BJR, 2024, **97**, 13–20
https://doi.org/10.1093/bjr/tqad018
Advance access publication: 12 December 2023
Review

Deep learning for automated segmentation in radiotherapy: a narrative review

Jean-Emmanuel Bibault , MD, PhD[1,2,*] and Paul Giraud, MD[2,3]

[1]Radiation Oncology Department, Georges Pompidou European Hospital, Assistance Publique—Hôpitaux de Paris, Université de Paris Cité, Paris, 75015, France
[2]INSERM UMR 1138, Centre de Recherche des Cordeliers, Paris, 75006, France
[3]Radiation Oncology Department, Pitié Salpêtrière Hospital, Assistance Publique—Hôpitaux de Paris, Paris Sorbonne Universités, Paris, 75013, France

*Corresponding author: Jean-Emmanuel Bibault, MD, PhD, Radiation Oncology Department, Hôpital Européen Georges Pompidou, 20 rue Leblanc, Paris, France (jean-emmanuel.bibault@aphp.fr)

Abstract

The segmentation of organs and structures is a critical component of radiation therapy planning, with manual segmentation being a laborious and time-consuming task. Interobserver variability can also impact the outcomes of radiation therapy. Deep neural networks have recently gained attention for their ability to automate segmentation tasks, with convolutional neural networks (CNNs) being a popular approach. This article provides a descriptive review of the literature on deep learning (DL) techniques for segmentation in radiation therapy planning. This review focuses on five clinical sub-sites and finds that U-net is the most commonly used CNN architecture. The studies using DL for image segmentation were included in brain, head and neck, lung, abdominal, and pelvic cancers. The majority of DL segmentation articles in radiation therapy planning have concentrated on normal tissue structures. *N*-fold cross-validation was commonly employed, without external validation. This research area is expanding quickly, and standardization of metrics and independent validation are critical to benchmarking and comparing proposed methods.

Keywords: machine learning; deep learning; radiation oncology; segmentation; contouring; delineation.

Introduction

Radiation therapy is a cornerstone of cancer treatment. Imaging is crucial in radiation therapy to stage patients, define volumes for treatment planning, and assess treatment outcomes. Radiotherapy planning is a lengthy process that requires optimizing the placement of radiotherapy beams to ensure sufficient dose coverage of the tumour while minimizing exposure to surrounding normal tissues, known as organs at risk (OAR). Automated segmentation of tumour and OAR volumes has shown promise in streamlining this process and has the potential to transform the workflow of radiation therapy planning. Computer vision and machine learning have a long history in the automation of segmentation tasks at multiple stages of cancer-related medical workflows. Traditional (non-learning based) segmentation algorithms, based on atlases, aim to distinguish abnormalities from the normal anatomical structures based on features such as intensity distributions, textures, and shape. The focus for improving these algorithms has been on defining better features or feature combinations based on knowledge of anatomy and physiology. In the last decade, deep learning (DL) techniques have been used for automatic contouring. Using DL could aid in improving contouring quality, decreasing interobserver variability, and reducing the time required for treatment planning.[1] Several experimental approaches and commercially available software have been proposed in the literature to assist clinicians.[2]

In this review, we will explain the technical basis of DL for the clinician. We will then review the existing studies published that assessed the performances of DL-based automatic segmentation in brain, head and neck, lung, abdominal, and pelvic tumours.

Methods

To perform this literature review on DL-based automatic contouring, we conducted a search on PubMed/Medline using the keywords "radiotherapy" and "deep learning," "segmentation," "contouring," and "delineation." The search was carried out in March 2023 and focused on studies published between 1997 and 2023. In addition, the reference lists of relevant articles were hand-searched to identify any additional studies of interest.

We assessed the relevance of the search results based on the title, abstract, and full text if necessary. To be included, studies had to be published as full articles in English and employ a DL technique in the field of radiation oncology for segmentation. Studies were excluded if they were written in a language other than English, did not use a DL technique, were not relevant to radiotherapy, lacked patient data, or did not have a clinical application focus. After screening, 38 studies were included in the analysis.

Results

What is DL?

A deep neural network is a type of artificial neural network that is composed of multiple layers of interconnected nodes,

Received: 18 May 2023; **Revised:** 10 August 2023; **Accepted:** 27 October 2023

or neurons. Each neuron takes in multiple inputs, performs a calculation, and produces an output. The outputs of neurons are then used as inputs to the next layer, and so on, until the final output (ie, segmentation, classification) is produced. The process of training a deep neural network involves adjusting the strengths of the connections between neurons so that the network can learn to recognize patterns in the input data. This is done using a process called backpropagation, where the network is fed a training example and the output is compared to the desired output. The difference between the two is used to adjust the weights of the connections in the network, with the goal of minimizing the difference over the entire training set.

