Technical Innovations & Patient Support in Radiation Oncology 26 (2023) 100211

Contents lists available at ScienceDirect

Technical Innovations & Patient Support in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/technical-innovations-and-patient-support-in-radiation-oncology

Clinical evaluation of a deep learning segmentation model including manual adjustments afterwards for locally advanced breast cancer

Nienke Bakx [a], Dorien Rijkaart [a], Maurice van der Sangen [a], Jacqueline Theuws [a], Peter-Paul van der Toorn [a], An-Sofie Verrijssen [a], Jorien van der Leer [a], Joline Mutsaers [a], Thérèse van Nunen [a], Marjon Reinders [a], Inge Schuengel [a], Julia Smits [a], Els Hagelaar [a], Dave van Gruijthuijsen [a], Hanneke Bluemink [a], Coen Hurkmans [a,b,*]

[a] Catharina Hospital, Department of Radiation Oncology, Eindhoven, the Netherlands
[b] Technical University Eindhoven, Faculties of Physics and Electrical Engineering, Eindhoven, the Netherlands

ARTICLE INFO

Keywords:
Deep learning
Auto-segmentation
Breast cancer
Radiotherapy

ABSTRACT

Introduction: Deep learning (DL) models are increasingly developed for auto-segmentation in radiotherapy. Qualitative analysis is of great importance for clinical implementation, next to quantitative. This study evaluates a DL segmentation model for left- and right-sided locally advanced breast cancer both quantitatively and qualitatively.

Methods: For each side a DL model was trained, including primary breast CTV (CTVp), lymph node levels 1–4, heart, lungs, humeral head, thyroid and esophagus. For evaluation, both automatic segmentation, including correction of contours when needed, and manual delineation was performed and both processes were timed. Quantitative scoring with dice-similarity coefficient (DSC), 95% Hausdorff Distance (95%HD) and surface DSC (sDSC) was used to compare both the automatic (not-corrected) and corrected contours with the manual contours. Qualitative scoring was performed by five radiotherapy technologists and five radiation oncologists using a 3-point Likert scale.

Results: Time reduction was achieved using auto-segmentation in 95% of the cases, including correction. The time reduction (mean ± std) was 42.4% ± 26.5% and 58.5% ± 19.1% for OARs and CTVs, respectively, corresponding to an absolute mean reduction (hh:mm:ss) of 00:08:51 and 00:25:38. Good quantitative results were achieved before correction, e.g. mean DSC for the right-sided CTVp was 0.92 ± 0.06, whereas correction statistically significantly improved this contour by only 0.02 ± 0.05, respectively. In 92% of the cases, auto-contours were scored as clinically acceptable, with or without corrections.

Conclusions: A DL segmentation model was trained and was shown to be a time-efficient way to generate clinically acceptable contours for locally advanced breast cancer.

Introduction

The process of radiotherapy (RT) treatment planning contains several steps, which are partly performed manually. One of these steps is segmentation of target contours and surrounding organs at risk (OARs). Several studies showed variability in delineation of target and OARs delineation for breast cancer, even in the presence of delineation atlases [1,2]. Moreover, manual delineation is a time consuming process. Therefore, several studies introduced automatic segmentation of contours. Previously, these segmentation delineation models were atlas-based, and several studies regarding delineation for breast cancer were published [3–5]. In recent years, the use of deep learning (DL) for automatic delineation increased [6–12]. For delineation of contours for breast cancer RT, several DL models are deployed and promising results are shown. Most studies limit the analysis to a quantitative score, however for successful clinical implementation a qualitative end-user scoring is of great importance [13,14]. This qualitative scoring includes assessment by end-users of the DL models to validate its use in a

Abbreviations: 95%HD, 95% Hausdorff distance; CTV, clinical target volume; CTVp, primary CTV; DL, deep learning; DSC, Dice similarity coefficient; OARs, organs at risk; RO, radiation oncologist; RT, radiotherapy; RTT, radiotherapy technologist; sDSC, surface DSC.

