Physics and Imaging in Radiation Oncology 34 (2025) 100749

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Development and external multicentric validation of a deep learning-based clinical target volume segmentation model for whole-breast radiotherapy

Maria Giulia Ubeira-Gabellini [a,*,1], Gabriele Palazzo [a,1], Martina Mori [a], Alessia Tudda [a], Luciano Rivetti [b], Elisabetta Cagni [c], Roberta Castriconi [a], Valeria Landoni [d], Eugenia Moretti [e], Aldo Mazzilli [f], Caterina Oliviero [g], Lorenzo Placidi [h], Giulia Rambaldi Guidasci [i], Cecilia Riani [a], Andrei Fodor [j], Nadia Gisella Di Muzio [j,k], Robert Jeraj [b,l], Antonella del Vecchio [a], Claudio Fiorino [a]

[a] Medical Physics, IRCCS San Raffaele Scientific Institute, Milan, Italy
[b] University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia
[c] Medical Physics Unit, Department of Advanced Technology, Azienda USL-IRCCS di Reggio Emilia, Reggio Emilia, Italy
[d] Istituto Nazionale dei tumori Regina Elena, Rome, Italy
[e] Department of Medical Physics, University Hospital, Udine, Italy
[f] Medical Physics Dept, University Hospital of Parma AOUP, Italy
[g] University Hospital, "Federico II", Naples, Italy
[h] Fondazione Policlinico Universitario A. Gemelli IRCCS, uosd medical physics and radioprotection, Rome, Italy
[i] UOC di Radioterapia Oncologica, Fatebenefratelli Isola Tiberina – Gemelli Isola, Rome, Italy
[j] Radiation Oncology, IRCCS San Raffaele Scientific Institute, Milan, Italy
[k] Vita-Salute San Raffaele University, Milan, Italy
[l] Department of Medical Physics, University of Wisconsin, Madison, USA

ARTICLE INFO

Keywords:
Segmentation
Breast clinical target volume
Radiotherapy
Deep learning
Artificial intelligence

ABSTRACT

Background and purpose: In order to optimize the radiotherapy treatment and minimize toxicities, organs-at-risk (OARs) and clinical target volume (CTV) must be segmented. Deep Learning (DL) techniques show significant potential for performing this task effectively. The availability of a large single-institute data sample, combined with additional numerous multi-centric data, makes it possible to develop and validate a reliable CTV segmentation model.

Materials and methods: Planning CT data of 1822 patients were available (861 from a single center for training and 961 from 8 centers for validation). A preprocessing step, aimed at standardizing all the images, followed by a 3D-Unet capable of segmenting both right and left CTVs was implemented. The metrics used to evaluate the performance were the Dice similarity coefficient (DSC), the Hausdorff distance (HD), and its 95th percentile variant (HD_95) and the Average Surface Distance (ASD).

Results: The segmentation model achieved high performance on the validation set (DSC: 0.90; HD: 20.5 mm; HD_95: 10.0 mm; ASD: 2.1 mm; epoch 298). Furthermore, the model predicted smoother contours than the clinical ones along the cranial–caudal axis in both directions. When applied to internal and external data the same metrics demonstrated an overall agreement and model transferability for all but one (Inst 9) center.

Conclusion: . A 3D-Unet for CTV segmentation trained on a large single institute cohort consisting of planning CTs and manual segmentations was built and externally validated, reaching high performance.

1. Introduction

A widely applied approach in the treatment of early breast cancer consists of breast-conserving surgery followed by radiotherapy (RT). To optimize RT, organs-at-risk and clinical target volume (CTV) must be segmented. This is a highly time-consuming task, currently performed manually, and prone to large inter and intra-observer variability, which is difficult to quantify [1–4], and negatively influences the consistency of the delivered dose distributions among different institutions [5].

* Corresponding author.
E-mail address: ubeira.mariagiulia@hsr.it (M.G. Ubeira-Gabellini).
[1] The two authors contributed equally to this research work.

https://doi.org/10.1016/j.phro.2025.100749
Received 10 September 2024; Received in revised form 24 February 2025; Accepted 14 March 2025
Available online 26 March 2025
2405-6316/© 2025 The Authors. Published by Elsevier B.V. on behalf of European Society of Radiotherapy & Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Moreover, the clinical inter-observer variability in the cranial direction is typically larger than that in the caudal direction [1,6]. Thus, interest in automatic tools supporting organs-at-risk/CTV segmentation is rapidly increasing, although few studies to date have dealt with CTV breast segmentation using large clinical cohorts [7]. Atlas-based methods were first introduced [5,8,9], while Deep Learning (DL) based methods (e.g., ResNet, U-net, nnUnet) are preferred to train/test reliable models [10–15], particularly for soft tissue-based regions. Although auto-contouring may help in optimizing and reducing the clinical workload in the RT workflow, particular attention must be given to high segmentation accuracy considering its direct impact on treatment planning and patient care. Moreover, in order to improve consistency between manual and automatic segmentation, inter-observer variability should be taken into account, a task not fully addressed by ATLAS based methods, but potentially manageable by DL models. Very importantly, the inter-observer (and inter-Institutional) variability in breast CTV segmentation is large and, to our knowledge, this issue has been considered by very few works into the development of DL tools for auto-segmentation [16–18]. Recent data using commercial systems [19], tested on a large population, reached relatively modest performance, with a Dice Similarity Coefficient (DSC) around 0.8, reinforcing the need to explore more advanced tools. Furthermore, to our knowledge, multi-institute external validations of DL models for breast CTV auto-segmentation have not been reported.

The aim of this work was to use a large patient cohort treated in a single institute (861 patients) to train a 3D DL model for segmentation of the CTV, using in-house open-source tools. The model was externally validated on a multicentric cohort of eight centers (including the same center on which training was performed) on 961 patients.

2. Materials and methods

2.1. Patient data, contouring, planning and delivery procedures

The dataset used to train and validate the model consisted of 861 patients treated at San Raffaele Institute, who underwent breast-conserving surgery followed by RT; 80% of the total was used for the training and 20% for internal validation. For each patient, 3D planning CT images and CTV contours delineated by the treating physicians were available. Patients collected were 450 for right-sided breast cancer (BC) and 411 for left-sided BC. All images were acquired with the same scanner (GE Medical System - HiSpeed NX/i) and showed variable 3D size and resolution (in the transverse plane the spacing between axes ranged from 0.97 mm to 1.95 mm; in the sagittal plane the distance between adjacent slices ranged from 3 to 5 mm). CTV delineation was performed by different clinicians in the period 2017–2021. This retrospective study was performed according to the principles of the Helsinki Declaration. Informed consent for treatment and disease-related information publication was obtained from all participants. Patients included consisted of both patients with (n = 181) and without (n = 680) boost, although only the CTV referring to the whole breast was considered. Patients with silicone implants were excluded due to their different topology and the small number of cases. This is explained in the Supplementary material (Section A.1.1). The dataset size, combined with the applied pre-processing, enhanced model generalizability.

