

Clinical validation of AI-based probabilistic CTV breast maps incorporating Inter-Observer Variability

M.G. Ubeira-Gabellini¹, C. Riani^{1,2}, G. Palazzo¹, D. Monticelli¹, A. Fodor³, N. G. Di Muzio^{3,4}, A. del Vecchio¹,

A. Balsamo⁵, A. Coniglio⁵, C. Fiorino¹, A. Palma⁶,

¹ Medical Physics, IRCCS San Raffaele Scientific Institute, Milan, Italy

² Radiation Biophysics and Radiobiology Laboratory, Physics Department, University of Pavia, Pavia, Italy

³ Radiation Oncology, IRCCS San Raffaele Scientific Institute, Milan, Italy

⁴ Vita-Salute San Raffaele University, Milan, Italy

⁵ Ministry of Health, Department of Human Health, Animal Health and Ecosystem (One Health) and International Relations (DOHRI), Rome, Italy

⁶ Centro Nazionale Intelligenza Artificiale, e Tecnologie Innovative per la Salute, Istituto Superiore di Sanità, Rome, Italy

alessandra.palma@iss.it

Abstract

Manual segmentation of the Clinical Target Volume (CTV) in breast radiotherapy is time-consuming and affected by inter-observer variability (IOV). This study proposes a deep learning (DL) pipeline based on transfer learning (TL) to generate probabilistic clinician-based segmentation maps, enabling quantification of systematic IOV, and validates it on an independent multi-observer cohort.

A dataset of 961 patients treated at San Raffaele Institute (2017–2022) with breast-conserving surgery and radiotherapy was analyzed. An in-house UNet (MONAI v1.3), previously validated [1], was fine-tuned using TL (200 epochs) on data from seven clinicians (≥ 70 cases each), generating clinician-specific models combined into probabilistic maps (14–100% isoproability range). Validation included 40 patients contoured by three expert clinicians and one resident. Agreement between probabilistic maps and clinical contours was assessed using ECDF (Empirical Cumulative Distribution Function) under/over-segmentation residuals, Dice Similarity Coefficient (DSC), and Average Surface Distance (ASD). Dosimetric impact was evaluated on 5 left-breast cases.

TL improved agreement with clinician contours compared to the baseline UNet across all metrics. The resulting 14–100% variability band was wider at cranial-caudal and lateral borders (mean volume difference $193 \pm 103 \text{ cm}^3$), consistent with expected IOV. ECDF reached 0.9 at ~ 9 –10% residuals. Expert contours largely fell within the predicted variability band, while the resident showed larger deviations, mainly along the cranial-caudal axis. Dosimetric evaluation confirmed that heart and lung DVHs remained within the predicted variability range.

The proposed TL approach provides robust and reliable probabilistic segmentation maps providing direct IOV impact estimation, alongside the mean auto-contour. The present study confirms that the model accurately captures IOV in a test cohort with multiple observers delineating the same patients. Applications encompass clinical trials QA, contour quality assessment, tutoring and education.

1] Ubeira-Gabellini MG et al. *Phys Imaging Radiat Oncol.* 2025;34:100749

Funding: Supported by CCM 2024 (Ministry of Health)