* Corresponding author at: Catharina Hospital, Department of Radiation Oncology, Eindhoven, the Netherlands.
E-mail address: coen.hurkmans@catharinaziekenhuis.nl (C. Hurkmans).

https://doi.org/10.1016/j.tipsro.2023.100211
Received 10 February 2023; Received in revised form 23 April 2023; Accepted 9 May 2023
Available online 13 May 2023

real-world clinical setting. Besides, the end-users will correct contours when found as not directly clinically acceptable, and a quantitative assessment is performed to compare the contours before and after correction. These measurements are additional to measurements performed in comparable studies [8,12]. Moreover, both the manual contouring as automatic contouring, including correction time, are timed since time efficiency is the primary endpoint, while still preserving quality. In addition, consistency is considered as an important endpoint. In this study, the performance of a DL segmentation model, trained separately for both left- and right-sided locally advanced breast cancer, including target and OARs volumes, is assessed. Both quantitative and qualitative scoring of the contours are included, as well as time efficiency.

Materials and methods

Patients

The patient dataset used in this study consists of a training set for the development of the DL models and an independent test set used for the quantitative and qualitative evaluation. All patients were treated for locally advanced breast cancer between February 2017 and May 2022. After surgery of the breast and axilla, radiotherapy of the breast including axillary node levels 1 and 2 or levels 1 to 4 was indicated. Data was anonymized according to Dutch data protection and privacy legislation. The training set contains 80 patients for both the left- and right-sided model, thus 160 patients in total. The test set contains 10 patients for each side. Of these 20 patients, respectively four and seven patients contained node levels 1 to 4 for the left- and right-sided model, where the other patients only contained node levels 1 and 2. Contouring was performed following the ESTRO guidelines [15,16], where targets were delineated by radiation oncologists (ROs), and OARs by radiotherapy technologists (RTTs). Experience of both ROs and RTTs differed. All contours were checked by an experienced RO before inclusion in either the training or test set. Patients were excluded from training or evaluation when contours differed from the guidelines due to the clinical nature, for example nodal tissue which was visible on PET-CT and included in the nodal target volume beyond the standard atlas boundary.

Deep learning model

The DL model framework is developed by RaySearch and training is performed in RayStation 9B, while testing is performed in version 10B-SP1 (RaySearch laboratories, AB, Sweden). The DL network used is a 3D U-net, based on the architecture described by Çiçek et al. [17]. One of the distinctive characteristics of a U-net is its ability to combine image features on different levels of resolution, which makes it suitable for segmentation. Data augmentation is performed during training to synthetically generate new data, to improve model performance. A more detailed description of the model and training procedure can be found in the supplementary materials.

Evaluation

Targets and OARs of the patients in the test set were contoured manually during clinical practice and thus delineated by one RTT and RO. Afterwards, the automatic segmentation was performed. Manual corrections could then be applied if needed. Both segmentation processes were timed. For full manual segmentation, this time included loading the structure template, delineation of the OARs by the RTTs and a check of the OARs delineation and delineation of the target volumes by the ROs. For automatic segmentation, the time needed to run the segmentation model, and to check and correct the contours by the corresponding user was measured. The outcomes were scored both quantitatively and qualitatively. For quantitative scoring, three metrics were used to compare the manual contours to both uncorrected and corrected auto-contours: (1) Dice Similarity Coefficient (DSC), (2) surface DSC (sDSC) [18] and (3) 95% Hausdorff Distance (95%HD). Significance in performance between the models for both sides was investigated with the Wilcoxon rank-sum test, whereas the Wilcoxon signed rank test was used to investigate differences between the uncorrected and corrected contours. For both tests, a p-value of 0.05 or lower was considered statistically significant.

