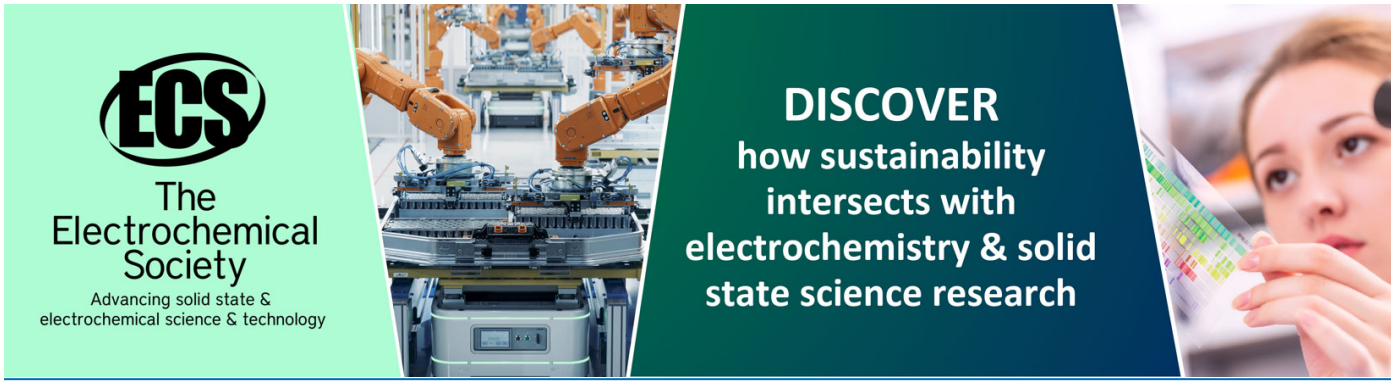
Training and validation of a commercial deep learning contouring platform



Training and validation of a commercial deep learning contouring platform

J Koo, J Caudell, V Feygelman, E Moros and K Latifi
Moffitt Cancer Center, Tampa, FL 33612, USA

E-mail: Kujtim.Latifi@moffitt.org

Abstract. We present our experience with training and validation of a commercially available deep learning algorithm for organs at risk(OAR) auto-contouring. Computed tomography(CTs) with OARs from a cohort of 213 head and neck(H&N) patients were used for training the deep learning model. A separate cohort of 85 CTs and structure sets was used for validation. All OARs (13) were contoured by a single physician. Metrics such as the DICE similarity coefficient (DSC), Jaccard similarity coefficient (JSC), and volumetric difference (VD) were used to analyze contouring variation. Mean DSC and JSC values ranged 0.48-0.89 and 0.32-0.8, respectively, depending on OAR. A DSC value ≥0.7 indicated low inter-observer variability. In our study, all but one of the contours were above this threshold. DSC for the middle pharyngeal constrictor had the lowest value of all the contours. This may be due to the small volume of this structure. Qualitative assessment of auto-segmented structure samples confirmed the reliability of DSC by demonstrating the compatibility between the expert's evaluation and DSC values. Overall, we found that deep learning auto contouring is a useful tool to speed up the process of contouring in radiotherapy treatment planning.

1. Introduction

Contouring of organs at risk (OARs) is a time-consuming part of radiotherapy treatment planning. There are many approaches to speed up this process. These range from atlas-based segmentation (ABS) algorithms, active contour model (ACM), machine learning, and others. ABS methods are often used and some software packages are commercially available; however, they are generally limited to certain OARs and local elastic registration is required for accurate results, which is time-consuming. [1-2] ACM delineates the outline of objects using energy constraint and image forces. Being a deformable model, ACM is frequently adopted in medical image processing; however, it is computationally intensive especially when the image size is large and accuracy demands aconvergence criteria. [3]

Machine learning techniques, deep learning in particular, have been rapidly growing in the last few years in many industries and are being adopted in radiotherapy for many tasks such as clinical outcome prediction, medical image analysis, dose-response modeling and image segmentation. [4-7] The main advantage of this technique is its ability to learn the most suitable representation of data for given tasks. In this paper, we present our experience with training and validation of a commercial deep learning automatic contouring software for H&N OARs.

2. Materials and Methods

2.1. Quantitative assessment

A commercial deep learning auto-segmentation software (DLCExpert™, Mirada DBx., Mirada Medical Ltd.) was utilized. CTs and OAR contours from a cohort of 213 H&N patients were used for training the deep learning model. A separate cohort of 85 CTs and contours were used for validation. Patients with tilted head position were excluded. Thirteen most common and important OARs in H&N radiation treatment planning were considered in this study (Table 1). All OARs were contoured by a single physician and were used to generate the radiation treatment plans with which the patients were treated. In order to quantitatively evaluate the contours generated from the model against the expert's contours, the following metrics were used; DICE similarity coefficient (DSC),

$$DSC = \frac{2(V_{gt} \cap V_{auto})}{V_{gt} + V_{auto}}$$

Jaccard similarity coefficient (JSC),

$$JSC = \frac{V_{gt} \cap V_{auto}}{V_{gt} \cup V_{auto}}$$

and volumetric difference (VD)

$$VD = \frac{V_{auto} - V_{gt}}{V_{gt}}$$

where V_{gt} and V_{auto} are the volume of organs of ground truth and auto-segmented contour.