The architecture of a deep neural network can vary widely depending on the task it is being used for. For example, a convolutional neural network (CNN) is often used for image recognition tasks, while a recurrent neural network is often used for sequence prediction tasks. One of the key benefits of deep neural networks is their ability to automatically learn from the input data. In traditional machine learning approaches, features must be manually extracted from the data and provided as input to the model. However, in a deep neural network, the features are learned automatically as the network is trained on the data. This allows the network to learn more complex and abstract representations of the input data, which can lead to better performance on the task at hand.

Overall, deep neural networks are a powerful tool for solving a wide range of machine learning tasks, from image recognition (image classification, segmentation, object detection) to natural language processing. However, training and tuning these networks can be a challenging and computationally intensive process, and careful attention must be paid to issues such as overfitting (where a model learns the training data too well, capturing noise and random fluctuations, rather than the actual underlying patterns) and data bias.

DL has numerous applications in radiation oncology, such as image segmentation and detection, image phenotyping and radiomic signature discovery, clinical outcome prediction, image dose quantification, dose-response modelling, radiation adaptation, and image generation. CNN is the most widely used DL technique because they require few parameters since convolutions are invariant by translation of the input. Since 2017, modern DL architectures, such as CNN and auto-encoder, have been increasingly utilized in radiation oncology studies compared to older neural networks such as deep belief network and fully connected neural network. Although diverse, DL applications share a similar framework. A dataset D: $\{X, Y\}$ is created, consisting of training examples X and their labels Y. The goal is to predict Y given X with an estimation function f: $X \rightarrow Y$ that is effectively implemented by the network. To detect overfitting, the dataset is split into training, validation, and test sets, with cross-validation used for smaller datasets. The process can be broken down into the following steps:

1. Data preparation

The first step in DL-based segmentation is to prepare the training data. This typically involves manually segmenting a set of images and using them as the ground truth labels for the training set. The training images are usually preprocessed to ensure consistency in size, orientation, and pixel values.

2. Network architecture

Next, a suitable architecture is selected. There are many types of neural network architectures available, and the choice will depend on the specific task and dataset. In general, deep CNNs with many layers have been shown to perform well on medical image segmentation tasks.

3. Training and validation

The network is trained on the prepared training dataset. During training, the network learns to map the input images to the corresponding segmentation labels. The process of training involves iteratively adjusting the weights of the network to minimize the difference between the predicted and actual labels. This is done using a loss function, such as dice loss or cross-entropy, which measures the difference between the predicted and actual labels.

Validation is a step that occurs during the training process. It is typically done on a separate portion of the dataset, which is distinct from the training set and is used to fine-tune hyperparameters and monitor the model's performance during training. Monitoring performance on the validation set helps prevent overfitting. By assessing how well the model generalizes to data it has not seen during training, it is possible to make adjustments like changing the learning rate or adjusting the model architecture.

4. Testing

Testing is the final evaluation step to assess how well the trained model performs on unseen, completely independent data that it has never encountered during training or validation. The test set is crucial for estimating the model's real-world performance and generalization ability.

DL-based segmentation typically involves using CNNs, which are a class of neural networks that have been specifically designed to work with image data. Within that class, U-Net is the most widely used type of networks. The U-Net architecture is a powerful tool for image segmentation tasks, particularly in the biomedical field, where it has been used for various applications such as brain tumour segmentation, cell segmentation, and organ segmentation. The combination of the encoder and decoder with skip connections enables the network to effectively capture low-level and high-level features, leading to accurate segmentation results. It was proposed by Ronneberger et al[3] in 2015, and it has been shown to achieve state-of-the-art performance in many segmentation tasks. The U-Net architecture is composed of two parts: an encoder and a decoder. The encoder consists of a series of convolutional layers that downsample the input image, while the decoder consists of a series of upsampling layers that produce the segmentation map. The architecture is named after its U-shape, where the encoder and decoder are connected by a bottleneck layer.