Qualitative scoring was measured by using a 3-point Likert scale for each contour:

1. Clinically acceptable, no correction is needed
2. Not clinically acceptable, but can be used as a starting point to create a clinically acceptable contour while still saving time
3. Not clinically acceptable, and cannot be used as a starting point to create a clinically acceptable contour

For qualitative scoring, anonymized patients of the test set were divided over the same group of five RTTs and five ROs, which performed the manual segmentation of the test set. To prevent bias, automatic segmentation and correction of corresponding patients was always performed by different RTTs and ROs than the those who performed the manual segmentation. Moreover, manual contours were hidden during the qualitative scoring.

To assess possibility of clinical implementation, the primary endpoint is time reduction. When leading to a mean reduction of time, the DL model will be considered as successful for clinical use.

Results

Time saving

Fig. 1 shows the time needed for each patient for manual and automatic segmentation, including corrections when needed. The mean time (hh:mm:ss) for manual delineation was 00:17:05 and 00:41:31 for respectively the OARs and CTVs. While using auto-segmentation, the total time spent including correction was 00:08:47 and 00:15:43 for OARs and CTVs, resulting in a reduction of 00:08:51 and 00:25:38, respectively. Only for one patient case out of 20, the time needed to correct the OARs took more time than the manual delineation (00:12:40 vs 00:10:15), leading to a decrease in time in 95% of the patients.

Quantitative evaluation

The automatically generated structures were quantitatively compared with the manually generated structures, of which the results are visualized in Fig. 2. The results can also be found in Table S2 of the supplementary materials. As can be observed, some outliers are present for the CTVp and CTVn1 for all three metrics. When statistically comparing the two models, a significant difference was found for both lungs for all metrics. However, as these quantitative metrics show a good performance, these differences are not considered as clinically relevant. Moreover, for the 95% HD and sDSC a significant difference was found for the esophagus, where the model for the right side outperforms the left-sided model. However, after visual inspection, it was observed that differences between manual and automatic contours were mostly due to a difference in length. When only considering the overlapping parts, a 95% HD of 3.78 $\pm$ 2.05 mm and 3.27 $\pm$ 1.38 mm were found for respectively the left- and right-sided model, which were not significantly different. Also for the sDSC, the scores were not significantly different, with mean scores of 0.96 $\pm$ 0.04 and 0.97 $\pm$ 0.04. The differences between the sDSC score of the heart also decreased when only considering overlapping parts, but were still statistically significant (0.89 $\pm$ 0.08 vs 0.78 $\pm$ 0.10).

The impact of the corrections made was also quantitatively measured, by calculating the metrics for both the automatically (not-