A commercial software (StructSure, Standard Imaging Inc.) was also used to validate the contouring variation. The program gives score (0-100) based on the similarity of two contours by penalizing each missing and extra voxel as a function of distance from the primary contours with forgiveness threshold distances;

$$Voxel\ penalty\ metric = A + Bd + Ce^{Dd}$$

where A is a constant penalty, B is a linear penalty according to the distance, C is a base penalty for the exponential function, D is an exponential penalty rate, and d is distance. [8]

2.2. Qualitative assessment

A sample of 20 CTs was selected from the validation CT set. 256 automatically segmented structures were qualitatively evaluated by the single physician on a scale of 0-5; 0: ideal, exactly what the ground truth would be, 1: acceptable, no edits necessary, 2: too big, needs minor edits, 3: too small, needs minor edits, 4: not close enough to the ground truth, too big, 5: unacceptable (0-3: clinically useful, 4-5: not clinically useful) (Table 2).

3. Results

Mean DSC values ranged from 0.48±0.14 for the middle pharyngeal constrictor to 0.89±0.03 for the cerebellum. Similarly, the JSC values ranged from 0.32±0.12 for the middle pharyngeal constrictor to 0.80±0.05 for the cerebellum. VD values ranged from 0.02±0.29 for the right submandibular gland to 17.17±59.40 for the middle pharyngeal constrictor (Table 1).

For 20 CTs with 13 OARs, a total of 256 automatically generated contours were evaluated by a physician to determine if they were clinically applicable (Table 2). None of the structures were evaluated to be ideal or acceptable without some edits, but 98% of contours, 251 out of 256, were still considered clinically useful with minor modifications. Among the clinically useful cases, 242 structures were bigger than the ground truth and the other 9 contours were smaller. Meanwhile, five automatically generated structures, 2% of all, were evaluated to be clinically unacceptable; one got a score of 4, much bigger than ground truth (Figure 1).

The contour graded as 4 was the brainstem (Table 3). When compared to the ground truth contour slice by slice, overall contour lines were posteriorly expanded and shifted with 2 and 4 extra slices at the top and bottom. It had a DSC value of 0.69, StructSure score of 0.00, and VD value of 0.80.

Table 1. A comparison of DLC generated contours to the expert contours.

	V_{gt} (cc)	V_{auto} (cc)	VD (%)	DSC	JSC	StructSure
Spinal cord	16.32±4.33	17.76±4.17	0.13±0.28	0.78±0.08	0.64±0.10	76.2±25.3
Parotid_R	26.80±9.40	32.47±10.1	0.24±0.22	0.79±0.06	0.66±0.07	88.0±5.9
Parotid_L	27.56±8.94	30.65±9.83	0.13±0.20	0.81±0.06	0.68±0.08	89.7±4.4
SPC[a]	8.92±2.55	9.78±2.23	0.13±0.23	0.61±0.07	0.44±0.07	60.4±21.6
MPC[b]	1.81±0.68	3.27±1.41	17.17±59.40	0.48±0.14	0.32±0.12	28.6±32.5
IPC[c]	8.38±2.48	8.85±2.86	0.06±0.20	0.74±0.09	0.60±0.10	76.3±23.6
Larynx	19.88±7.08	20.32±6.20	0.04±0.19	0.82±0.12	0.71±0.13	86.7±15.7
SMG_R[d]	8.73±2.46	8.66±2.72	0.02±0.29	0.76±0.10	0.62±0.12	85.7±9.4
SMG_L[e]	9.00±2.38	8.73±2.32	0.04±0.60	0.78±0.11	0.64±0.12	82.9±20.0
Brainstem	27.57±3.04	37.37±4.15	0.37±0.17	0.79±0.04	0.65±0.05	63.9±21.7
Mandible	57.48±12.45	73.10±16.57	0.27±0.10	0.86±0.04	0.75±0.05	69.7±30.2
Cerebellum	125.92±14.19	137.47±12.96	0.10±0.12	0.89±0.03	0.80±0.05	85.3±9.8
Oral cavity	211.55±45.23	233.99±44.90	0.13±0.23	0.84±0.07	0.73±0.10	65.4±22.7

[a] superior pharyngeal constrictor
[b] middle pharyngeal constrictor
[c] inferior pharyngeal constrictor
[d] right submandibular gland
[e] left submandibular gland

Table 2. Distribution of scores evaluated by a physician.

