ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Towards meaningful evaluation of uncertainty-aware segmentation workflows for medical applications

Dany Rimez[1,2], John A. Lee[1,2] and Ana Maria Barragan-Montero[1,2] *

1- UCLouvain – IREC – MIRO
Avenue Hippocrate 55 B1.54.07, 1200 Brussels - Belgium
2- UCLouvain – ICTEAM

Abstract. In radiation oncology, image segmentation with deep learning models must reduce clinician workload without compromising patient safety. Uncertainty quantification is therefore essential to provide reliable error estimates and to determine which segmentations require correction. Despite progress, we argue that current evaluation protocols enable only general comparison and ignore the definition of decision thresholds used in practice. We propose a new paradigm for the evaluation of such automated systems through the quantification of clinical outcomes. We compute the fraction of high-confidence segmentation prediction meeting quality standards and their corresponding average performance, based on a decision threshold. We calibrate this threshold to bound the amount of segmentations inferior to standards below a risk tolerance specified by clinicians. Through experiments across four medical datasets, we show our approach delivers meaningful performance guarantees essential for regulatory compliance and building trust in automated systems.

Keywords: Uncertainty quantification, Evaluation

1 Introduction

In radiotherapy treatment planning, automatic segmentation with deep learning reduces clinician workload and accelerates treatment workflows. However, reliability issues have raised interest in Uncertainty Quantification (UQ) to flag poor segmentation results for expert review, allowing safer and more efficient use of deep learning models [1]. We call the association of a segmentation model and an UQ method an Uncertainty-aware Segmentation Workflows (USw).

For a given prediction, an UQ method will estimate uncertainty and flag this prediction for review if it exceeds a given uncertainty threshold. A perfect UQ method should identify all predictions that are below a certain quality.

In reality, UQ methods are not perfect (i.e. no identity relationship, and some erroneous predictions are associated to low uncertainty), and USw are rarely actionable in clinical practice [2]. Current approaches aim to evaluate USw with summary metrics across all possible decision thresholds (e.g., AURC [3] or correlation). While valid to compare USw in general, these metrics are not

*D.R. is a PhD Fellow funded by Télévie grant 7.4511.25. A.M.B.M. is a Research Associate with the Belgian F.R.S.-FNRS. J.A.L. is a Research Director with the F.R.S.-FNRS.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

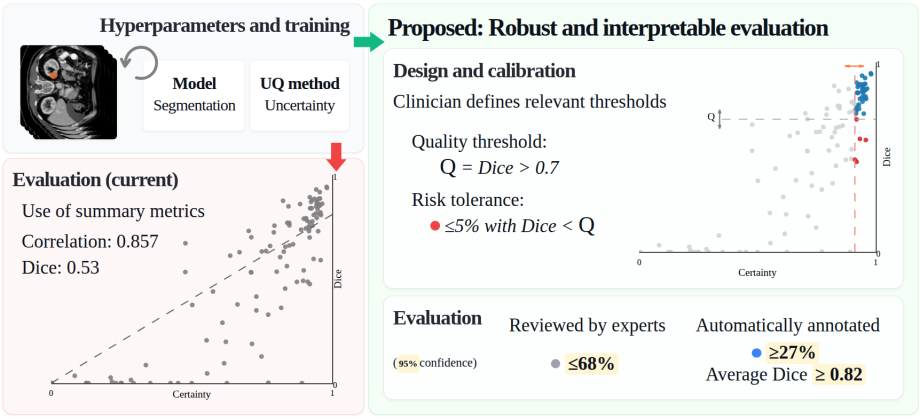

Fig. 1: Illustration of our evaluation workflow. We chose the Dice score to quantify segmentation quality. Graphs are generated from results of our use-case: TTA with the 0.2 dropout model on ADBO1k (see Section 2).

suitable for predicting practical outcomes like the number of images that can be automatically segmented without compromising patient safety.

To bridge these gaps, we propose a new paradigm for the design and evaluation of USw. Our approach, illustrated in Fig. 1, first calibrates a decision threshold on uncertainty under a specified risk tolerance. Then, the automation gain quantifies the amount of confident predictions that can be trusted at this tolerance level.

Confidence intervals can be computed for both gain and segmentation performance of the trusted predictions to obtain a robust lower bound on expected performance in small datasets. Quantifying these variables is also essential to satisfy regulatory frameworks like the EU AI Act and clinical trial guidelines [4], which demand quantifiable performance indicators and safety guarantees for automated systems.

