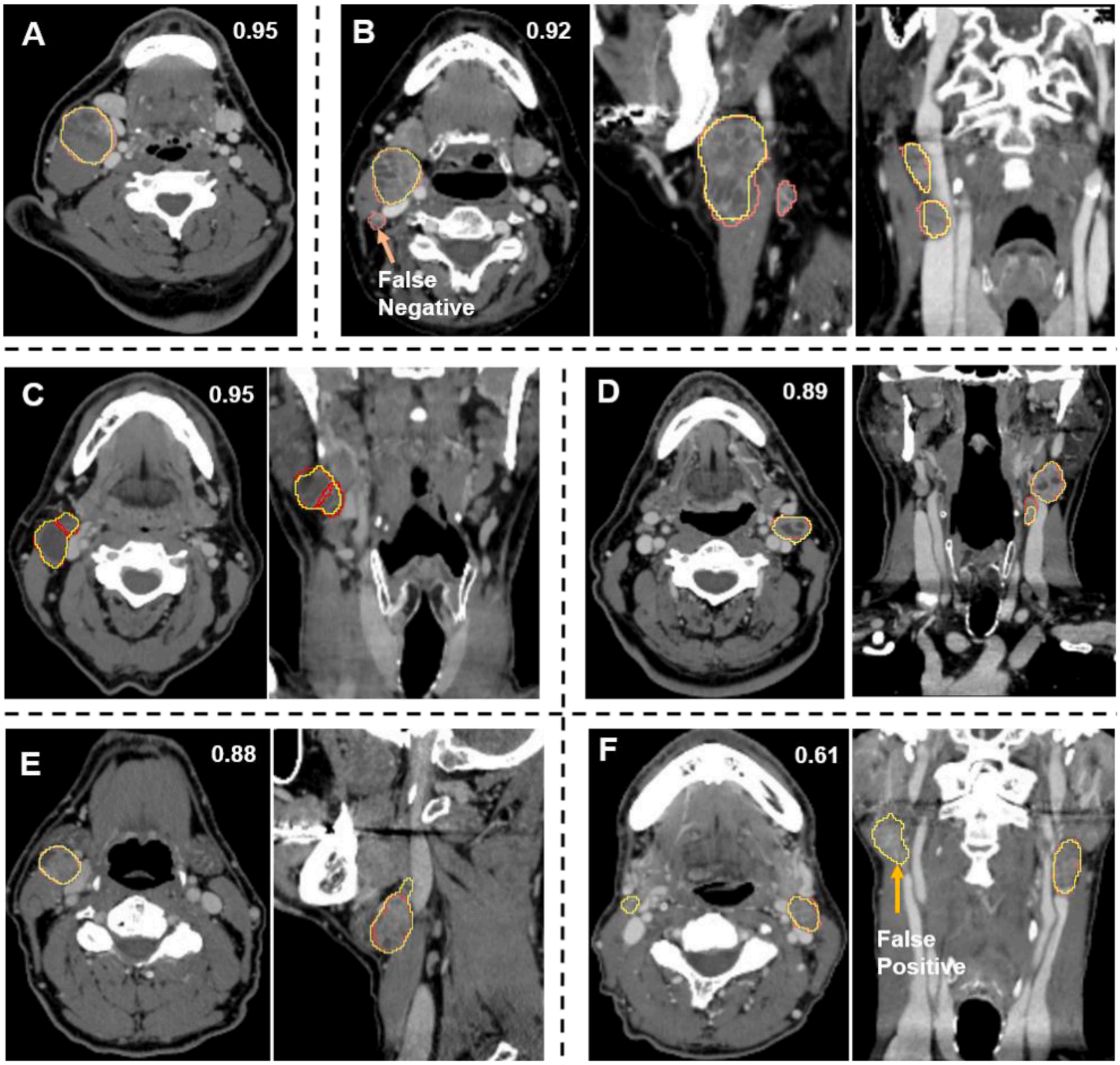

Fig. 4. Comparison of consensus segmentation masks (yellow) to ground-truth masks (red) for a subset of testing dataset patients with greater or equal Dice similarity coefficients (A, B, C; 1 involved lymph node, 3 involved lymph nodes, and 2 involved lymph nodes, respectively), slightly lower Dice similarity coefficients (D, E; 2 involved lymph nodes and 1 involved lymph node, respectively), and much lower Dice similarity coefficient (F; 1 involved lymph node) than the median Dice similarity coefficient value of 0.92. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2. When compared to ground-truth masks, sub-model #4 achieved the highest median DSC, with a score of 0.92 (IQR, 0.90–0.94) for the validation segmentation masks. All the 5 sub-models generated validation segmentation masks with a median DSC of at least 0.90. Similarly, all the 5 sub-models generated validation segmentation masks with a median volume similarity score of at least 0.95, with sub-model #1 achieving the highest median volume similarity score and narrowest volume similarity IQR.

DL-CNN testing performance