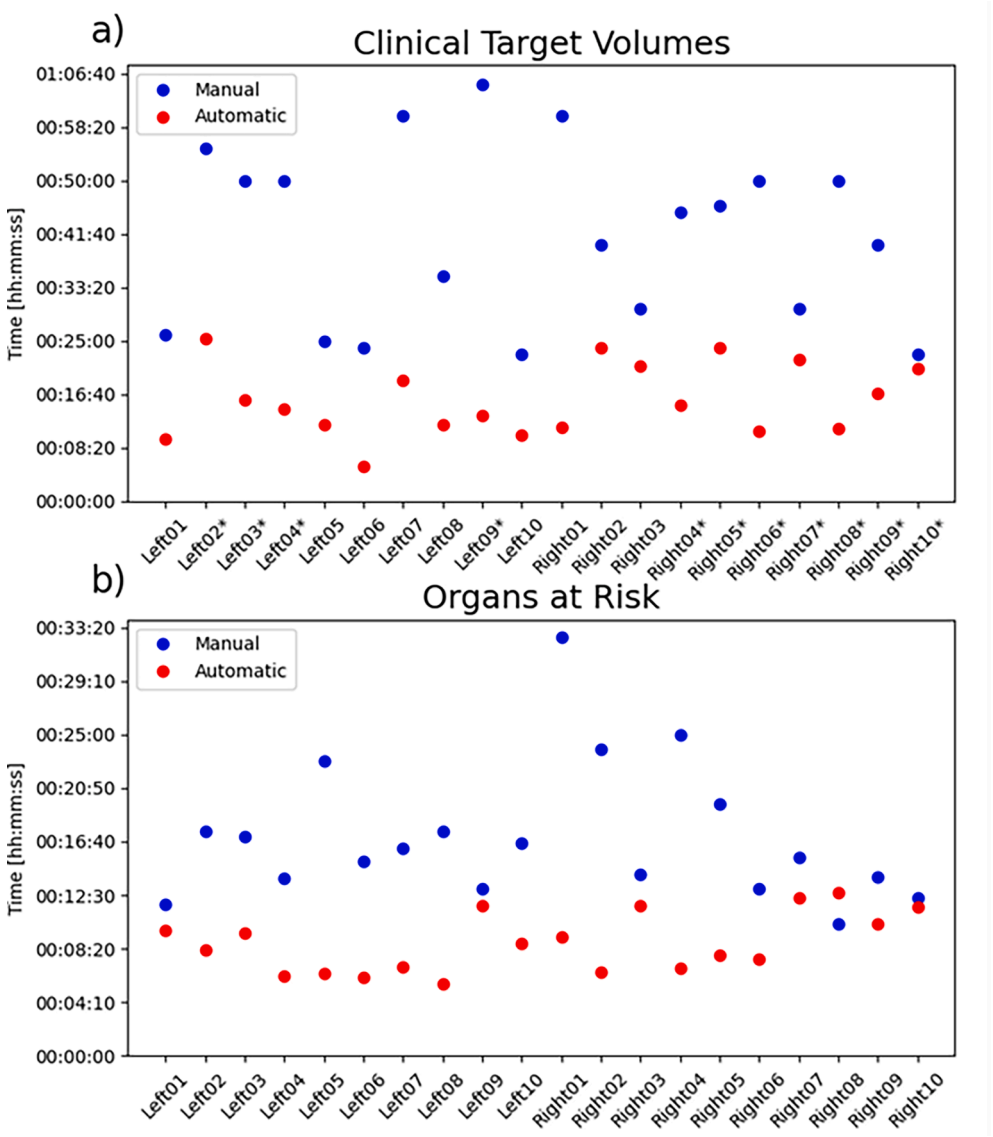

Fig. 1. Visualization of the time spent per patient on manual and automatic (including correction) segmentation of (a) clinical target volumes and (b) organs at risk. Patients including clinical target volume of node levels 1 to 4 are indicated with an asterisk, other patients only include node levels 1 and 2.

corrected) and corrected structures, using the manually generated contours as ground-truth. The difference between these metrics were calculated by subtracting the not-corrected metrics from the corrected metrics, and therefore a better agreement with the manual contour of the corrected contours is indicated by a positive DSC or sDSC value or a negative 95% HD value. The outcomes for the three metrics are visualized as boxplots in Fig. 3, and can also be found in Table S3 of the supplementary materials. For the left-sided CTVn1, a significant difference was observed for all three metrics, indicating an significant impact of the corrections made. For the right-sided CTVp and the heart, this significance was only observed for the DSC and sDSC scores. These results indicate that, although corrections were made in most of the cases, only for a small number of cases these corrections were actually significantly different.

Qualitative evaluation

The automatically generated contours were qualitatively scored by five different RTTs and five ROs. For both models, the scores are visualized in Fig. 4. While for the OARs, only the heart and thyroid both got scored as not usable in only one patient, this was more often the case for one of the CTVs. The CTVp was found not to be usable in a total of seven cases, CTVn1 in six cases and CTVn4 in four cases. A few observations can be made. First of all, the correction needed for the left and right lungs in respectively 20 and 50% of the cases for the right-sided model is remarkable, given the high quantitative score for these ROIs. Except for one case, these scores were assigned by the same observer. Something similar can be observed for CTVp and CTVn1, which were always assigned a score 3 by one of the observers, which scored four cases in total. No correlation was found between the quantitative metrics and the assigned scores, except for assigning score 3 to cases which were outliers in terms of DSC scores and HD95%. However, score 3 was also assigned to ROIs which had high quantitative results.