With experiments in four medical image datasets, we demonstrate how our paradigm not only supports the design and evaluation of USw, but also helps build trust in automated systems with interpretable metrics and statistical guarantees.

By convention, we will talk about certainty rather than uncertainty in the rest of the paper. Section 2 details our safety-constrained threshold calibration and evaluation protocol, Section 3 describes the experimental setup and baselines, and Section 4 presents results and analysis.

2 Proposed evaluation protocol

In a USw, a segmentation is accepted if its estimated certainty is above a threshold th ; otherwise, it is flagged for expert review. Reducing th increases the amount of predictions that are not reviewed (automation), but also increases

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Algorithm 1 Decision threshold calibration and USw evaluation algorithm.

With samples S $\leftarrow \{(certainty_i, quality_i)\}_{i=1}^{N}$, quality threshold $\mathcal{Q}$, risk tolerance $\mathcal{R}$, confidence level α, N_b bootstrap resampling

Define candidate thresholds set $Th \leftarrow \{\, c \mid (c, q) \in S, q < \mathcal{Q}\}$
Define gain function $g(\text{th}) = \mathbb{P}(c \geq \text{th}, \text{q} \geq \mathcal{Q})$
Define risk function $r(\text{th}) = \mathbb{P}(c \geq \text{th}, \text{q} < \mathcal{Q})$
Find $\mathcal{C}^* \in Th$ that maximizes $r(\mathcal{C}^*)$ s.t. $r(\mathcal{C}^*) \leq \mathcal{R}$
for $b = 1$ to N_b **do**
    Sample population B $\leftarrow \{(c_{i_j},\ q_{i_j}) \mid (c_{i_j}, q_{i_j}) \in S,\ i_j \sim \text{Uniform}\{1, \ldots, N\}\}$
    Compute $\mathcal{G}_b = g(\mathcal{C}^*)$ in population B
    Compute $\mathcal{Q}uality_b^{\text{accurate}} = \text{average}(\{q_{b_i} \mid c_{b_i} \geq \mathcal{C}^*, q_{b_i} \geq \mathcal{Q}, (c_{b_i}, q_{b_i}) \in B)$
Compute automation gain: $\mathcal{G}^{\text{low}} = \text{Quantile}_{1-\alpha}(\{\, \mathcal{G}_b \,\})$
Compute segmentation performances: $\mathcal{Q}uality_{\text{low}}^{\text{accurate}} = \text{Quantile}_{1-\alpha}(\{\, \mathcal{Q}uality_b^{\text{accurate}} \,\})$

the risk of accepting a poor segmentation. The goal is then to maximize useful automation while limiting this risk, through the definition of a clinically motivated quality threshold $\mathcal{Q}$ over a segmentation quality metric (e.g., Dice score).

We then need to define the automation gain function $g(\text{th})$ and the risk function $r(\text{th})$ as the proportion of predictions that are not reviewed and of high or low quality, respectively. Formally:

$$g(t) = \mathbb{P}(\text{certainty} \geq t, \text{Quality} \geq \mathcal{Q}); \quad r(t) = \mathbb{P}(\text{certainty} \geq t, \text{Quality} < \mathcal{Q})$$

Using the Dice score to quantify the segmentation quality (e.g. $\mathcal{Q} = 0.8$), and given a risk tolerance $\mathcal{R}$ (e.g., 0.05 or 5% failure rate), we select the optimal certainty threshold $\mathcal{C}^*$ and the corresponding automation gain $\mathcal{G}^*$:

$$\mathcal{C}^* := \arg\max_t g(t) \quad \text{subject to} \quad r(t) \leq \mathcal{R}, \qquad \mathcal{G}^* = g(\mathcal{C}^*)$$

Low bound estimation. Because estimates of $\mathcal{G}^*$ depend on limited test data, we compute a lower confidence bound using bootstrap resampling to ensure robustness under sampling variability [5]. With confidence level $\alpha = 0.95$, this gives the minimum automation gain we can confidently expect in real-world deployment, under the assumption that real-world data distribution is similar to the one of the test set.

Segmentation performance. Similarly, the USw can be further evaluated by computing the confidence interval of the average Dice scores of accepted segmentation with sufficient quality $\text{Dice}^{\text{accurate}}$: Dice $\geq \mathcal{Q}$. This allows us to confidently estimate the average segmentation performance on accepted images $\text{Dice}_{\text{low}}^{\text{accurate}}$, while knowing it will always be higher than $\mathcal{Q}$.