Segmentation mask metrics for model testing are presented in Table 3. When compared to ground-truth masks, the median DSC for testing segmentation masks was greater than 0.90 for all sub-models. The median DSC for consensus segmentation masks was 0.92 (IQR, 0.89–0.95). Comparisons between the testing segmentation masks and ground-truth masks for a subset of cases based on DSC are depicted in Fig. 4. A maximum volume similarity score of 1.0 was achieved by all sub-models for testing segmentation masks, with sub-model #4 achieving the highest minimum volume similarity score and a median volume similarity score of 0.97. The median volume similarity score for consensus segmentation masks was 0.97 (IQR, 0.94–0.99). All sub-models achieved a median HD less than 6 mm, with a median HD for consensus segmentation masks of 4.52 mm (IQR, 1.22–8.38). The median CKC for testing segmentation masks was nearly identical across the sub-models, and the median CKC for consensus segmentation masks was 0.92 (IQR, 0.89–0.95). There was no statistically significant difference in DSC, volume similarity, HD, or CKC when cases pertaining to smokers were compared to non-smokers (**Figure S1**).

DL-CNN discrimination performance

Confusion matrices and receiver operating characteristic curves for the three imaging resolutions are presented in Fig. 5. The medium resampled resolution model achieved the most optimal identification of node-positive HN-CT scans (AUC = 0.98), with 20 of 20 HN-CT scans