In most studies, segmentation performances are reported using Dice similarity coefficient (DSC), which is a measure of overlap between two sets of contours (A being the automatic segmentation and B the ground truth segmentation). The DSC is calculated as the area of overlap between the contours divided by their mean area.

$$\frac{2 \times |A \cap B|}{|A| + |B|}$$

The DSC ranges from 0 to 1, where 0 indicates no overlap between the analysed structures and 1 indicates complete overlap. Other metrics have been proposed to assess the performances of automatic contouring algorithm, such as Added Path Length and Surface DSC to better represent clinical usefulness.[4]

Brain

Table 1 summarizes the characteristics of the brain auto-contouring studies.[5-9] DL techniques have been extensively studied in diagnostic neuro-radiology for brain tumour and secondary lesion segmentation, but their direct use in radiotherapy is still limited. Recently, Liu et al developed a method for segmenting brain metastases on contrast-enhanced T1w MRI datasets. Their network architecture consisted of four sections: input, convolution, fully connected, and classification sections. The approach was validated on Multimodal Brain Tumor Image Segmentation Challenge (BRATS) data, consisting of 65 patients, and 240 brain metastases patients with T1c MRI scans collected at the University of Texas Southwestern Medical Center.

Liu et al's[5] study reported DSC values of 0.75 ± 0.07 in the tumour core and 0.81 ± 0.04 in the enhancing tumour for the BRATS data, outperforming most techniques in the 2015 Brain Tumor Image Segmentation Challenge. The study also showed that the segmentation results of patient cases had an average DSC of 0.67 ± 0.03 and achieved an area under the receiver operating characteristic curve of 0.98 ± 0.01. Charron et al also used a similar approach by adapting an existing 3D CNN algorithm, Deep-Medic, to detect and segment brain metastases on MRI scans of patients undergoing stereotactic treatments. The dataset consisted of 182 patients with three MRI modalities (T1w3D, T2w2D, and T1w2D) and was split into training, validation, and test sets. The benchmark segmentation was carried out manually by up to four radiation oncologists and compared to the DL output. The results obtained were promising and indicated the potential application of DL techniques in the identification and segmentation of brain metastases on multimodal MR images.[6]

Head and neck

Table 2 summarizes the characteristics of the head and neck auto-contouring studies.[10-19] Segmenting images of head and neck malignancies is a challenging and time-consuming task in the field of radiotherapy. This complexity can impede the progress of adaptive approaches in this area.[20]

The normal anatomy of the head and neck region can be significantly altered by the presence of large primary or nodal lesions, as well as by surgical procedures. As a result, manual segmentation can be a time-consuming and challenging task, as automated segmentation methods may not be able to handle these anatomical variations. However, DL techniques and prior knowledge can be leveraged to address these difficulties and improve the accuracy of image segmentation in this area. CNNs have been utilized to improve the accuracy and speed of organs at risk (OAR) delineation in head and neck cancer patients. The structure of a typical CNN involves repeating blocks, each containing a convolutional layer, a batch normalization layer, a rectified linear unit activation layer, a dropout layer, and a pooling layer. Ibragimov and Xing employed a tri-planar patch-based network with these repeating blocks on 50 patients who were scheduled for head and neck radiotherapy. The performance of the CNN was comparable or even better than the reference segmentation for various organs, including the spinal cord, mandible, parotid glands, larynx, pharynx, eye globes, and optic nerves. However, the results for the sub-mandibular glands and optic chiasm were less satisfactory due to their size and location.[15]

Men and colleagues have used a deep deconvolutional neural network for the accurate delineation of nasopharyngeal gross tumour volume (GTV), metastatic lymph node GTV, clinical target volumes (CTVs), and OAR in planning CT images of 230 patients diagnosed with stage I or II nasopharyngeal carcinoma. The study demonstrated that DL techniques can enhance the consistency of contouring performance and optimize radiotherapy treatment workflow when compared to other segmentation methods.[14]

Cardenas et al have presented a novel approach for auto-delineation of high-risk CTVs for head and neck tumours using deep neural networks. Their study showcases the potential of this technique in reducing variability in target design and improving clinical practice for radiation oncologists. This approach saves time and provides more reliable data for multi-institutional studies where clinical practices may vary widely. The results of the study demonstrate that the proposed DL-based approach can produce comparable results to inter- and intraobserver studies for manual delineation of these complex volumes, with DSC values ranging from 0.62 to 0.90 and a median mean surface distance of 2.75 mm.[21]

Thorax

Table 3 summarizes the characteristics of the thorax auto-contouring studies.[22-27] The auto-segmentation of the thoracic site using traditional semi-automatic tools has demonstrated good overall performance, with DSC values surpassing 0.9 when compared to manual benchmarking. This

Table 1. Summary of the study characteristics by the anatomical region in the brain.