Discussion

This study explored the clinical feasibility of an auto-segmentation model for breast cancer. In 95% of the cases, time reduction was achieved while including corrections of the contours when needed. During qualitative measurements, the auto-contours got scored as clinical acceptable, with or without corrections, in 92% of the cases.

Several other studies are published which use auto-segmentation for

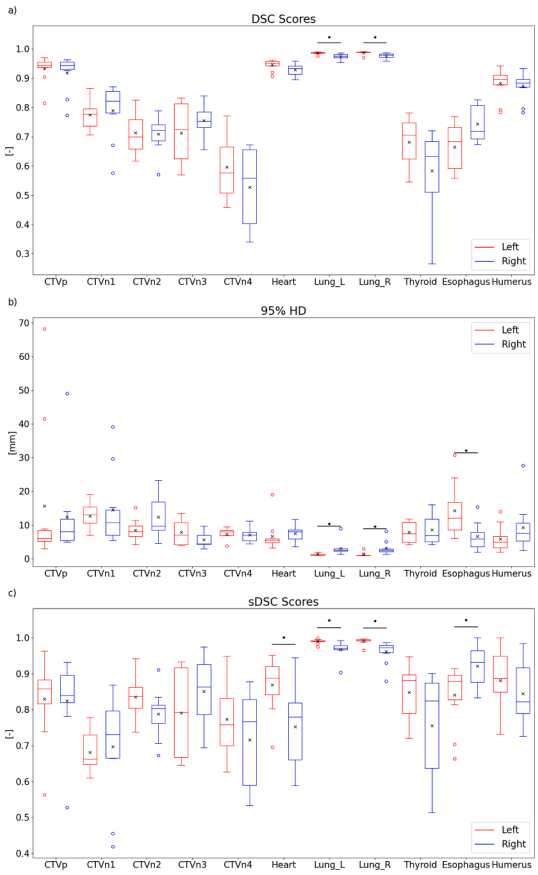

Fig. 2. Visualization of the (a) DSC scores, (b) 95% HD and (c) sDSC scores of the comparison between automatically and manually generated contours. Horizontal lines in boxes are medians, crosses are means, dots are outliers. Statistically significant differences are indicated with an asterisk ($p < 0.05$).

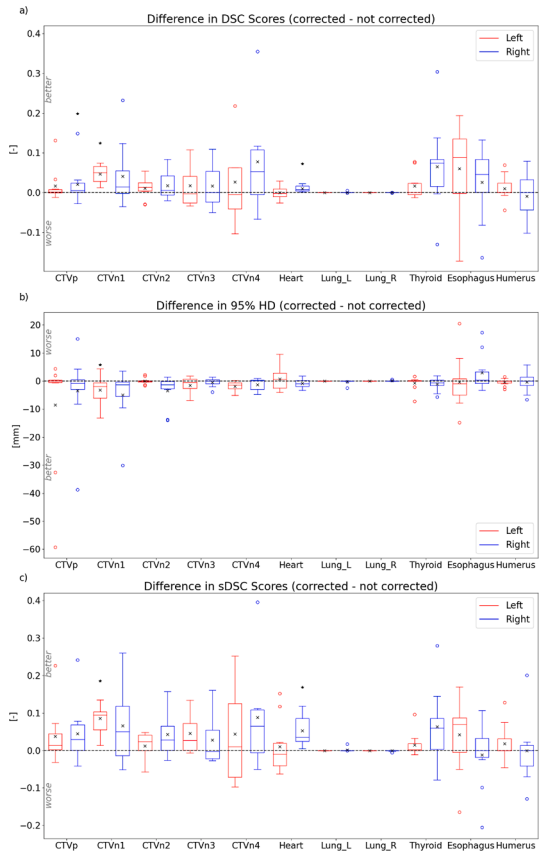

Fig. 3. visualization of the difference in (a) DSC scores, (b) 95% HD and (c) sDSC scores of the automatically generated contours (not corrected) and corrected contours, using the manually generated contours as ground-truth. Horizontal lines in boxes are medians, crosses are means, dots are outliers. Statistically significant differences between the contours are indicated with an asterisk ($p < 0.05$). Note: y-axis between Figs. 2 and 3 differ.