Contouring was performed according to national guidelines [20,21]: the contour, initially manually drawn adjacent to the skin, was automatically cropped by all centers based on the outer body contour at 5 mm from the skin surface. Details of planning and delivery approaches are reported elsewhere [22–24]. In short, planning target volume (PTV) was generated by expanding the CTV with a 5–10 mm margin in all directions, except for the lung (5 mm); PTV was cropped to avoid the first 5 mm of skin. A median of four segments was used within a tangential 3D Conformal Radiation Therapy technique. Regarding target delineation and RT technique, clinical practice has remained stable over the years.

As an external validation dataset, CTs from the MIKAPOCo study were used, with 969 patients across several centers. More details are provided in Section 2.3.

Preliminary evaluation of inter-observer variability (IOV) in our institute was calculated: three expert clinicians were asked to delineate the same CTV structures for ten new patients (five right and five left-sided breast cancer cases).

2.2. Building the network for CTV auto-contouring

A pre-processing phase was first performed, aiming to make the image format optimal for training a neural network (i.e.: crop, resize, scaling). All images were re-sampled in order to have uniform images of 512*512*256 voxels (voxel space: 1*1*1.5 mm^3). To help in the CTV side identification (right or left) and to reduce the input parameters without reducing voxel resolution, an automatic cropping of the left/right quadrant was applied, resulting in final image dimensions of 256*256*256 voxels. To do this, the xy coordinates of the body segmentation center were computed for each slice along the cranial–caudal axis. The median value of these coordinates was used as the lower left/right (depending on the side) corner of the crop rectangle for every slice with the addition of a fixed extra shift of 25 pixels, including more tissue in the image fed to the neural network (see Supplementary Figure S3 in the Supplementary material). Moreover, to ensure optimal network performance and enhance contrast between glandular and fat tissue, a minimum of −200 HU and maximum of 100 HU were chosen and the values subsequently normalized to a range of 0 to 1.

All images (i.e., with left/right BC) were then used to train a global model. A 3D-Unet, enriched with dropout and residual units, from the MONAIlibrary (v1.3), was implemented. All procedures were implemented in-house in a python 3.10.9 based platform developed for machine learning (ML)/DL radiotherapy and medical physics applications, included in the specific library of mAItre [25] (the documentation is freely accessible under the name libsrdl). The best parameters of the 3D-Unet were obtained after hyperparameter optimization performed through "optuna" [26]. Other architectures were tested (e.g. DynUnet, SegResNetDS), obtaining comparable or lower performance. Regions with rescaled intensity value of zero were removed in order to focus on the valid body area. Data augmentation (random flip on x, rotation, translation, scaling, Gaussian sharpening) was performed. The same pre-processing was also applied without cropping the images, followed by two separate models for right and left breast. Those results are shown in the Supplementary material (Supplementary Figures S4, S5). Their results are similar, i.e.: with similar metrics performance, even though the model with the crop is preferred because it includes both breast sides in a single model, it does not lose resolution and it converges faster.

The segmentation performance of the validation dataset was assessed using DSC, Symmetric Average Surface Distance (ASD), Hausdorff distance (HD) and its 95th percentile variant (HD_95). Each 3D segmentation prediction was found to have more than one single continuous volume, with the CTV contour being the most extended. For this reason, a selection of the largest continuous 3D volume for each prediction was implemented a posteriori. This procedure was also applied during training and, thus, all the plots show the results with this post-processing applied. Finally, the predicted segmentation was co-registered back to the original image dimensions needed in future for clinical purpose. Additionally, an assessment of the computational resources required for model deployment was performed. More details are provided in supplementary material (see Section A.1.3).

2.3. Multiple independent validation

As an external validation dataset, planning CTs from the MIKAPOCo study (Multi Institutional Knowledge-based Approach to Plan Optimization for the Community, Institutional Board approval: 248/2021)