Publication	Year	Image modality	Patients/plans	Delineation type	Outcome
Liu et al[5]	2017	MRI	490	Tumour	Mean DSC = 0.75
Charron et al[6]	2018	MRI	182	Tumour	Mean DSC = 0.77
Naceur et al[7]	2018	MRI	285	Tumour	Mean DSC = 0.88
Deng et al[8]	2019	MRI	100	Tumour	Mean DSC = 0.91
Sun et al[9]	2019	MRI	384	Tumour	Mean DSC = whole tumour (0.84), tumour core (0.72)

Abbreviation: DSC = Dice similarity coefficient.

Table 2. Summary of the study characteristics by the anatomical region in head and neck.

Publication	Year	Image modality	Patients/ plans	Cancer Site	Delineation type	Outcome
Liang et al[10]	2019	CT	185	NPC	OARs	DSC = 0.86
Lin et al[11]	2019	MRI	1021	NPC	GTV	DSC = 0.79
van Rooij t al.[12]	2019	CT	157	H&N	OARs	DSC = 0.60 (oesophagus) to 0.83 (right parotid)
Chan et al[13]	2019	CT	200	H&N	OARs	DSC = 0.84 (left temporomandibular joint) to 0.91 (mandible)
Men et al[14]	2019	CT	100	H&N	OARs	Mean DSC = 0.90
Ibragimov et al[15]	2017	CT	50	H&N	OARs	DSC = 0.37 (optic chiasm) to 0.89 (mandible)
Zhu et al[16]	2018	CT	271	H&N	OARs	Mean DSC = 0.79%
Men et al[14]	2017	CT	230	NPC	GTVn and CTV	Mean DSC: GTVnx (0.81), GTVnd (0.62), CTV (0.82)
Cardenas et al[17]	2018	CT	285	Oropharyngeal	CTV	DSC > 0.75 on 96% of the cases
van Dijk et al[18]	2020	CT	589	H&N	OARs	DSC: atlas vs CNN (0.59 vs 0.74)
Zhong et al[19]	2019	CT	140	NPC	OARs	Mean DSC = optic nerves (0.89) to thyroids (0.92)

Abbreviations: CNN = convolutional neural network; CTV = clinical target volume; DSC = Dice similarity coefficient; GTVn = Growth tumour volume; GTVnd = metastatic lymph node tumour total volume; GTVnx = nasopharyngeal tumour total volume; H&N = head and neck; NPC = nasopharyngeal carcinoma; OARs = organ at risks.

Table 3. Summary of the study characteristics by the anatomical region in the thorax.

Publication	Year	Image modality	Patients/ plans	Cancer Site	Delineation type	Outcome
Men et al[22]	2018	CT	800	Breast	CTV	DSC = Right-sided breast cancer (0.91); left-sided breast cancer (0.91)
Bi et al[23]	2019	CT	250	Lung	CTV	DSC = 0.75
Fechter et al[24]	2017	CT	50	Thorax	Oesophagus	Mean DSC = 0.76
Yang et al[25]	2018	CT	60	Thorax	OARs	DSC = 0.72 (oesophagus) to 0.97 (left lung)
Liu et al[26]	2021	CT	110	Breast	CTV	DSC = 0.90
Primakov et al[27]	2022	CT	1328	Lung	GTV	DSC = 0.82

Abbreviations: CTV = clinical target volume; DSC: Dice similarity coefficient; GTV = total tumour volume; OARs = organ at risks.

success can be largely attributed to the unique anatomy of this site, which exhibits naturally high contrast at the air/tissue interfaces.[28] Despite these good performances, some DL approaches have been proposed to further enhance radiotherapy-oriented auto-segmentation performance. These approaches add to the significant radiological evidence in this field, which has traditionally focused on nodule classification and conventional computer-aided diagnosis support.[1,29,30]

Lustberg and colleagues conducted a study on the time-saving potential of using software-generated contouring as a starting point for manual segmentation in thoracic organ segmentation. They tested a commercially available atlas-based software and a CNN for the segmentation of thoracic OAR (including lungs, oesophagus, spinal cord, heart, and mediastinum) in 20 CT scans of stage I-III lung cancer patients. They found that using user-adjusted software-generated contours as a starting point for manual segmentation resulted in a significant reduction in contouring time.[31]