breast cancer. Recently, Almberg *et al.* performed a study in collaboration with RaySearch, evaluating a model for the same target areas [12]. Similar quantitative results were achieved, except for CTVn4, thyroid and humerus, which scored better in their study. However, in our study these ROIs were found to be useful as a starting point for correction in most cases, while still saving time, showing clinical feasibility. The study of Almberg *et al.* did not include a measurement of contouring time needed. In a future study, an extensive comparison could be made to assess the differences in performance in more depth and explore the origin of these differences.

During analysis of the qualitative results, it was observed that different scores were assigned by different RTTs and ROs for comparable corrections made. For example, one RO would assign a score 2 if he/she had to correct eight slices of the CTVp, while another RO would assign a score 3, emphasizing the subjectivity of these scores. This inter-observer difference in scoring could be a result of using a 3-point Likert scale, which is less able to capture extreme values than for example a 5-point Likert scale. Therefore, it was important to further investigate the actual corrections made during the study. It was found that, even when a score 3 was assigned, in some cases the auto-contour was still used as a starting point, and still lead to a reduction in time.

Besides, the study only involved scoring of auto-contours, which could induce subjectivity. Observers could judge auto-contours differently, then they would judge manual contours by a change in perception towards the use of AI. This difference could be overcome by performing a head-to-head comparison, such as the Turing test, in which the user has to identify the origin of the contour [19].

An evaluation method of the model performance which was not used in this study, but is used in other studies involving auto-segmentation, is dosimetric evaluation [13,20]. This method gives an indication of the clinical relevance of variations in contours and could therefore quantify the clinical relevance of corrections made. In a future study, the dosimetric impact of this auto-segmentation model could be evaluated. In order to perform such an analysis without any interobserver bias, ideally this should be done by the use of automatically generated treatments plans [21,22]. Moreover, these plans should be based on predefined clinical goals which are widely accepted to make such a comparison more generally useful [23]. Ideally, this dosimetric information could be incorporated in the auto-segmentation tool, indicating in which regions the uncertainty of the auto-contour is large, and in which regions corrections could be considered clinically relevant.

The results of this study led to the clinical implementation of the DL models for auto-segmentation. To fully assess the clinical use, the time needed for correction and amount of corrections will be monitored during the first period after clinical introduction to determine the efficiency of the use of the models. A difference between the clinical setting and the study setting could for example emerge when people gain more trust in the DL module. Besides, these outcomes could give insight in which ROIs need the most corrections, and therefore might require retraining of the model to improve the model outcome. This improvement could for example be achieved by the use of more and more consistent data.

Conclusions

A DL segmentation model was developed for both left- and right-sided locally advanced breast cancer. The primary endpoint of time