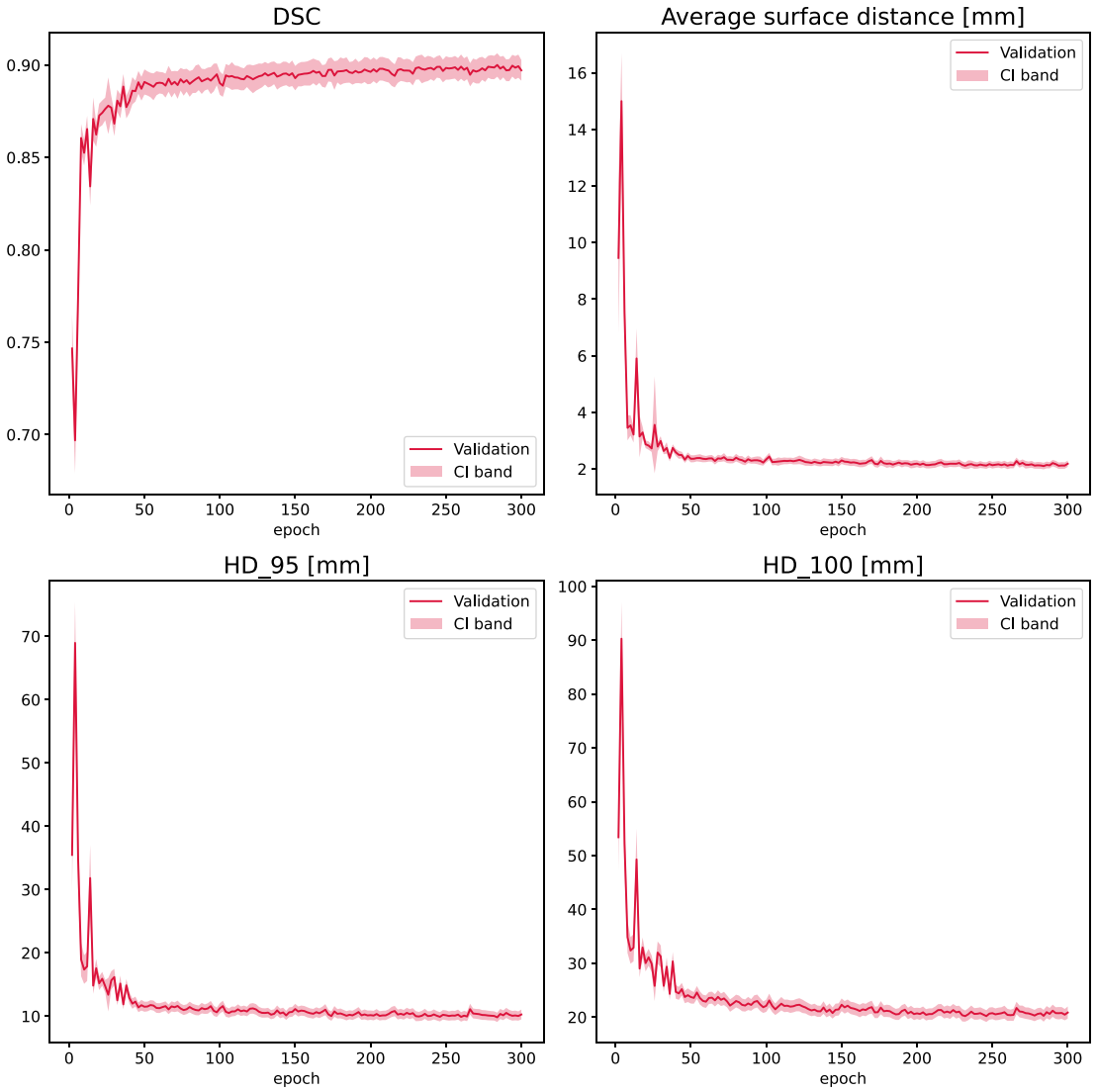

Fig. 1. Dice Similarity Coefficient (DSC), Symmetric Average Surface Distance (ASD), 95th and 100th percentile Hausdorff distance (HD_95 and HD) metrics over 300 epochs during the validation process. The corresponding bands represent the 95% confidence interval (CI) for each metric. The model converged after 298 epochs. The best metric values are listed in Supplementary Table S3. Data augmentation was not applied to the validation set shown in this plot.

were used [27–30]. The original dataset included patients from 10 centers (Inst 1–10); lung/heart dose-volume histograms were modeled to generate plan prediction models [29,30]. CT and contours of the patients of 8 of the 10 centers were available, while Inst 6 and Inst 8 did not join the current (optional) study. Within the MIKAPOCo activities already completed, plan prediction models were developed to quantify inter-institutional variability in plan prediction, suggesting high transferability for both right and left breast scenarios [29]. Inst 3 corresponded to the Institution to which the training dataset referred: this additional group did not overlap with the training set, namely 37 for left and 44 for right. This group was retained to compare the network's performance on external datasets against that of the training center. Overall, the considered patients were 509 for right-sided BC and 452 for left-sided BC, with a minimum number of patients per center per side of 54 (right) and 65 (left): the available number of patients from each center is reported in the Supplementary material, including those with silicone (Supplementary Table S1).

The pre-processing applied to the external validation dataset was the same as that applied to the training set and the global model was applied to the images for prediction. From the generated CTV prediction, quantitative features were extracted to assess contour quality. The metrics chosen for comparisons between centers were DSC, HD, HD_95 and the volumes difference (VD) between CTVs predicted and segmented by radiation oncologists. The cranial and caudal border distances between clinical and model-based segmentations were also assessed. As for training, post-processing was applied to select the largest continuous contour for each patient, avoiding mispredictions due to cranial–caudal and contralateral uncertainties. Inter-Institute differences in the various metrics were analyzed, and statistically significant values were assessed using the non-parametric Mann–Whitney U-test.

3. Results

The best 3D-Unet model obtained was able to reach, after 298 epochs, high metric performance on the validation cohort (DSC: 0.90 ± 0.01; HD: 20.5 ± 1.0 mm; HD_95: 10.0 ± 0.7 mm) as shown in Fig. 1 and Supplementary Figure S2. The results were in line with the preliminary IOVs calculated at our institute, which were as follows: DSC: 0.89, HD: 20.0 mm, and HD 95: 12.7 mm. The predicted segmentation accurately matched the clinicians' contours on the skin and chest side, while it was more uncertain, as expected, on the lateral side (see Fig. 2).In particular, the median absolute difference between predicted and observed segmentation was 7.5 mm (IQR: 3.0–10.5) in the caudal and 9.0 mm (IQR: 3.0–16.5) in the cranial directions for left side CTVs,