In 2022, Primakov et al published a study presenting a fully automated pipeline for detecting and volumetrically segmenting non-small-cell lung cancer, validated on 1328 thoracic CT scans from 8 different institutions. An *in silico* prospective clinical trial demonstrated that the proposed method is faster and more reproducible compared to human experts. In addition, radiologists and radiation oncologists preferred the automatic segmentations in 56% of the cases on average.[27]

Abdomen

Table 4 summarizes the characteristics of the abdominal region auto-contouring studies.[32-36] Auto-segmentation software, including those based on DL techniques, faces challenges in the abdomen due to the high anatomical variability in this region. Factors such as the displacement of hollow organs and bowel loops, as well as interpatient variability, limit the efficacy of auto-segmentation, resulting in relatively poor results. However, the liver is a promising candidate for auto-segmentation applications in the abdomen, as it tends to have a more regular shape and position. Ibragimov and colleagues proposed an approach for segmenting the intrahepatic portal vein (PV) as part of the planning process for stereotactic body radiation therapy (SBRT).[32] This approach represents the first attempt at using DL techniques for PV segmentation and has the potential to improve the accuracy and efficiency of SBRT planning in liver cancer patients. Since the PV is a critical structure that needs to be spared during radiation therapy, accurate segmentation is crucial for minimizing the risk of complications and improving treatment outcomes. Segmenting the PV in planning images can be difficult due to poor visibility caused by artefacts, fiducials, stents, or variable anatomy, even for experienced operators. However, DL-based segmentation algorithms have satisfactory results with DSC ranging from 0.7 to 0.83 when compared to manual segmentation benchmarks.

Table 4. Summary of the study characteristics by the anatomical region in the abdomen.

Publication	Year	Image modality	Patients/ plans	Cancer Site	Delineation type	Outcome
Ibragimov et al[32]	2017	CT	72	Liver	Portal vein	Median DSC = 0.83
Qin et al[33]	2018	CT	100	Liver	Liver	Median DSC = 0.97
Ahn et al[34]	2019	CT	70	Liver	OARs	DSC (atlas) = 0.60 (stomach) to 0.93 (liver) DSC (DL) = 0.73 (stomach) to 0.94 (heart)
Fu et al[35]	2018	MRI	120	Abdominal	OARs	Mean DSC = 0.65 (duodenum) to 0.95 (liver)
Hu et al[36]	2017	CT	132	Abdominal	Liver, spleen, and kidney	Mean DSC = 0.96 (liver), 0.94 (spleen), 0.95 (kidneys)

Abbreviations: DL = deep learning; DSC: Dice similarity coefficient; OARs = organ at risks.

Table 5. Summary of the study characteristics by the anatomical region in the pelvis.

Publication	Year	Image modality	Patients/ plans	Cancer Site	Delineation type	Outcome
Men et al[37]	2017	CT	278	Rectal	CTV, OARs	Mean DSC = 0.87 (CTV), 0.93 (bladder), 0.92 (femoral heads) (92.1), 0.65 (small intestine), 0.62 (colon)
Men et al[38]	2018	CT, MRI	70	Rectal	CTV	DSC = 0.78 and 0.85 for MRI and CT
Trebeschi et al[39]	2017	MRI	140	Rectal	Tumour	DSC = 0.68 and AUC = 0.99
Wang et al[40]	2018	MRI	93	Rectal	GTV	Mean DSC = 0.74
Song et al[41]	2020	CT	199	Rectal	CTV, OARs	Volumetric Dice coefficient: model 1 vs model 2 = 0.88 vs 0.87
Balagopal et al[42]	2018	CT	136	Prostate	CTV, OARs	Mean DSC = 0.9 (prostate), 0.95 (femoral heads), 0.95 (bladder), 0.84 (rectum)
Karimi et al[43]	2019	TRUS images	675	Prostate	CTV	DSC = 0.94
Ju et al[44]	2021	CT	133	Prostate Cervix	CTV	DSC = 0.82
Liu et al[45]	2020	CT	237	Cervix	CTV, OARs	Mean DSC = 0.86 (CTV), 0.91 (bladder), 0.85 (bone marrow), 0.9 (femoral heads), 0.82 (rectum), 0.85 (bowel bag), 0.82 (spinal cord)
Sartor et al[46]	2020	CT	266	Cervix and anorectal	CTV	Cervical cancer: median DSC = 0.93 (femoral heads), 0.84 (bladder), 0.88 (bowel bag), 0.82 (CTV) Anorectal cancer: median DSC = 0.92 (femoral heads), 0.94 (bladder), 0.83 (bowel bag), 0.82 (CTV)
Wang et al[47]	2020	CT	125	Cervical	CTV, OARs	DSC = 0.86 (CTV), 0.91 (bladder), 0.88 (femoral heads), 0.86 (small intestine), 0.81 (rectum)
Zhang et al[48]	2020	CT	91	Cervical	CTV, OARs	DSC = 0.82 (CTV), 0.87 (bladder), 0.8 (small intestine), 0.65 (sigmoid), 0.82 (rectum)

Abbreviations: CTV = clinical target volume; DSC = Dice similarity coefficient; GTV = total tumour volume; OARs = organ at risks; TRUS = transrectal ultrasound.