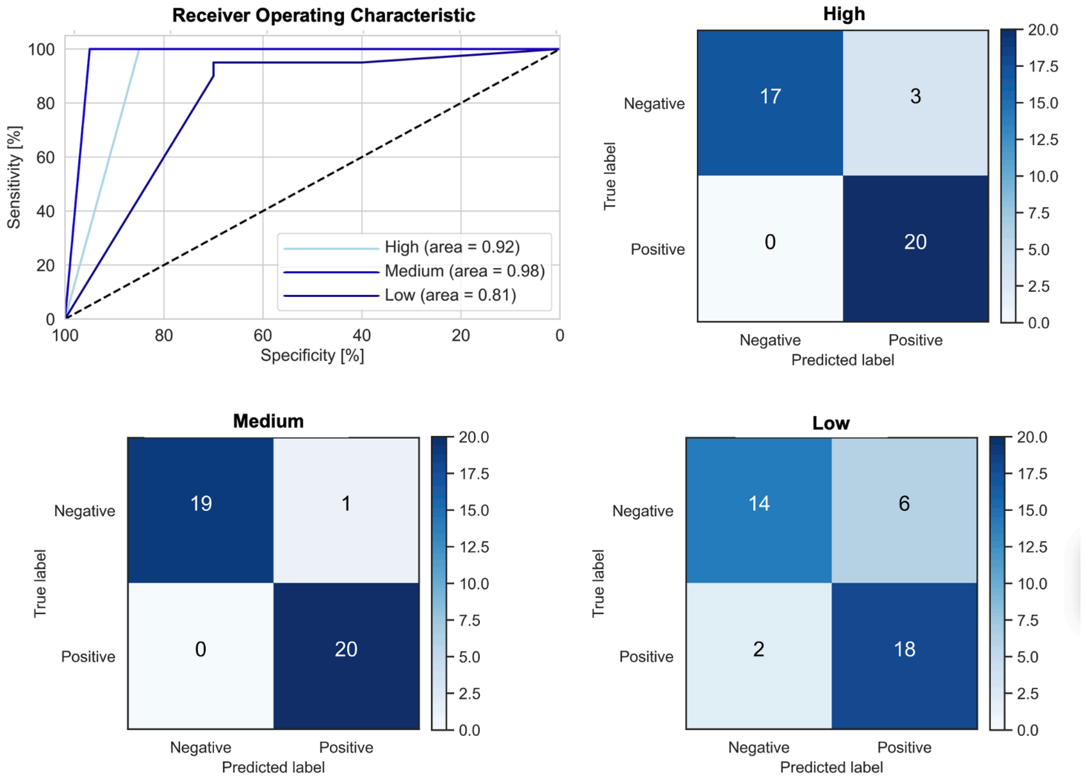

Fig. 5. Receiver operating characteristic curves for node-positive versus node-negative head and neck computed tomography scan discrimination comparing three resampled image resolutions (High, 1.0 mm; Medium, 1.5 mm; and Low, 2.0 mm) and their corresponding confusion matrices.

Table 4
Physician observer evaluations of manually generated (i.e., "human") and model generated (i.e., "computer") segmentations.

Clinical Acceptability of Segmentations				Source of Segmentations				Confidence in Source of Segmentations			
	Computer n (%)	Human n (%)	*P* value		Computer n (%)	Human n (%)	*P* value		Computer n (%)	Human n (%)	*P* value
Observer 1			0.83				0.66				0.72
Clinically acceptable, highly accurate	13 (0.65)	13 (0.65)		Computer	8 (0.40)	9 (0.45)		Very confident	2 (0.10)	6 (0.30)	
Clinically acceptable, errors not significant	3 (0.15)	3 (0.15)		Human	12 (0.6)	11 (0.55)		Somewhat confident	14 (0.70)	9 (0.45)	
Requires correction, minor errors	2 (0.10)	3 (0.15)						Somewhat unconfident	4 (0.20)	4 (0.20)	
Requires correction, large errors	2 (0.10)	1 (0.05)						Very unconfident	0 (0.0)	1 (0.05)	
Observer 2			0.75				0.82				0.046
Clinically acceptable, highly accurate	4 (0.20)	8 (0.4)		Computer	11 (0.55)	11 (0.55)		Very confident	0 (0.0)	1 (0.05)	
Clinically acceptable, errors not significant	8 (0.40)	10 (0.50)		Human	9 (0.45)	9 (0.45)		Somewhat confident	4 (0.20)	9 (0.45)	
Requires correction, minor errors	6 (0.30)	1 (0.05)						Somewhat unconfident	15 (0.75)	10 (0.50)	
Requires correction, large errors	2 (0.10)	1 (0.05)						Very unconfident	1 (0.05)	0 (0.0)	
Observer 3			0.41				0.26				0.28
Clinically acceptable, highly accurate	14 (0.70)	10 (0.50)		Computer	7 (0.35)	7 (0.35)		Very confident	0 (0.0)	0 (0.0)	
Clinically acceptable, errors not significant	5 (0.25)	10 (0.50)		Human	13 (0.65)	13 (0.65)		Somewhat confident	3 (0.15)	1 (0.0)	
Requires correction, minor errors	1 (0.05)	0 (0.0)						Somewhat unconfident	7 (0.35)	6 (0.30)	
Requires correction, large errors	0 (0.0)	0 (0.0)						Very unconfident	10 (0.50)	13 (0.65)	

with involved lymph nodes correctly identified as node-positive and 19 of 20 of the remaining HN-CT scans correctly identified as node-negative. In contrast, the low resampled resolution model had the worst classification of HN-CT scans (AUC = 0.81), with 2 of 20 HN-CT scans with involved lymph nodes incorrectly identified as node-negative and 6 of 20 of HN-CT scans with no involved lymph nodes incorrectly identified as node-positive. Illustrative examples of the detection process and individual test case predictions using the best-performing model (medium resolution) are shown in Figure S2.