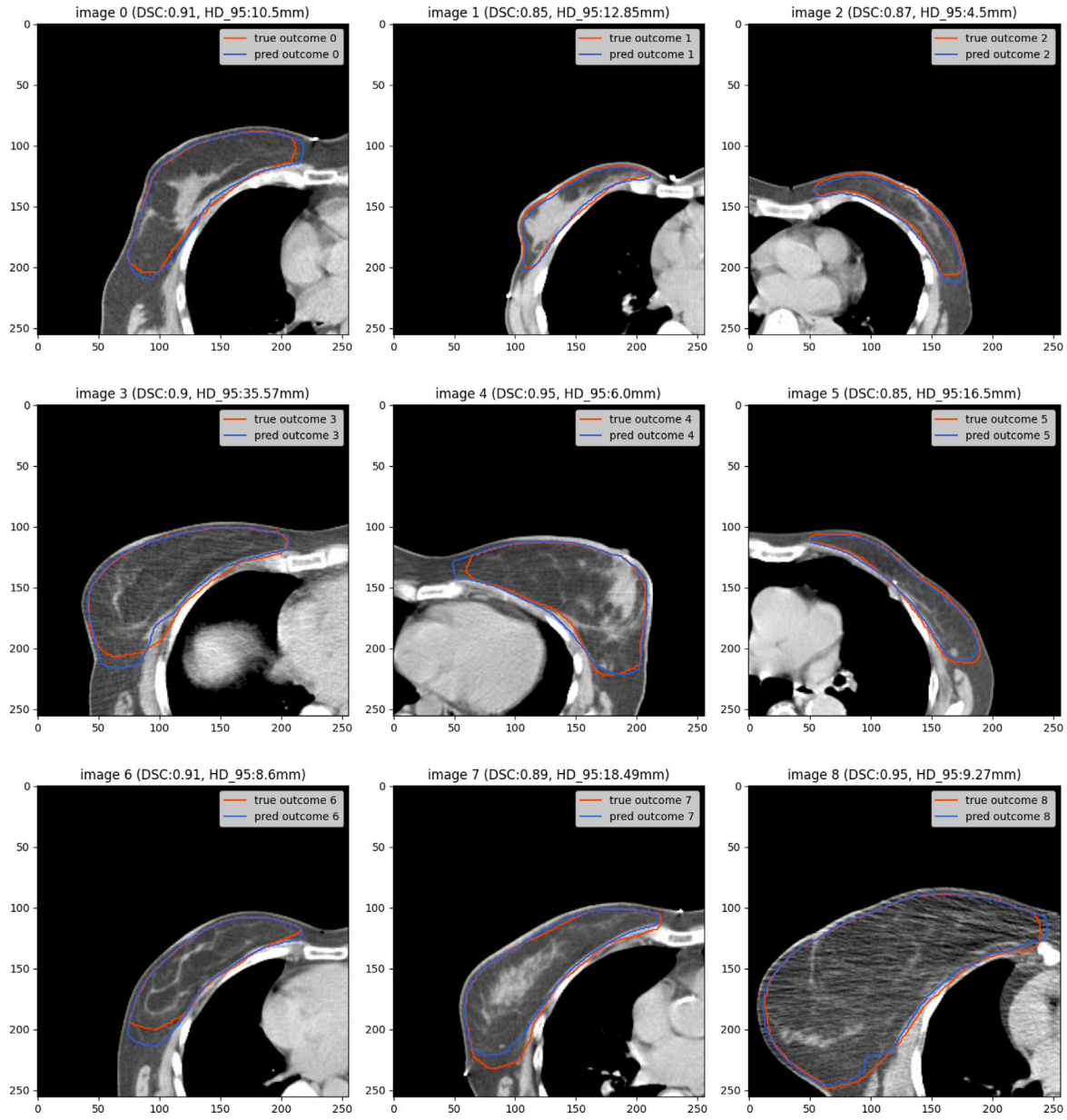

Fig. 2. Nine randomly chosen patient CTs right and left samples taken at the mean cranial–caudal slices from training data of Inst 3. The clinical CTV delineation (red) is superimposed on the observed predicted (blue) segmentation ROI. The DSC and HD_95 values for each patient are shown above the images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and 6.7 mm (IQR: 3.0–10.9) and 9.0 mm (IQR: 1.5–13.9) for right side CTVs. Thus they had a larger prediction along the cranial–caudal border compared with the clinical contours. However, the predicted segmentation volume was found to be globally smaller than the clinical contours for Inst 3 (−20 cm^3 for left and −26 cm^3 for right, $p < 0.05$).

The behavior of DSC, HD, HD_95, VD metrics in the external MIKAPOCo dataset is shown in Fig. 3. However, to enhance the difference on the lateral part and for better visualization, a further crop on the cranial–caudal direction to match clinical contour was performed; the cranial–caudal behavior was thus separate and shown in Fig. 4. The performances for Inst 3 (i.e.: the same from which training data are) were in quite good agreement with the values obtained in the training (DSC: 0.89, HD: 28.7 mm, HD_95: 14.4 mm; both sides): the average values referred to training are shown as black lines in Fig. 3. Inst 4 and 9 left differed from Inst 3 in two out of three metrics ($p < 0.05$). Considering that Inst 4 had better performance than Inst 3, the performances may be considered highly satisfactory. Our model tended to predict CTVs larger than those of the radiation oncologists, mainly in the cranial and caudal borders for almost all centers. Excluding the cranial/caudal agreement, performances slightly varied among centers: Inst 4 (right side), 5 and 9 volume had predictions typically larger than clinical CTV, while the other centers were statistically in line with clinical volumes. Comparing volume differences to those of Inst 3, all centers except for Inst 2 and 10 had smaller differences ($p < 0.05$). The different behavior of Inst 9 was not due to a local interpretation of the guidelines, as described in Section 2. Fig. 4 shows how CTV prediction varies in comparison to clinical contours in both the caudal and cranial direction. Results show how, for almost all centers, the extension in both cranial and caudal direction was larger for predictions than for clinical CTVs. Caudal and cranial difference on the training dataset were on average of 3.7 mm and 5.5 mm, respectively. To complete the results, for each metric a table with the mean and standard deviation of patients per center is shown (Table 1). The predicted segmentation was post-processed to retain the largest continuous contour, however,