Pelvis

Table 5 summarizes the characteristics of the main pelvic region auto-contouring studies.[37-48] Several automatic strategies for pelvic auto-segmentation have been proposed in recent years, but there is still room for improvement with the development of more efficient DL techniques.[49,50] Men et al utilized DL for the segmentation of OAR and CTVs in planning CTs of rectal cancer patients in the pelvic region. Their approach achieved a high concordance of 87.7% for target volumes (TVs), with a very fast segmentation speed of 45 s.[14] Trebeschi et al[39] developed a CNN approach for segmenting primary locally advanced rectal cancer lesions on T2- and DWI-MRI images. They reported a DSC of 0.7 and an AUC (Area Under the Curve) of 0.99 for the CNN-generated contours compared to manually obtained contours. Wang et al developed an auto-segmentation model for GTV segmentation on T2 MR images of 93 rectal cancer patients. The model achieved a segmentation performance similar to manual interobserver variability, with a DSC of 0.74.[40] A recent study from Liang et al, introduced a Deep Unsupervised Learning framework based on regional deformable model for automated prostate contour propagation from planning computed tomography to cone-beam CT. The average DSCs between DUL-based prostate contours and reference contours were 0.83 ± 0.04.[51] This method could be used for adaptive radiotherapy of prostate cancer with daily replanning.

Commercial solutions

There are several commercial solutions currently available: MVision (Helsinki, Finland), AutoContour (Radformation, New York, USA), RayStation (RaySearch, Stockholm, Sweden), and Annotate (Therapanacea, Paris, France). A recent study aimed to assess the effectiveness of these auto-segmentation solutions in improving contouring accuracy, reducing variability among observers, and saving time in medical image analysis.[52] The solutions were evaluated on OAR contours for various patient groups, including breast, head and neck, lung, and prostate cases. All AI (Artificial Intelligence) systems produced contours with good quality, with median volumetric DSCs comparable to existing literature. Additionally, significant time savings were observed

across all systems in contouring tasks, ranging from 14 to 93 minutes for different patient groups.

Discussion

Time-saving

Auto-delineation offers the key benefit of reducing the amount of time required for manual delineation of the TV and OARs. For instance, Liang et al[10] employed DL techniques to develop a framework that can automatically segment and detect OARs in nasopharyngeal carcinoma. Using this DL-based framework, a single CT image can be delineated in roughly 30 s. In another study, Chan et al[13] utilized deep lifelong learning to design a CNN algorithm for the automatic segmentation of OARs in the head and neck region. The networks were able to predict all the OARs within 20 s. Similarly, Men et al[37] employed a CNN to automatically segment the clinical TV and OARs in rectal cancer. The time required for segmentation of all the CTV, colon, intestine, right and left femoral heads, and bladder was 45 s per patient during the test phase. However, future studies on the clinical implementation of automatic contouring, beyond accuracy studies based on DSC, should assess the actual time saved.

Intra- and interobserver variability

Besides its time-saving benefits, DL technology can also improve delineation accuracy and standardization across different operators and centres. Lin et al[11] developed and validated an AI contouring tool for automating the main GTV contouring in patients with nasopharyngeal carcinoma. The researchers collected nasopharyngeal MRI data from 1021 patients with nasopharyngeal carcinoma to delineate GTV contours, which were jointly defined by two experts. The study found that the contouring accuracy was improved with the assistance of AI, and the intra- and interpractitioner variabilities were significantly reduced (by 36.4% and 54.5%, respectively) during multicentre evaluation. In the study published by Primakov et al,[27] authors showed that the automatic method stratified patients into low and high survival groups by applying RECIST criteria more consistently than methods based on manual contours.

Future studies should assess contour changes performed by the clinicians to automatic contouring: even if an automatic contouring model has good DSC performances *in silico*, it remains to be demonstrated that physicians will not modify them greatly to their liking afterwards, since they provide the final validation.