Clinical performance

Physician observer evaluations of manually generated (i.e., "human") and DL-CNN generated (i.e., "computer") segmentations are presented in Table 4**.** There was no statistically significant difference in the clinical acceptability of manually generated versus DL-CNN generated segmentations among observers. Although observer 1 was found to have a statistically significant difference in reported confidence in segmentation source favoring manually generated segmentations, there was no statistically significant difference in the correct identification of manually generated versus DL-CNN generated segmentations for any of the three observers. Additionally, there was no difference in observer confidence in source of segmentation for the remaining 2 observers.

Discussion

Deep learning is a subset of machine learning that uses neural networks to learn and classify data [7]. Within the context of radiation oncology, deep learning algorithms have emerged as a mechanism by which to automate image-based segmentation of normal structures and target volumes–thereby improving segmentation reproducibility and efficiency. We used supervised learning, contrast-enhanced HN-CT scans with corresponding ground-truth masks, and a patch-based approach to training a DL-CNN to auto-segment involved lymph node GTVs for patients with early-stage HPV-associated OPC. In the validation phase, we found that our DL-CNN achieved median DSC and volume similarity scores of at least 0.90 and 0.95, respectively. When tested on unseen data, our DL-CNN was notable for a median DSC of 0.92 for consensus segmentation masks and AUC of 0.98 for successful identification of node-positive HN-CT scans. Physician observers were unable to distinguish between manually generated and DL-CNN generated lymph node segmentations and there was no statistically significant difference between the clinical acceptability of the two groups of segmentations.

Our use of a DL-CNN to auto-segment involved lymph node GTVs on unseen HN-CT scans builds on the work of Cardenas et al. in the auto-segmentation of lymph node level CTVs for head and neck radiotherapy treatment planning [17]. Using 71 HN-CT scans acquired during radiotherapy simulation and their respective ground truth contours for model training and validation, they found that their DL-CNN achieved a DSC of 0.90 for the auto-segmentation of CTVs for lymph node neck levels II-IV in the testing phase. Physician review of an independent dataset of 32 HN-CT scans found that over 99% of the DL-CNN auto-segmented lymph node CTVs were either sufficient for clinical use or required minor revisions [17]. Taken together with our results, these findings suggest that, for patients with early-stage HPV-associated OPC, DL-CNNs may be used to auto-detect node-positive HN-CT scans, auto-segment lymph node GTVs, and auto-segment lymph node level CTVs as part of the radiation oncology treatment planning workflow with a high degree of fidelity.

Prior studies on the role of DL-CNNs in the auto-segmentation of involved lymph nodes of the head and neck are limited. Bielak et al. investigated the impact of various magnetic resonance imaging sequences on auto-segmentation of lymph nodes for 24 patients with squamous cell carcinoma of the head and neck and found a maximum DSC of 0.58 [18]. Similarly, Wang et al. integrated the extraction of various imaging features from contrast-enhanced HN-CT scans for 82 patients with squamous cell carcinoma of the head and neck into a DL-CNN and achieved a mean DSC score of 0.94 for the highest performing model [19]. As iodinated contrast is often administered during radiotherapy simulation to enhance vascular visibility and facilitate target volume segmentation, we used 90 contrast-enhanced HN-CT scans pertaining to patients with early-stage OPC to train, validate, and test our model [20,21]. Additionally, as lymph node metastases in HPV-associated disease are characterized by several distinct features on clinical imaging, including cystic composition and matted conglomeration, we limited our cohort to this subset of head and neck cancers [22].

Our results have implications for treatment delivery and toxicity reduction among patients treated with definitive radiation for early-stage HPV-associated OPC. In a study of radiotherapy treatment tumor response for this population, Kabarriti and colleagues found that the mean lymph node volume decreased by 12.9 cm^3 on mid-radiotherapy HN-CT scans when compared to baseline simulation HN-CT scans [23]. Additionally, studies of adaptive planning for OPC have demonstrated that the majority of patients will require their radiotherapy to replanned at least once during the course of treatment, principally due to changes in neck soft tissue that impact the robustness of radiation dose to both targeted and non-targeted tissues [24]. Incorporation of our DL-CNN into primary and adaptive radiotherapy planning for early-stage HPV-associated OPC may facilitate efficacy of nodal GTV delineation, improvement in target coverage, sparing of normal tissues, and efficiency of workflow [25].