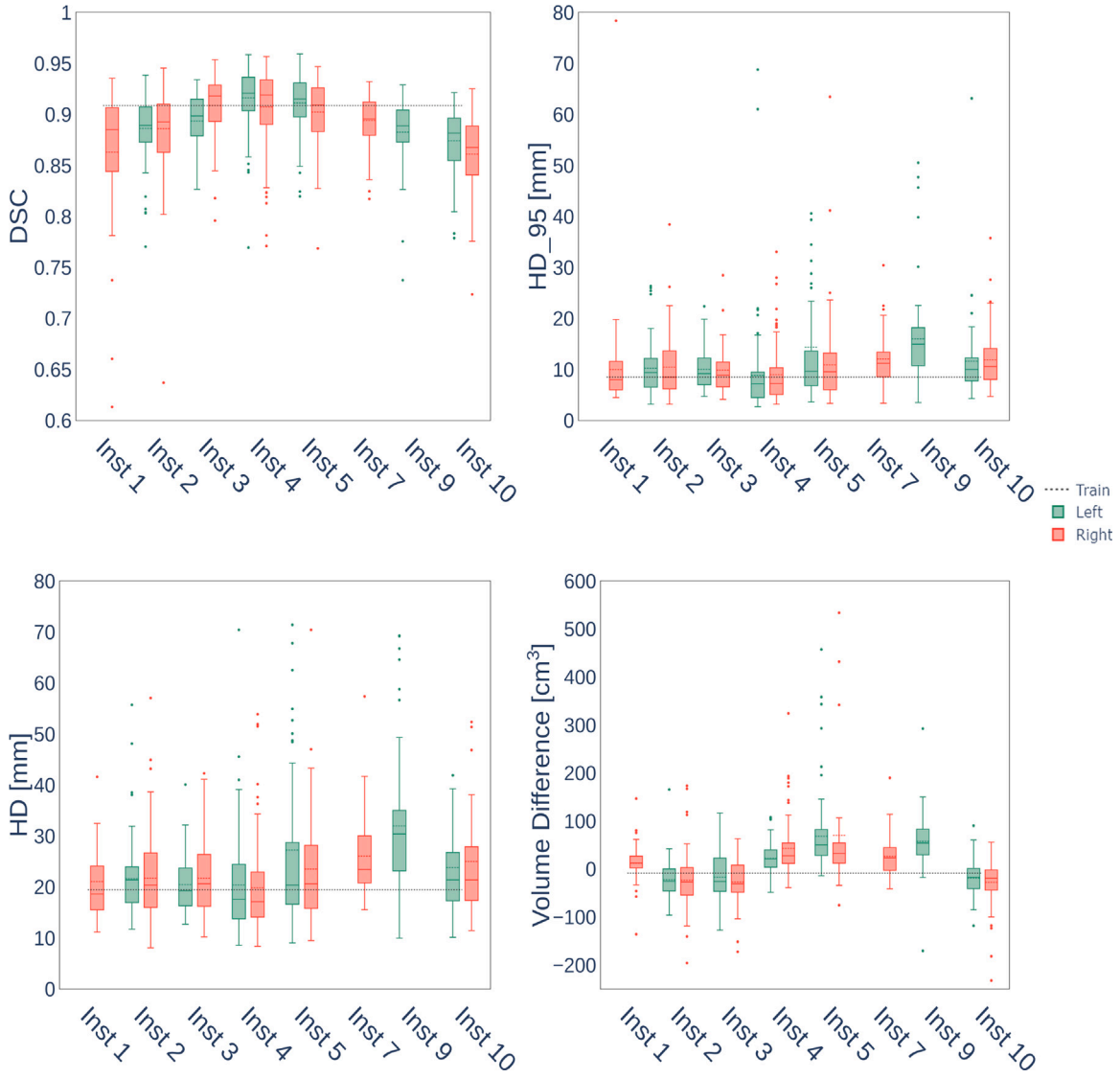

Fig. 3. DSC, HD, HD_95 and VD trend across different centers, with Inst 3 being the training center. VD negative signifies smaller CTV prediction. Box plots are shown with median (solid line) and mean (dashed lines) values. Outliers are shown with dots. Black lines show the best metric values obtained for training (applying cranial–caudal crop). Post-processing was applied to the CTV prediction as described in Materials and Methods, followed by clinical cranial–caudal cropping to enhance the main central difference.

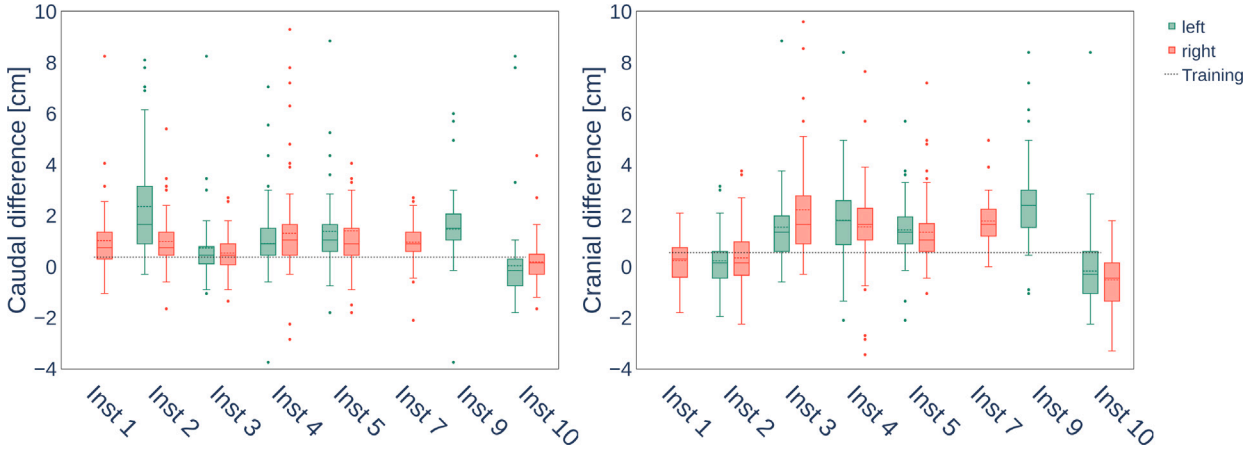

Fig. 4. Difference in the sagittal plane of clinical and predicted contours in the caudal (left) and cranial (right) direction for all MIKAPOCo centers considered on both right (red) and left (green) breast segmented. Post-processing was applied as described in Section 2 with no extra cropping. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
For each metric mean and standard deviation are shown for patients divided for center. Predicted segmentation was post-processed to keep the largest continuous contour, but with no crop to match clinical contours was applied, as instead done in Fig. 3. The mean values are thus larger than what shown in the correspondent plot.

Center	mean_left	std_left	mean_right	std_right
	DSC			
Inst 1	–	–	0.86	0.09
Inst 2	0.89	0.03	0.89	0.04
Inst 3	0.89	0.03	0.91	0.03
Inst 4	0.92	0.03	0.91	0.04
Inst 5	0.91	0.03	0.90	0.03
Inst 7	–	–	0.89	0.03
Inst 9	0.88	0.03	–	–
Inst 10	0.87	0.03	0.86	0.04
	Hausdorff distance [mm]			
Inst 1	–	–	21.1	10.8
Inst 2	21.6	7.1	21.7	8.3
Inst 3	20.5	5.6	21.7	7.2
Inst 4	20.4	11.1	19.8	9.0
Inst 5	27.2	20.1	23.5	11.7
Inst 7	–	–	26.1	7.7
Inst 9	32.0	14.0	–	–
Inst 10	23.8	11.2	25.0	13.8
	95% Hausdorff distance [mm]			
Inst 1	–	–	10.0	9.0
Inst 2	10.2	4.8	10.5	5.8
Inst 3	10.0	4.0	9.9	4.7
Inst 4	8.8	9.0	9.0	5.6
Inst 5	14.4	17.4	10.9	8.2
Inst 7	–	–	12.0	4.9
Inst 9	16.0	8.7	–	–
Inst 10	11.6	7.6	11.9	5.6
	Volume difference [cm^3]			
Inst 1	–	–	12.02	32.84
Inst 2	−22.39	36.11	−23.67	55.13
Inst 3	−17.11	52.04	−27.53	50.17
Inst 4	21.96	28.59	42.97	52.83
Inst 5	67.91	76.67	70.10	181.77
Inst 7	–	–	26.09	40.30
Inst 9	57.36	52.03	–	–
Inst 10	−18.98	35.38	−27.96	47.79

no cropping to match clinical contours was applied as was done in Fig. 3.