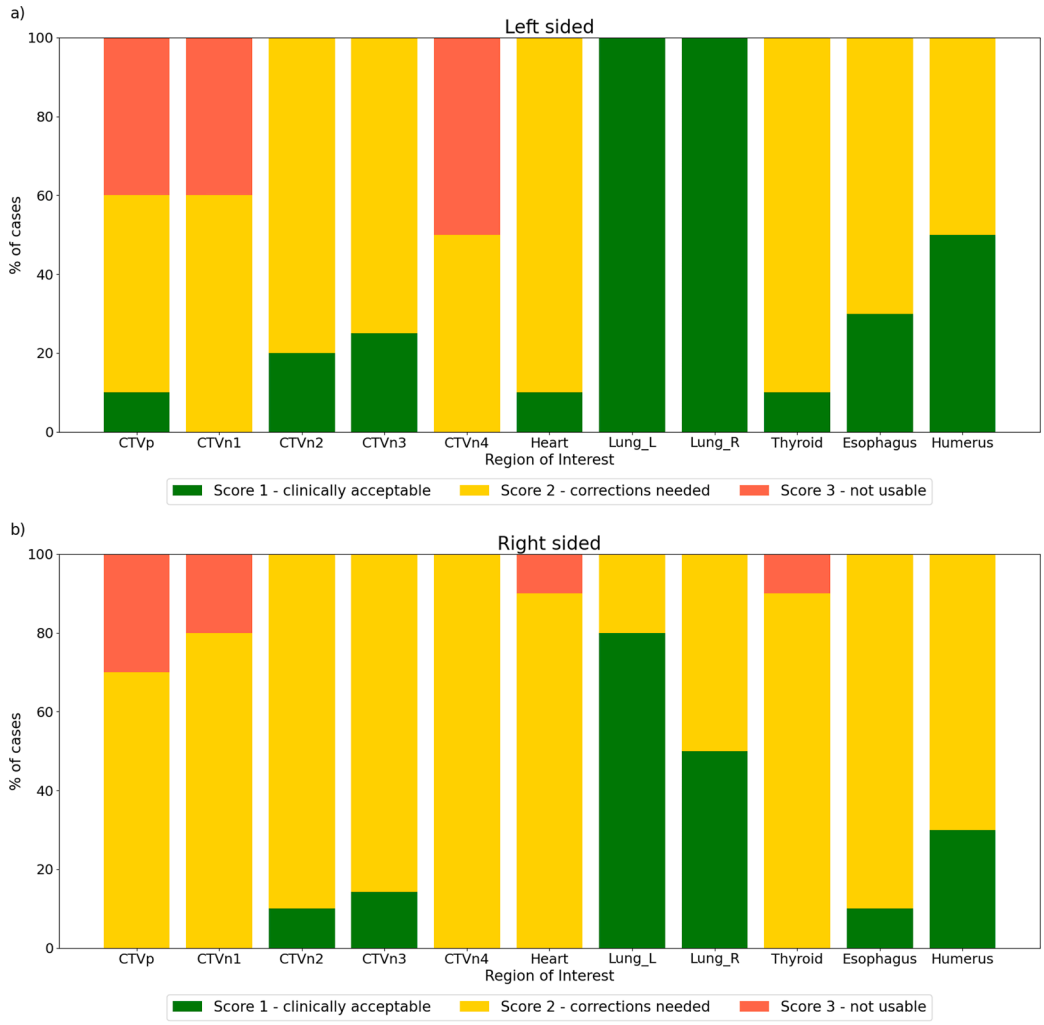

Fig. 4. Qualitative results for (a) left-sided and (b) right-sided model. For each ROI, the percentage of cases receiving one of the scores is indicated. For both sides, 10 patients were included, which contain CTVn3 and CTVn4 in 4 and 7 cases, for respectively the left- and right-sided model.

efficiency was fulfilled, while also improved consistency as an inherent effect of using auto segmentation. Besides, both quantitative as qualitative measurements showed high clinical potential. As a result of this study, the DL models will be clinically implemented in the near future in our clinic.

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.

Acknowledgements

We would like to thank Fredrik Löfman, Elin Samuelsson and Jonas Söderberg from the machine learning department of RaySearch Laboratories AB for their contribution by providing a training framework, integration of the AI model into RayStation and many fruitful discussions. A research grant from Raysearch Laboratories AB is also acknowledged.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tipsro.2023.100211.

References

[1] Li XA, Tai A, Arthur DW, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG multi-institutional and multiobserver study. Int J Radiat Oncol Biol Phys 2009;73:944–51. https://doi.org/10.1016/j.ijrobp.2008.10.034.

[2] Ciardo D, Argenone A, Boboc GI, et al. Variability in axillary lymph node delineation for breast cancer radiotherapy in presence of guidelines on a multi-institutional platform. Acta Oncol (Madr) 2017;56:1081–8. https://doi.org/10.1080/0284186X.2017.1325004.