There are several limitations to our study. We included patients with early-stage HPV-associated OPC who had undergone TORS of the primary tumor and lymph node dissection. Therefore, it is possible that our results may not be fully generalizable to patients with more locoregionally advanced disease, including patients with 3 or more radiographically involved lymph nodes and/or radiographic evidence of extranodal extension. As we combined all involved lymph nodes into a single ground truth segmentation mask, our use of the HD metric may not appropriately reflect comparisons between manually generated and DL-CNN generated segmentation masks inclusive of multiple lymph nodes separated spatially. Nonetheless, our use of HD is complemented by other reported metrics that demonstrate the strength of the model. Our findings represent the results of a small cohort of HN-CT scans obtained at a single institution. Additional studies are needed to externally validate the model in a larger dataset of HN-CT scans performed at other institutions, Our DL-CNN was trained, validated, and tested on contrast-enhanced HN-CT scans. However, as iodine-based contrast may impact dose calculations, non-contrast HN-CT scans are the gold standard for radiotherapy treatment planning [26]. Further studies incorporating HN-CT scans acquired without contrast may facilitate the future integration of the DL-CNN into online and offline adaptive radiotherapy workflows for which intravenous contrast is not administered.

Conclusion

Patients diagnosed with early-stage HPV-associated OPC are often found to have clinical evidence of lymph node involvement at the time of diagnosis. Manual segmentation of radiographically involved lymph nodes is an integral part of treatment planning for those patients dispositioned to definitive radiotherapy. Here we have presented a DL-CNN that can be used to generate auto-segmentations of involved lymph nodes that cannot be reliably distinguished from human segmentations by physician observers. Future studies on the validation of the DL-CNN on larger external datasets of HN-CT scans, on HN-CT scans acquired without contrast, and on HN-CT scans pertaining to patients with surgically unresectable disease will further clarify the role of our DL-CNN in the larger radiation oncology workflow.

Funding statement

Nicolette Taku received funding from the National Institutes of Health Research Education Program (R25EB025787). Kareem A. Wahid is supported by the Dr. John J. Kopchick Fellowship through The University of Texas MD Anderson UTHealth Graduate School of Biomedical Sciences, the American Legion Auxiliary Fellowship in Cancer Research, and an NIH/National Institute for Dental and Craniofacial Research (NIDCR) F31 fellowship (1 F31DE031502-01). Clifton David Fuller received funding from the National Institute for Dental and Craniofacial Research Award [1R01DE025248-01/R56DE025248] and Academic-Industrial Partnership Award [R01 DE028290], the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant [NSF 1557679], the NIH Big Data to Knowledge (BD2K) Program of the National Cancer Institute (NCI) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award [1R01CA214825], the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program [1R01CA218148], the NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program [P30CA016672], the NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award [P50 CA097007] and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program [R25EB025787]. He has also received direct industry grant support from Elekta. Mohamed Naser received funding from the National Institutes of Health (R01DE028290-01).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2022.06.007.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7–33.

[2] Centers for Disease Control and Prevention. Cancers Associated with Human Papillomavirus, United States—2013–2017. USCS Data Brief, no 18. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2020.

[3] Adelstein DJ, Ridge JA, Brizel DM, Holsinger FC, Haughey BH, O'Sullivan B, et al. Transoral resection of pharyngeal cancer: summary of a National Cancer Institute Head and Neck Cancer Steering Committee Clinical Trials Planning Meeting, November 6–7, 2011, Arlington. Virginia *Head Neck* 2012;34(12):1681–703.

[4] Nichols AC, Theurer J, Prisman E, Read N, Berthelet E, Tran E, et al. Radiotherapy versus transoral robotic surgery and neck dissection for oropharyngeal squamous cell carcinoma (ORATOR): an open-label, phase 2, randomised trial. Lancet Oncol 2019;20(10):1349–59.

[5] Ferris RL, Flamand Y, Weinstein GS, Li S, Quon H, Mehra R, et al. Phase II Randomized Trial of Transoral Surgery and Low-Dose Intensity Modulated Radiation Therapy in Resectable p16+ Locally Advanced Oropharynx Cancer: An ECOG-ACRIN Cancer Research Group Trial (E3311). J Clin Oncol 2022;40(2): 138–49.