4. Discussion

The issue of clinical segmentation manually performed with significant effort by clinicians is still open, particularly concerning CTV segmentation. Advances in DL now enable realistic automation. The large dataset collected both from a single center and from multiple centers allowed us to train a reliable 3D neural network and validate it both internally and externally. Our model's performance was in line with the results reported by other groups [31–33]. However, our dataset was much larger than those used in [15,31–33]. Additionally, our approach to using a 3D model differed from the methods described in [11]. A 3D model enabled us to analyze the cranial–caudal axis together while predicting a smooth volume . Additionally, external validation of a large dataset strengthened the model's stability and generalizability. Moreover, our model performed significantly better than Atlas methods [5] and reported performance of a commercial system [19] (DSC∼0.8), aligning with the IOV of our department. The need for manual review and correction of CTVs by clinicians as previously suggested, along with evaluation of time savings [31] and dosimetric impact [32,33], was also evident. It is worth noting that with clinicians refining our model's predictions, the results would align even more closely with the initial clinical contours. A follow-up paper will assess our model's performance on a small sub-set after editing, providing a qualitative score and evaluating the time savings and physicians' acceptance. Considering that DSC is not the best metric — being sensitive to volume effects [34] — yet remains the most typically used, and given the wide variety of metrics present in literature [35, 36], it is highly important to highlight that our model's performance (see Supplementary Table S3) in terms of Intersection Over Union (IOU∼0.82) and ASD (around 2 mm) was quite consistent with values reported for IOV [2–4]. Even among observers from the same center, the conformity index, defined as the IOU, reached about 0.85, with an ASD of 3.9 mm [3]. The visual comparison between predicted and clinical CTVs delineated by our physicians showed no clear distinctions from clinical CTV delineated by our physicians. Results regarding the cranial and caudal borders deserve specific discussion. While a large IOV is well-documented [6] (cranial: 28 mm; caudal: 15 mm), a systematic deviation was found between the model's predictions and the clinical delineations. This highlights a tendency for clinical CTVs to be truncated, especially at the cranial border, while the predicted CTV exhibits a rounded and more regular shape at the borders. This variability along the cranial–caudal direction was also confirmed when running the same model with different random seeds for the dropout layers. It remains unclear whether this difference has any clinical impact, aside from the fact that automatically generated CTVs were larger at the borders, reducing the risk of missing the target.

Although predicted segmentation volumes were smaller than clinical contours, their larger extent along the cranial–caudal border indicates that the model was unable to compensate, on average, for its smaller prediction on the transversal plane.

Very importantly from the point of view of generalizability and usability of the model, this study represents, for the first time, an external validation conducted across multiple centers demonstrating the model's consistency across quite a large number of institutions in the same country. Notably, all centers showed results largely in line with the training center, with the exception of Inst 9, which seems to follow a different contouring approach (smaller CTV, [29]), despite all centers claiming adherence to national delineation guidelines. Based on this, contours from all centers — except Inst 9 — could be included in a future benchmark model. It is important to note that patients with silicone implants were excluded from the dataset, see Section A.3.1. Comparing cranial and caudal distances with the training population, caudal results of Inst 3 from MIKAPOCo were in line with those values, while cranial distances were larger. Cranial contouring differences were statistically significant for all centers except Inst 4, 5, and 7. Caudal differences were notable, except for Inst 1 (right side CTVs,$p = 0.09$). Outliers were investigated (see Section A).

More details about model deployment are provided in supplementary material (see Section A.2.3).

In conclusion, a 3D-Unet for CTV segmentation trained on 861 patients of a single center was built. The training performance was assessed using validation metrics, obtaining mean DSC: 0.90 and HD_95: 10.0 mm at epoch 298. The model was internally and externally validated through the MIKAPOCo dataset. Internal MIKAPOCo validation confirmed the results obtained in the training. The metrics used for comparison demonstrated overall agreement and model transferability for all centers, with the exception of Inst 9. Inst 10 and 1 exhibited greater variability in some metrics, though overall performance remained comparable with the internal MIKAPOCo validation set. Our model tended to predict larger contours than those delineated by radiation oncologists along the cranial–caudal axis, while for the lateral side, it depended on the center, with Inst 9 displaying the most significant discrepancy, with clinical contours significantly smaller than those from the other centers. Future works will test these results by incorporating a quality score assigned by radiation oncologists to assess the impact of the predicted CTVs on clinical workflow, giving an indication of whether, and to what degree, manual editing is required to ensure clinical acceptability.

CRediT authorship contribution statement

Maria Giulia Ubeira-Gabellini: Conceptualization, Data organization, Formal analysis, Methodology, software, Visualization, Writing – original draft. **Gabriele Palazzo:** Conceptualization, Data organization, Formal analysis, Methodology, software, Visualization, Writing – original draft. **Martina Mori:** Data curation, Data organization. **Alessia Tudda:** Data curation. **Luciano Rivetti:** Conceptualization, Methodology, software. **Elisabetta Cagni:** Data curation. **Roberta Castriconi:** Data curation. **Valeria Landoni:** Data curation. **Eugenia Moretti:** Data curation. **Aldo Mazzilli:** Data curation. **Caterina Oliviero:** Data curation. **Lorenzo Placidi:** Data curation. **Giulia Rambaldi Guidasci:** Data curation. **Cecilia Riani:** Formal analysis, Methodology, software. **Andrei Fodor:** Conceptualization, Writing – review. **Nadia Gisella Di Muzio:** Writing – review. **Robert Jeraj:** Supervision. **Antonella del Vecchio:** Writing – review. **Claudio Fiorino:** Conceptualization, Writing – original draft, Supervision, Funding acquisition.

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.

Acknowledgments

The study was supported by an AIRC (Associazione Italiana per la Ricerca sul Cancro) grant (IG23150). The authors want to thank prof. S. Carrazza (TIF Lab, Dipartimento di Fisica, Università degli Studi di Milano, Italy) for useful discussions.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.phro.2025.100749.

References

[1] Pitkänen M, Holli K, Ojala A, Laippala P. Quality assurance in radiotherapy of breast cancer: Variability in planning target volume delineation. Acta Oncol 2001;40(1):50–5. http://dx.doi.org/10.1080/028418601750071055, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035076811&doi=10.1080%2f028418601750071055&partnerID=40&md5=928489075baf8660dd5efd848c08e1f9.

[2] Struikmans H, Wárlám-Rodenhuis C, Stam T, Stapper G, Tersteeg RJ, Bol GH, Raaijmakers CP. Interobserver variability of clinical target volume delineation of glandular breast tissue and of boost volume in tangential breast irradiation. Radiother Oncol 2005;76(3):293–9, URL http://dx.doi.org/10.1016/j.radonc.2005.03.029.