Training data quality and ground truth

Despite the promising results of DL-based auto-delineation of TV and OARs, there are still several unresolved challenges in applying DL to clinical practice. DL frameworks need to be trained with a large and diverse range of representative examples to increase their accuracy and reliability in real-world operations. This is difficult because of the scarcity of labelled high-quality data. To this end, solutions that can achieve cross-institutional and international sharing of data should be promoted and adopted. Privacy concerns remain an obstacle for that. Another challenge is the intra and inter-rater variability: different annotators (and even the same annotator) may annotate the TV and OAR differently, which results in inconsistent supervision during the training process. The adoption of international consensus guidelines across centres would help to reduce this variability. The performance evaluation could also introduce some tolerance to reflect that the ground truth annotations are imperfect.

Conclusion

DL has the potential to revolutionize the field of RT and improve patient outcomes. However, it is important to acknowledge that DL is not a panacea, and there are still many challenges that need to be addressed before its widespread implementation in clinical practice. It is crucial to ensure that the DL models are validated thoroughly and meet the highest standards of accuracy and safety. Additionally, it is essential to continue to improve the quality and quantity of annotated datasets to facilitate the training of DL models. As DL technology advances and more research is conducted, we can expect to see more reliable and accurate auto-delineation tools that can be integrated into routine clinical practice.

Funding

None declared.

Conflict of interest

None declared.

References

1. Lee JG, Jun S, Cho YW, et al. Deep learning in medical imaging: general overview. *Korean J Radiol*. 2017;18(4):570-584.
2. Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. *Radiother Oncol*. 2014;112(3):317-320.
3. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham: Springer International Publishing; 2015 :234-241.
4. Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*. eCollection 2020; 13:1-6.
5. Liu Y, Stojadinovic S, Hrycushko B, et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. *PLoS One*. 2017;12 (10):e0185844.
6. Charron O, Lallement A, Jarnet D, Noblet V, Clavier JB, Meyer P. Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput Biol Med*. 2018;95:43-54.
7. Naceur MB, Saouli R, Akil M, Kachouri R. Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images. *Comput Methods Programs Biomed*. 2018; 166:39-49.
8. Deng W, Shi Q, Luo K, Yang Y, Ning N. Brain tumor segmentation based on improved convolutional neural network in combination with non-quantifiable local texture feature. *J Med Syst*. 2019; 43(6):152.
9. Sun J, Chen W, Peng S, Liu B. DRRNet: dense residual refine networks for automatic brain tumor segmentation. *J Med Syst*. 2019; 43(7):221.
10. Liang S, Tang F, Huang X, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur Radiol*. 2019;29(4):1961-1967.