Algorithm 1 summarizes the complete procedure. Taken together, $\mathcal{C}^*$, $\mathcal{G}^{\text{low}}$ and $\text{Dice}_{\text{low}}^{\text{accurate}}$, form the basis of our evaluation protocol and guarantees the following, with example values from Figure 1 ($\mathcal{R} = 0.05$, $\mathcal{Q} = 0.7$, confidence of 95%, $\mathcal{G}^{\text{low}}$ of 27% and $\text{Dice}_{\text{low}}^{\text{accurate}}$ of 0.82):

> **Less than** 5% of automated segmentations will have a Dice < 0.7.
>
> With 95% confidence, **at least** 27% of automated segmentations will have a Dice ≥ 0.7, and even **higher than** 0.82 on average.
>
> Remaining $\leq 68\%$ of automated segmentations will be flagged for expert review.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

3 Experimental setup

Models (Residual U-Net) were trained for tumor segmentation on four datasets: BraTS[1] (brain MRI), Pancreas[1], LUNG[2], and ABDO1k[3] (CT scans).

We evaluate four UQ methods for segmentation: Monte Carlo Dropout (MCDO)[6] with 20 inference passes (dropout was added in the residual blocks of the model's encoder with rates ranging from 0.1 to 0.5); Deep Ensembles (DE)[7] using 10 models trained from distinct initializations; Test-Time Augmentation (TTA)[8] with 8 input variants using training data augmentations. Certainty of the model for the average prediction is then quantified using the average of the Dice scores between it and individual predictions [9].

A last method is used, Out-Of-distribution Detection (OOD) using input reconstruction similarly to [10]. Here, a reconstruction branch is added to the model and takes the skip connections from the encoder as input. A first inference produces the segmentation prediction and a reconstruction of the input. A second inference produces another segmentation used to quantify certainty by taking the Dice between the two segmentations.

We compare our calibration-evaluation method with current evaluation paradigms: Spearman correlation and Dice (separate evaluation of UQ and segmentation model), and AURC [3]. AURC integrates the segmentation error across all possible certainty thresholds.

Code and supplementary material, like training schedules and data preprocessing are available, at `https://github.com/Dany546/esann2026`.

4 Results

We first evaluate UQ methods through Spearman correlation between certainty and segmentation error, and the average Dice score (Fig. 2). On ABDO1k, correlation ranks DE the best of the UQ methods, and so does the average Dice. AURC brings the same conclusions, with the interest of being single-metric and measuring if a UQ method is better than random selection. However, none of these approaches answers the question: 'At a 5% failure tolerance, what fraction of segmentation results can be trusted without review ?'.

However, when evaluated under clinical safety constraints (Fig. 3) with $\mathcal{Q} = 0.7$, we can see that the models achieve a high automation gain for BraTS ($\mathcal{G}_{\text{low}} > 0.5$), but much lower for other datasets. For Pancreas, most USw cannot achieve a $\mathcal{G}_{\text{low}}$ higher than the safety constraint, i.e. the gain is lower than the risk taken and the USw is unsafe. While differences between datasets could be anticipated from the differences in Dice (Fig. 2), such gaps could not be quantified by current evaluation approaches.

Moreover, the comparison between UQ methods significantly depends on both $\mathcal{Q}$ and $\mathcal{R}$, highlighting the importance of evaluating USw with carefully

[1] http://medicaldecathlon.com/

[2] https://www.cancerimagingarchive.net/collection/nsclc-radiomics/

[3] https://github.com/JunMa11/AbdomenCT-1K

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

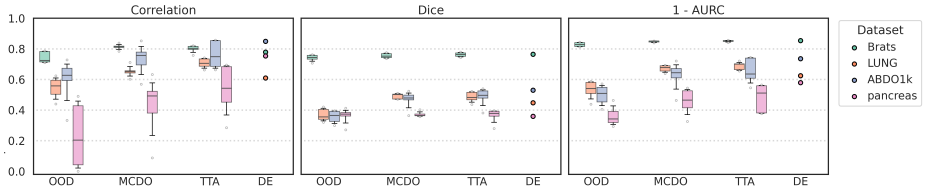

Fig. 2: Evaluation of UQ methods, across all models and methods. For illustration purposes, fliers are the minimum and maximum values while IQR marks are unconventionally replaced here with the second and penultimate values. We showed 1-AURC instead of AURC.

chosen thresholds, or to use several thresholds in the evaluation depending on the application. With a higher quality threshold ($\mathcal{Q} = 0.8$), the automation gain is heavily reduced even for BraTS ($\mathcal{G}_{\text{low}} \leq 0.4$).