[3] Li XA, Tai A, Arthur DW, Buchholz TA, Macdonald S, Marks LB, Moran JM, Pierce LJ, Rabinovitch R, Taghian A, Vicini F, Woodward W, White JR. 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(3):944–51, URL http://dx.doi.org/10.1016/j.ijrobp.2008.10.034.

[4] van Mourik AM, Elkhuizen PH, Minkema D, Duppen JC, van Vliet-Vroegindeweij C. Multiinstitutional study on target volume delineation variation in breast radiotherapy in the presence of guidelines. Radiother Oncol 2010;94(3):286–91, URL http://dx.doi.org/10.1016/j.radonc.2010.01.009.

[5] Ciardo D, Gerardi MA, Vigorito S, Morra A, Dell'acqua V, Diaz FJ, Cattani F, Zaffino P, Ricotti R, Spadea MF, Riboldi M, Orecchia R, Baroni G, Leonardi MC, Jereczek-Fossa BA. Atlas-based segmentation in breast cancer radiotherapy: Evaluation of specific and generic-purpose atlases. Breast J 2017;32:44–52, URL http://dx.doi.org/10.1016/j.breast.2016.12.010.

[6] Hurkmans CW, Borger JH, Pieters BR, Russell NS, Jansen EP, Mijnheer BJ. Variability in target volume delineation on CT scans of the breast. Int J Radiat Oncol Biol Phys 2001;50(5):1366–72, URL http://dx.doi.org/10.1016/S0360-3016(01)01635-2.

[7] Eldesoky AR, Yates ES, Nyeng TB, Thomsen MS, Nielsen HM, Poortmans P, Kirkove C, Krause M, Kamby C, Mjaaland I, Blix ES, Jensen I, Berg M, Lorenzen EL, Taheri-Kadkhoda Z, Offersen BV. 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(3):424–30, URL http://dx.doi.org/10.1016/j.radonc.2016.09.005.

[8] Velker VM, Rodrigues GB, Dinniwell R, Hwee J, Louie AV. Creation of RTOG compliant patient CT-atlases for automated atlas based contouring of local regional breast and high-risk prostate cancers. Radiat Oncol 2013;8(1). URL http://dx.doi.org/10.1186/1748-717X-8-188.

[9] Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, Chang JS, Kim JS. 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, URL http://dx.doi.org/10.1016/j.radonc.2020.09.045.

[10] Bell LR, Dowling JA, Pogson EM, Metcalfe P, Holloway L. Atlas-based segmentation technique incorporating inter-observer delineation uncertainty for whole breast. J Phys Conf Ser 2017;777:012002, URL http://dx.doi.org/10.1088/1742-6596/777/1/012002.

[11] Men K, Zhang T, Chen X, Chen B, Tang Y, Wang S, Li Y, Dai J. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Phys Med 2018;50:13–9, URL http://dx.doi.org/10.1016/j.ejmp.2018.05.006.

[12] Byun HK, Chang JS, Choi MS, Chun J, Jung J, Jeong C, Kim JS, Chang Y, Chung SY, Lee S, Kim YB. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol 2021;16(1). URL http://dx.doi.org/10.1186/s13014-021-01923-1.

[13] Chung SY, Chang JS, Choi MS, Chang Y, Choi BS, Chun J, Keum KC, Kim JS, Kim YB. 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). URL http://dx.doi.org/10.1186/s13014-021-01771-z.

[14] Liu Z, Liu F, Chen W, Tao Y, Liu X, Zhang F, Shen J, Guan H, Zhen H, Wang S, Chen Q, Chen Y, Hou X. Automatic segmentation of clinical target volume and organs-at-risk for breast conservative radiotherapy using a convolutional neural network. Cancer Manag Res 2021;Volume 13:8209–17, URL http://dx.doi.org/10.2147/cmar.s330249.

[15] Baroudi H, Huy Minh Nguyen CI, Maroongroge S, Smith BD, Niedzielski JS, Shaitelman SF, Melancon A, Shete S, Whitaker TJ, Mitchell MP, Yvonne Arzu I, Duryea J, Hernandez S, El Basha D, Mumme R, Netherton T, Hoffman K, Court L. Automated contouring and statistical process control for plan quality in a breast clinical trial. Phys Imaging Radiat Oncol 2023;28:100486, URL http://dx.doi.org/10.1016/j.phro.2023.100486.

[16] Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing (Amst) 2019;338:34–45, URL http://dx.doi.org/10.1016/j.neucom.2019.01.103.

[17] Poortmans PM, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: The use of artificial intelligence to individualise radiation therapy for breast cancer. Breast J 2020;49:194–200, URL http://dx.doi.org/10.1016/j.breast.2019.11.011.

[18] Zhong Y, Guo Y, Fansg Y, Wu Z, Wang J, Hu W. Geometric and dosimetric evaluation of deep learning based auto-segmentation for clinical target volume on breast cancer. J Appl Clin Med Phys 2023;24(7). URL http://dx.doi.org/10.1002/acm2.13951.

[19] Lee Bm, Kim JS, Chang Y, Choi SH, Park JW, Byun HK, Kim YB, Lee IJ, Chang JS. Experience of implementing deep learning-based automatic contouring in breast radiation therapy planning: Insights from over 2, 000 cases. Int J Radiat Oncol Biol Phys 2024. URL http://dx.doi.org/10.1016/j.ijrobp.2024.02.041.

[20] Best clinical practice nella radioterapia dei tumori della mammella 2019, gruppo di coordinamento AIRO mammella. 2019.

[21] Ciabattoni A, Gregucci F, De Rose F, Falivene S, Fozza A, Daidone A, Morra A, Smaniotto D, Barbara R, Lozza L, Vidali C, Borghesi S, Palumbo I, Huscher A, Perrucci E, Baldissera A, Tolento G, Rovea P, Franco P, De Santis MC, Grazia AD, Marino L, Meduri B, Cucciarelli F, Aristei C, Bertoni F, Guenzi M, Leonardi MC, Livi L, Nardone L, De Felice F, Rosetto ME, Mazzuoli L, Anselmo P, Arcidiacono F, Barbarino R, Martinetti M, Pasinetti N, Desideri I, Marazzi F, Ivaldi G, Bonzano E, Cavallari M, Cerreta V, Fusco V, Sarno L, Bonanni A, Mangiacotti MG, Prisco A, Buonfrate G, Andrulli D, Fontana A, Bagnoli R, Marinelli L, Reverberi C, Scalabrino G, Corazzi F, Doino D, Di Genesio-Pagliuca M, Lazzari M, Mascioni F, Pace MP, Mazza M, Vitucci P, Spera A, Macchia G, Boccardi M, Evangelista G, Sola B, La Porta MR, Fiorentino A, Levra NG, Ippolito E, Silipigni S, Osti MF, Mignogna M, Alessandro M, Ursini LA, Nuzzo M, Meattini I, D'Ermo G. AIRO breast cancer group best clinical practice 2022 update. Tumori 2022;108(2_suppl):1–144, URL http://dx.doi.org/10.1177/03008916221088885.