[3] Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. Radiother Oncol 2012;102:68–73. https://doi.org/10.1016/j.radonc.2011.08.043.

[4] Eldesoky AR, Yates ES, Nyeng TB, et al. Internal and external validation of an ESTRO delineation guideline – dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer. Radiother Oncol 2016;121:424–30. https://doi.org/10.1016/j.radonc.2016.09.005.

[5] Ciardo D, Gerardi MA, Vigorito S, et al. Atlas-based segmentation in breast cancer radiotherapy: evaluation of specific and generic-purpose atlases. Breast 2017;32:44–52. https://doi.org/10.1016/j.breast.2016.12.010.

[6] Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiother Oncol 2020;153:139–45. https://doi.org/10.1016/j.radonc.2020.09.045.

[7] Qi X, Hu J, Zhang L, Bai S, Yi Z. Automated segmentation of the clinical target volume in the planning CT for breast cancer using deep neural networks. IEEE Trans Cybern 2020. i:1–11. https://doi.org/10.1109/tcyb.2020.3012186.

[8] Chung SY, Chang JS, Choi MS, et al. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. Radiat Oncol 2021;16:1–10. https://doi.org/10.1186/s13014-021-01771-z.

[9] Liu Z, Liu F, Chen W, et al. Automatic segmentation of clinical target volume and organs-at-risk for breast conservative radiotherapy using a convolutional neural network. Cancer Manag Re 2021;13:8209–17. https://doi.org/10.2147/CMAR.S330249.
[10] Byun HK, Chang JS, Choi MS, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol 2021;16:1–8. https://doi.org/10.1186/s13014-021-01923-1.
[11] Buelens P, Ir SW, I LV, Crijns W, Ir FM, Weltens CG. Clinical Evaluation of a deep learning model for segmentation of target volumes in breast cancer radiotherapy. Radiother Oncol 2022;171:84–90. doi:10.1016/j.radonc.2022.04.015.
[12] Almberg SS, Lervåg C, Frengen J, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol 2022;173:62–8. https://doi.org/10.1016/j.radonc.2022.05.018.
[13] Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.
[14] Mcintosh C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27:999–1005. https://doi.org/10.1038/s41591-021-01359-w.
[15] Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. Radiother Oncol 2015;114:3–10. https://doi.org/10.1016/j.radonc.2014.11.030.
[16] Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Sola AB, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer, version 1.1. Radiother Oncol 2016;118:205–8. https://doi.org/10.1016/j.radonc.2015.12.027.
[17] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Conf Med Image Comput Comput-Assist Intervent 2016:424–32. https://doi.org/10.1007/978-3-319-46723-8_49.
[18] Nikolov S, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Int Res 2021;23:e26151.
[19] Gooding MJ, Smith AJ, Tariq M, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test. Med Phys 2018;45:5105–15. https://doi.org/10.1002/mp.13200.
[20] Sherer MV, Lin D, Elguindi S, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review. Radiother Oncol 2021;160:185–91. https://doi.org/10.1016/j.radonc.2021.05.003.
[21] van de Sande D, Sharabiani M, Bluemink H, et al. Artificial intelligence based treatment planning of radiotherapy for locally advanced breast cancer. Phys Imag Radiat Oncol 2021;20:111–6. https://doi.org/10.1016/j.phro.2021.11.007.
[22] Kneepkens E, Bakx N, van der Sangen M, et al. Clinical evaluation of two AI models for automated breast cancer plan generation. Radiat Oncol 2022;17:1–9. https://doi.org/10.1186/s13014-022-01993-9.
[23] Hurkmans C, Duisters C, Peters-Verhoeven M, et al. Harmonization of breast cancer radiotherapy treatment planning in the Netherlands. Tech Innov Patient Support Radiat Oncol 2021;19:26–32. https://doi.org/10.1016/j.tipsro.2021.06.004.