Scale			count
0	ideal	clinically useful	0
1	acceptable, no edits necessary		0
2	too big, needs minor edits		242
3	too small, needs minor edits		9
4	not close to ground truth, too big	not clinically useful	1
5	Unacceptable		4

Three of the lowest-rated structures were in the oral cavity. The oral cavity is an OAR with the largest volume in H&N cases with complex structures, thereby carries a larger element of risk of variation in contour. One of the three oral cavity contours had irregular shapes, with large missing (67.66cc) and extra (83.59cc) volume. VD was small, 0.07, but less important in this case as the VD formula only considers the quantitative difference between the volumes, not the spatial concordance. On the other hand, the StructSure score, which gives a penalty on the distance between each voxel, was 0.00. The other two had DSC>0.7 (0.81 and 0.80), but the StructSure scores were 32.03 and 49.96 (median

70.05) as one was missing 11 slices of contours (VD=-0.22) and the other had irregular shapes. The other lowest-rated contour was a left submandibular gland. In this case, the structure was adjacent to the target and DLC generated contour had 1 and 11 extra contoured slices at the top and bottom. Therefore V_{auto} was 5.7 times larger than the V_{gt}; VD=4.67. DSC value and StructSure score were 0.24 and 0.00, respectively, which showed that the two contours were substantially different.

Table 3. Details of not clinically useful DLC generated contours.

Score	Organ	V_{gt} (cc)	V_{auto} (cc)	VD (%)	Missing (cc)	Extra (cc)	DSC	StructSure Score
4	**Brainstem**	29.27	52.70	0.80	0.87	24.30	0.69	0.00
5	**SMG_L**	2.12	12.01	4.67	4.67	0.43	0.24	0.00
	Oral cavity	223.91	175.73	0.07	67.66	83.59	0.67	0.00
	Oral cavity	298.66	284.13	-0.22	62.49	14.31	0.81	32.03
	Oral cavity	223.91	239.84	-0.05	66.23	51.70	0.80	49.96

In most of the structures, except for the pharyngeal constrictors, DSC values agreed well with the score given by the physician. Only two out of 197 contours were evaluated as clinically useful when DSC values were below 0.7 (0.68 and 0.62). However, DSC value ≤0.4 was generally considered as a large variation, therefore those two contours also required expert evaluation. Thus, qualitative assessment substantiated the reliability of DSC by demonstrating the compatibility between the expert's evaluation and DSC values.

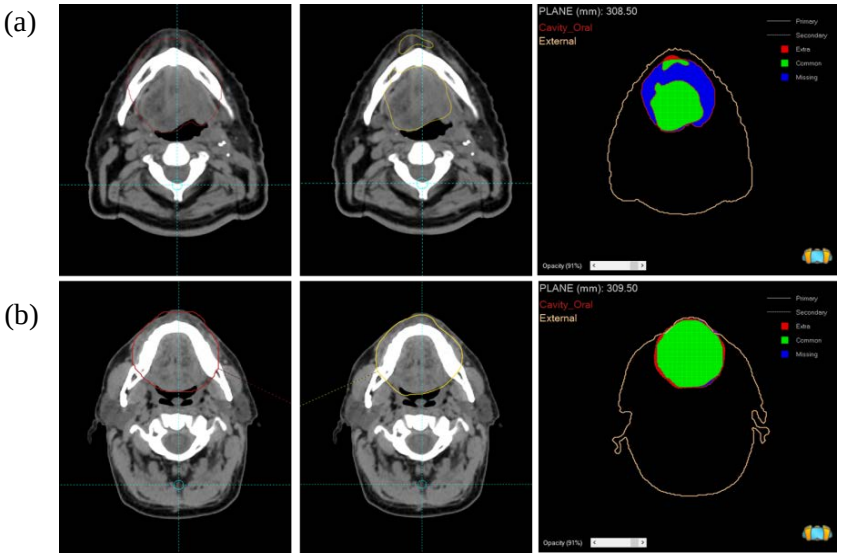


Figure 1. Comparison between the ground truth and DLC generated contour of the oral cavity. Red, green and blue areas on the left figure represent extra, common and missing volume, respectively. (a) An example of a not clinically useful case (b) a clinically useful case

4. Conclusions

A DSC value ≥0.7 indicated good concurrence between automated segmentation and expert contours. [9] DSC for middle pharyngeal constrictor had the lowest value of all the contours. This may be because of the small volume of this structure. In the majority of cases, DSC showed good agreement with the expert's judgment of the quality of the auto-segmented structures. In some cases of discordance, other metrics substantiated the qualitative evaluation as different metrics take different

factors into account. Due to the complexity of the appearance and shape of anatomical structures, auto-segmentation remains challenging. But overall, we found that deep learning auto-contouring was a useful tool to speed up the process of contouring in radiotherapy treatment planning.

5. References

[1] Ayman E *et al* 2011 *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies* vol 1 ed Jasjit S
[2] Tim L *et al* 2018 *Radiother. Oncol.* **126** 312-7
[3] Abhinav C and Bharat D 2012 *Int. J. Comput. Eng. Sci.* **2** 819-22
[4] Li L *et al* 2019 *Radiology* **291** 677-86
[5] Nan B *et al* 2019 *Font. Oncol.* **9** 1192
[6] Luca B *et al* 2019 *Font. Oncol.* **9** 977
[7] Berkman S *et al* 2019 *Med. Phys.* **46** 1
[8] Abhirami H *et al* 2012 *J Med. Imaging and Radiat. Oncol.* **56** 679-88
[9] Dolz J *et al* 2016 *Med. Phys.* **43** 2569