[6] 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(5):578–95.

[7] Shrestha A, Mahmood A. Review of Deep Learning Algorithms and Architectures. IEEE Access 2019;7:53040–65.

[8] Cardenas CE, Anderson BM, Aristophanous M, Yang J, Rhee DJ, McCarroll RE, et al. Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks. Phys Med Biol 2018;63(21):215026.

[9] Bodensteiner D. RayStation: External beam treatment planning system. Med Dosim 2018;43(2):168–76.

[10] Naser MA, Wahid KA, Grossberg AJ, et al. Deep Learning Auto-Segmentation of Cervical Neck Skeletal Muscle for Sarcopenia Analysis Using Pre-Therapy CT in Patients with Head and Neck Cancer. *medRxiv.* 2021:2021.2012.2019.21268063.

[11] Consortium M. MONAI: Medical Open Network for AI. In:2020.

[12] Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. Clin Transl Radiat Oncol 2022;32:6–14.

[13] Naser MA, et al. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A, editors. Head and Neck Cancer Primary Tumor Auto Segmentation Using Model Ensembling of Deep Learning in PET/CT Imagesvol 13209. Cham: Springer; 2022. https://doi.org/10.1007/978-3-030-98253-9_11.

[14] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23(7):903–21.

[15] Aydin OU, Taha AA, Hilbert A, Khalil AA, Galinovic I, Fiebach JB, et al. On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking. Eur Radiol Exp 2021;5(1).

[16] Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, Stoep J, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. Med Phys 2018;45(11):5105–15.

[17] Cardenas CE, Beadle BM, Garden AS, Skinner HD, Yang J, Rhee DJ, et al. Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach. Int J Radiat Oncol Biol Phys 2021;109(3):801–12.

[18] Bielak L, Wiedenmann N, Berlin A, Nicolay NH, Gunashekar DD, Hägele L, et al. Convolutional neural networks for head and neck tumor segmentation on 7-channel multiparametric MRI: a leave-one-out analysis. Radiat Oncol 2020;15(1).

[19] Wang Y, Zamiela C, Thomas TV, et al. 3D Texture Feature-Based Lymph Node Automated Detection in Head and Neck Cancer Analysis. Paper presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 16-19 Dec. 2020; 2020.

[20] Biau J, Lapeyre M, Troussier I, Budach W, Giralt J, Grau C, et al. Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 Update. Radiother Oncol 2019;134:1–9.

[21] Merlotti A, Alterio D, Vigna-Taglianti R, Muraglia A, Lastrucci L, Manzo R, et al. Technical guidelines for head and neck cancer IMRT on behalf of the Italian association of radiation oncology - head and neck working group. Radiat Oncol 2014;9(1).

[22] Joo L, Bae YJ, Choi YJ, Lee YS, Chung SR, Suh CH, et al. Prediction model for cervical lymph node metastasis in human papillomavirus-related oropharyngeal squamous cell carcinomas. Eur Radiol 2021;31(10):7429–39.

[23] Kabarriti R, Brodin NP, Lundgren G, Ohri N, Tomé WA, Kalnicki S, et al. Early Response Assessment on Mid-treatment Computed Tomography Predicts for Locoregional Recurrence in Oropharyngeal Cancer Patients Treated With Definitive Radiation Therapy. Int J Radiat Oncol Biol Phys 2018;102(4):1036–45.

[24] Deiter N, Chu F, Lenards N, Hunzeker A, Lang K, Mundy D. Evaluation of replanning in intensity-modulated proton therapy for oropharyngeal cancer: Factors influencing plan robustness. Med Dosim 2020;45(4):384–92.

[25] Green OL, Henke LE, Hugo GD. Practical Clinical Workflows for Online and Offline Adaptive Radiation Therapy. Semin Radiat Oncol 2019;29(3):219–27.

[26] Yamada S, Ueguchi T, Ogata T, Mizuno H, Ogihara R, Koizumi M, et al. Radiotherapy treatment planning with contrast-enhanced computed tomography: feasibility of dual-energy virtual unenhanced imaging for improved dose calculations. Radiat. Oncol. 2014;9(1).