[22] Fodor A, Brombin C, Mangili P, Borroni F, Pasetti M, Tummineri R, Zerbetto F, Longobardi B, Perna L, Dell'Oca I, Deantoni CL, Deli AM, Chiara A, Broggi S, Castriconi R, Esposito PG, Slim N, Passoni P, Baroni S, Villa SL, Rancoita PM, Fiorino C, Del Vecchio A, Bianchini G, Gentilini OD, Di Serio MS, Di Muzio N. Impact of molecular subtype on 1325 early-stage breast cancer patients homogeneously treated with hypofractionated radiotherapy without boost: Should the indications for radiotherapy be more personalized? Breast J 2021;55:45–54, URL http://dx.doi.org/10.1016/j.breast.2020.12.004.

[23] Fodor A, Brombin C, Mangili P, Tummineri R, Pasetti M, Zerbetto F, Longobardi B, Galvan AS, Deantoni CL, Dell'Oca I, Castriconi R, Esposito PG, Deli AM, Rancoita PMV, Fiorino C, Vecchio AD, Di Serio MS, Di Muzio NG. Toxicity of hypofractionated whole breast radiotherapy without boost and timescale of late

skin responses in a large cohort of early-stage breast cancer patients. Clin Breast Cancer 2022;22(4):e480–7, URL http://dx.doi.org/10.1016/j.clbc.2021.11.008.

[24] Cicchetti A, Mangili P, Fodor A, Gabellini MGU, Chiara A, Deantoni C, Mori M, Pasetti M, Palazzo G, Rancati T, del Vecchio A, Gisella Di Muzio N, Fiorino C. Skin dose-volume predictors of moderate-severe late side effects after whole breast radiotherapy. Radiother Oncol 2024;194:110183, URL http://dx.doi.org/10.1016/j.radonc.2024.110183.

[25] Ubeira-Gabellini MG, Mori M, Palazzo G, Cicchetti A, Mangili P, Pavarini M, Rancati T, Fodor A, del Vecchio A, Di Muzio NG, Fiorino C. Comparing performances of predictive models of toxicity after radiotherapy for breast cancer using different machine learning approaches. Cancers 2024;16(5). http://dx.doi.org/10.3390/cancers16050934, URL https://www.mdpi.com/2072-6694/16/5/934.

[26] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. 2019, http://dx.doi.org/10.1145/3292500.3330701.

[27] Esposito PG, Castriconi R, Mangili P, Fodor A, Pasetti M, Di Muzio NG, Del Vecchio A, Fiorino C. Virtual tangential-fields arc therapy (ViTAT) for whole breast irradiation: Technique optimization and validation. Phys Med 2020;77:160–8, URL http://dx.doi.org/10.1016/j.ejmp.2020.08.011.

[28] Castriconi R, Esposito PG, Tudda A, Mangili P, Broggi S, Fodor A, Deantoni CL, Longobardi B, Pasetti M, Perna L, del Vecchio A, Di Muzio NG, Fiorino C. Replacing manual planning of whole breast irradiation with knowledge-based automatic optimization by virtual tangential-fields arc therapy. Front Oncol 2021;11. URL http://dx.doi.org/10.3389/fonc.2021.712423.

[29] Tudda A, Castriconi R, Benecchi G, Cagni E, Cicchetti A, Dusi F, Esposito PG, Guernieri M, Ianiro A, Landoni V, Mazzilli A, Moretti E, Oliviero C, Placidi L, Rambaldi Guidasci G, Rancati T, Scaggion A, Trojani V, Fiorino C. Knowledge-based multi-institution plan prediction of whole breast irradiation with tangential fields. Radiother Oncol 2022;175:10–6, URL http://dx.doi.org/10.1016/j.radonc.2022.07.012.

[30] Castriconi R, Tudda A, Placidi L, Benecchi G, Cagni E, Dusi F, Ianiro A, Landoni V, Malatesta T, Mazzilli A, Meffe G, Oliviero C, Rambaldi Guidasci G, Scaggion A, Trojani V, del Vecchio A, Fiorino C. Inter-institutional variability of knowledge-based plan prediction of left whole breast irradiation. Phys Med 2024;120:103331, URL http://dx.doi.org/10.1016/j.ejmp.2024.103331.

[31] Buelens P, Willems S, Vandewinckele L, Crijns W, Maes F, Weltens C. Clinical evaluation of a deep learning model for segmentation of target volumes in breast cancer radiotherapy. Radiother Oncol 2022;171:84–90. http://dx.doi.org/10.1016/j.radonc.2022.04.015, URL https://www.sciencedirect.com/science/article/pii/S0167814022002213.

[32] Almberg SS, Lervåg C, Frengen J, Eidem M, Abramova TM, Nordstrand CS, Alsaker MD, Tøndel H, Raj SX, Wanderås AD. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol 2022;173:62–8. http://dx.doi.org/10.1016/j.radonc.2022.05.018, URL https://www.sciencedirect.com/science/article/pii/S0167814022022691.

[33] Mikalsen SG, Skjøtskift T, Flote VG, Hämäläinen NP, Heydari M, Rydén-Eilertsen K. Extensive clinical testing of deep learning segmentation models for thorax and breast cancer radiotherapy planning. Acta Oncol 2023;62(10):1184–93, URL https://doi.org/10.1080/0284186X.2023.2270152. PMID: 37883678.

[34] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15(1). URL http://dx.doi.org/10.1186/s12880-015-0068-x.

[35] Sherer MV, Lin D, Elguindi S, Duke S, Tan L-T, Cacicedo J, Dahele M, Gillespie EF. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol 2021;160:185–91. http://dx.doi.org/10.1016/j.radonc.2021.05.003, URL https://www.sciencedirect.com/science/article/pii/S0167814021062289.

[36] Mackay K, Bernstein D, Glocker B, Kamnitsas K, Taylor A. A review of the metrics used to assess auto-contouring systems in radiotherapy. Clin Oncol (R Coll Radiol) 2023;35(6):354–69. http://dx.doi.org/10.1016/j.clon.2023.01.016, URL https://www.sciencedirect.com/science/article/pii/S0936655523000213.