11. Lin L, Dou Q, Jin YM, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology*. 2019;291(3):677-686.
12. van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *Int J Radiat Oncol Biol Phys*. 2019;104(3):677-684.
13. Chan JW, Kearney V, Haaf S, et al. A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning. *Med Phys*. 2019;46(5):2204-2213.
14. Men K, Chen X, Zhang Y, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol*. 2017;7:315.
15. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys*. 2017;44(2):547-557.
16. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys*. 2019;46(2):576-589.
17. Cardenas CE, Anderson BM, Aristophanous M, et al. Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks. *Phys Med Biol*. 2018;63(21):215026. Nov
18. van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol*. 2020;142:115-123.
19. Zhong T, Huang X, Tang F, Liang S, Deng X, Zhang Y. Boosting-based cascaded convolutional neural networks for the segmentation of CT organs-at-risk in nasopharyngeal carcinoma. *Med Phys*. 2019;46(12):5602-5611.
20. Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. *Acta Oncol*. 2016;55(7):799-806.
21. Cardenas CE, McCarroll RE, Court LE, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. *Int J Radiat Oncol Biol Phys*. 2018;101(2):468-478.
22. Men K, Zhang T, Chen X, et al. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. *Phys Med*. 2018;50:13-19.
23. Bi N, Wang J, Zhang T, et al. Deep learning improved clinical target volume contouring quality and efficiency for postoperative radiation therapy in non-small cell lung cancer. *Front Oncol*. 2019;9:1192. https://www.frontiersin.org/articles/10.3389/fonc.2019.01192
24. Fechter T, Adebahr S, Baltas D, Ben Ayed I, Desrosiers C, Dolz J. Esophagus segmentation in CT via 3D fully convolutional neural network and random walk. *Med Phys*. 2017;44(12):6341-6352.
25. Yang J, Veeraraghavan H, Armato IS, et al. Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. *Med Phys*. 2018;45(10):4568-4581.
26. Liu Z, Liu F, Chen W, et al. Automatic segmentation of clinical target volumes for post-modified radical mastectomy radiotherapy using convolutional neural networks. *Front Oncol*. 2020;10:581347. https://www.frontiersin.org/articles/10.3389/fonc.2020.581347
27. Primakov SP, Ibrahim A, van Timmeren JE, et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat Commun*. 2022;13(1):3423.
28. Zhu M, Bzdusek K, Brink C, et al. Multi-institutional quantitative evaluation and clinical validation of Smart Probabilistic Image Contouring Engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography images in the head and neck, thorax, liver, and male pelvis areas. *Int J Radiat Oncol Biol Phys*. 2013;87(4):809-816.
29. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets Ther*. 2015;8:2015-2022.
30. McBee MP, Awan OA, Colucci AT, et al. Deep learning in radiology. *Acad Radiol*. 2018;25(11):1472-1480.
31. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol*. 2018;126(2):312-317.
32. Ibragimov B, Toesca D, Chang D, Koong A, Xing L. Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. *Phys Med Biol*. 2017;62(23):8943-8958.
33. Qin W, Wu J, Han F, et al. Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation. *Phys Med Biol*. 2018;63(9):095017.
34. Ahn SH, Yeo AU, Kim KH, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiat Oncol*. 2019;14(1):213.
35. Fu Y, Mazur TR, Wu X, et al. A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. *Med Phys*. 2018;45(11):5129-5137.
36. Hu P, Wu F, Peng J, Bao Y, Chen F, Kong D. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. *Int J Comput Assist Radiol Surg*. 2017;12(3):399-411.
37. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys*. 2017;44(12):6377-6389.
38. Men K, Boimel P, Janopaul-Naylor J, et al. Cascaded atrous convolution and spatial pyramid pooling for more accurate tumor target segmentation for rectal cancer radiotherapy. *Phys Med Biol*. 2018;63(18):185016. Sep
39. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci Rep*. 2017;7(1):5301.
40. Wang J, Lu J, Qin G, et al. Technical note: a deep learning-based autosegmentation of rectal tumors in MR images. *Med Phys*. 2018;45(6):2560-2564.
41. Song Y, Hu J, Wu Q, et al. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy. *Radiother Oncol*. 2020;145:186-192.
42. Balagopal A, Kazemifar S, Nguyen D, et al. Fully automated organ segmentation in male pelvic CT images. *Phys Med Biol*. 2018;63(24):245015.
43. Karimi D, Zeng Q, Mathur P, et al. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Med Image Anal*. 2019;57:186-196.
44. Ju Z, Guo W, Gu S, et al. CT based automatic clinical target volume delineation using a dense-fully connected convolution network for cervical cancer radiation therapy. *BMC Cancer*. 2021;21(1):243.
45. Liu Z, Liu X, Guan H, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiother Oncol*. 2020;153:172-179.
46. Sartor H, Minarik D, Enqvist O, et al. Auto-segmentations by convolutional neural network in cervical and anorectal cancer with clinical structure sets as the ground truth. *Clin Transl Radiat Oncol*. 2020;25:37-45.
47. Wang Z, Chang Y, Peng Z, et al. Evaluation of deep learning-based auto-segmentation algorithms for delineating clinical target volume and organs at risk involving data for 125 cervical cancer patients. *J Appl Clin Med Phys*. 2020;21(12):272-279.
48. Zhang D, Yang Z, Jiang S, Zhou Z, Meng M, Wang W. Automatic segmentation and applicator reconstruction for CT-based brachytherapy of cervical cancer using 3D convolutional neural networks. *J Appl Clin Med Phys*. 2020;21(10):158-169.

49. Gambacorta MA, Boldrini L, Valentini C, et al. Automatic segmentation software in locally advanced rectal cancer: READY (REsearch program in Auto Delineation sYstem)-RECTAL 02: prospective study. *Oncotarget*. 2016;7(27):42579-42584.
50. Gambacorta MA, Valentini C, Dinapoli N, et al. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system. *Acta Oncol*. 2013;52(8):1676-1681.
51. Liang X, Bibault JE, Leroy T, et al. Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning. *Med Phys*. 2021;48(4):1764-1770.
52. Doolan PJ, Charalambous S, Roussakis Y, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Front Oncol*. 2023 [cited 2023;13:1213068. https://www.frontiersin.org/articles/10.3389/fonc.2023.1213068