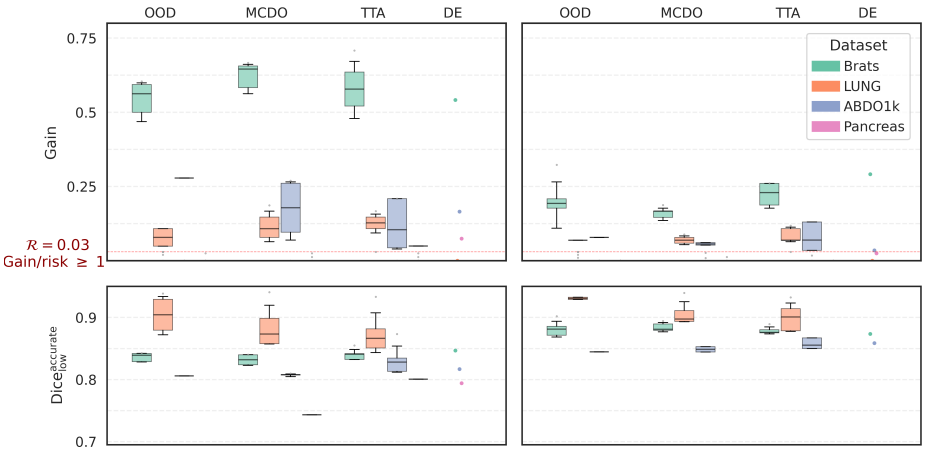

Fig. 3: We fixed $\mathcal{R} = 0.03$, and $\mathcal{Q} = 0.7$ (left) or $\mathcal{Q} = 0.8$ (right). For illustration purposes, only models with $\mathcal{G}_{\text{low}} \geq \mathcal{R}$ are used to generate the boxes according to the same convention as Fig. 2.

Finally, $\text{Dice}_{\text{low}}^{\text{accurate}}$ is much higher than average Dice in all datasets and even significantly higher than the quality threshold used (0.7 or 0.8). This shows that the average Dice of automated segmentation will be confidently and significantly higher than the specified threshold, providing a supplementary measure to build trust in the USw.

5 Conclusion

We introduce a new paradigm for the design and evaluation of Uncertainty-aware Segmentation Workflows under explicit clinical safety constraints. This

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

paradigm delivers interpretable performance evaluation with statistical guarantees.

Our approach expresses performance directly in terms of automation gain subject to user-defined safety constraints, providing an actionable measure of deployment readiness, and supporting model selection. To avoid over-optimistic conclusions on small datasets, we report a conservative 95% lower confidence bound on automation gain. It enables application-specific evaluation through configurable quality thresholds, risk tolerance, and gain–risk functions, addressing practical questions that conventional evaluations cannot answer, like 'How many cases can be auto-accepted without exceeding 5% failures?'.

Validation on four medical imaging datasets reveals critical limitations of standard performance reporting. Despite high Dice scores and strong correlation, our analysis identifies workflows that may remain clinically unsafe depending on the selected threshold and tolerated risk.

By anchoring evaluation to application-specific practical outcomes, this paradigm advances toward meaningful and trustworthy evaluation of automated systems in radiotherapy, and represents a first step toward evaluation procedures aligned with emerging regulatory requirements such as the EU AI Act for medical AI.

References

[1] M. et al Huet-Dastarac. Quantifying and visualising uncertainty in deep learning-based segmentation for radiation therapy treatment planning: What do radiation oncologists and therapists want? *Radiotherapy and Oncology*, 201:110545, December 2024.

[2] Stine Sofia Korreman and Jintao Ren. Understanding and leveraging uncertainties in autosegmentation for radiotherapy. *BJR—Artificial Intelligence*, 2(1):ubaf013, 07 2025.

[3] M. et al Zenk. Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. *Medical Image Analysis*, 101:103392, April 2025.

[4] European Parliament and Council of the European Union. Eu ai act or artificial intelligence act, 2024.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), January 1979.

[6] Y. et al. Gal. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[7] B. et al. Lakshminarayanan. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[8] M. S. Ayhan and P. Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *International Conference on Medical Imaging with Deep Learning*, 2018.

[9] Ng M. et al. Estimating uncertainty in neural networks for cardiac mri segmentation: A benchmark study. *IEEE Transactions on Biomedical Engineering*, 70:1955–1966, 2020.

[10] M. et al Huet-Dastarac. Can input reconstruction be used to directly estimate uncertainty of a dose prediction u-net model? *Medical Physics*, 51(10):7369–7377, July 2024.