Lost in the Folds: When Cross-Validation Is Not a Deep Ensemble for Uncertainty Estimation

Tristan Kirscher[1,2,3]*, Markus Bujotzek[3,4], Yannick Kirchhoff[3,5], Maximilian Rokuss[3,5], Fabian Isensee[3,6], Kim-Celine Kahl[3,5]†, Balint Kovacs[3,4]†, and Klaus Maier-Hein[3,7]†

[1] ICube Laboratory, CNRS UMR-7357, University of Strasbourg, Strasbourg, France
[2] CLCC Institut-Strauss, Strasbourg, France
[3] German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany
[4] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[5] Faculty of Mathematics and Computer Science, University of Heidelberg, Germany
[6] Helmholtz Imaging, German Cancer Research Center, Heidelberg, Germany
[7] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany

Abstract. Ensemble disagreement is widely used as a proxy for epistemic uncertainty in medical image segmentation. In practice, many studies form ensembles via K-fold cross-validation (CV), yet refer to them as "deep ensembles" (DE). Because CV members are trained on different data subsets, their disagreement mixes seed-driven variability with data-exposure effects, which can change how uncertainty should be interpreted. We audit recent segmentation uncertainty studies and find that terminology–implementation mismatches are common. We then compare a standard 5-fold CV ensemble to a 5-member DE (fixed training set, different random seeds) under otherwise identical configurations on three multi-rater segmentation datasets spanning three modalities. We evaluate uncertainty for calibration, failure detection, ambiguity modeling, and robustness under distribution shift. DE match segmentation accuracy while improving calibration and failure detection, whereas CV ensembles sometimes correlate more strongly with inter-rater variability on the studied datasets. Thus, ensemble construction should be chosen to match the research question: DE for reliability-oriented use (e.g., selective referral/failure detection) and CV ensembles as a proxy for ambiguity. We provide a lightweight nnU-Net modification enabling DE training within the default pipeline; code is available at https://github.com/Kirscher/LostInFolds.

Keywords: Uncertainty Quantification · Deep Ensembles · Medical Image Segmentation

* Corresponding author: tristan.kirscher@unistra.fr
†These authors share last authorship.

1 Introduction

Ensemble methods are widely used in medical image segmentation for robustness and uncertainty quantification, particularly in reliability-critical clinical pipelines. Among these methods, deep ensembles (DE), i.e. multiple instances of the same model trained on the same training set with different random initializations, have emerged as a simple and effective approximation to Bayesian model averaging [17,21]. In high-capacity neural networks, deep ensembles consistently outperform alternatives such as Monte Carlo dropout in terms of calibration, robustness under dataset shift, and failure detection [21].

In parallel, nnU-Net has become a de facto standard framework for medical image segmentation due to its automated configuration, strong empirical performance, and broad adoption across modalities and tasks [9,10]. By default, nnU-Net employs five-fold cross-validation (CV) to estimate in-distribution performance and averages fold-specific predictions at inference time to maximize segmentation accuracy. Practitioners often reuse these CV ensembles for uncertainty estimation rather than retraining deep ensembles (DE).

Increasingly, disagreement among these CV models is reused as a proxy for epistemic uncertainty in downstream applications such as quality control and human-in-the-loop review [11,18,15]. However, CV ensembles and DE differ in a fundamental and often overlooked way. In a deep ensemble, all models are trained on the same training set, and prediction variability reflects uncertainty in model parameters given the available data. In contrast, each model in a CV ensemble is trained on a different subset of the training data. As a result, disagreement within a CV ensemble conflates epistemic uncertainty with variability induced by incomplete or heterogeneous data exposure. This distinction is particularly relevant in medical imaging, where rare structures, underrepresented pathologies, or uneven fold composition can cause elevated disagreement that reflects missing training examples rather than genuine uncertainty under the full data distribution [21]. Our goal is to investigate these ensemble formulations to determine whether they induce systematically different uncertainty behavior, and which is preferable for a given downstream task.

In practice, uncertainty estimates derived from segmentation models are often thresholded, ranked, or aggregated to trigger concrete actions, such as selective referral to human experts [15], automated failure detection [27], or quality control and identification of reduced model performance [3,1]. Therefore, it is crucial to develop an understanding of how different ensemble constructions affect these follow-up tasks.

Contributions. This paper clarifies the distinction between CV ensembles and DE and addresses their conflation in the literature. The contributions are:

1. **Terminology audit.** We document, via a survey of uncertainty in segmentation papers, that the term deep ensemble is frequently used to describe a K-fold CV ensemble, similar to the default nnU-Net behavior, despite its members being trained on different data subsets (Table 1). We therefore distinguish CV ensembles from DE (same data, different seeds) throughout.

2. **Controlled comparison across downstream tasks.** Using the same network architecture and hyperparameters, we compare 5-fold CV ensembles against 5-member DE on multi-rater datasets spanning different modalities and evaluate segmentation, calibration, ambiguity modeling, failure detection, and shift robustness under a unified protocol (Table 3, Fig. 1-2).
3. **Practical implementation.** We provide a minor modification to the nnU-Net pipeline to support deep ensembles as originally described in [17].
4. **Task-dependent recommendations.** We show that the ensemble construction materially affects uncertainty behavior: DE provide more reliable calibration and failure detection at comparable Dice, while CV can better reflect annotation ambiguity in some settings. Consequently, we provide task-dependent recommendations for using and reporting ensemble uncertainty in reliability-oriented segmentation studies.

Table 1. Use of ensemble methods for uncertainty estimation in segmentation studies (2020–2025). Papers were identified via a literature search, followed by manual screening. SD: same training set across ensemble members. Align: method–claim alignment.

Paper	Task	Claim	Actual	SD	Align
Jungo et al. [11]	Brain MRI	DE	10-fold CV	✗	✗
Alves et al. [3]	Multi-organ CT	DE	5-fold CV	✗	✗
van Aalst et al. [1]	Head & Neck OAR CT	DE	5-fold CV	✗	✗
Khalili et al. [15]	CAMELYON16 (WSI)	DE	5-fold CV	✗	✗
Gotkowski et al. [6]	Multi-dataset (CT/MRI)	DE	5-fold CV	✗	✗
Schwab et al. [24]	Cardiac MRI (LGE)	DE	5-fold CV	✗	✗
Rosas-Gonzalez et al. [23]	Brain MRI	DE	5-fold CV	✗	✗
Göttlich et al. [7]	Multi-organ CT	CV	5-fold CV	✗	✓
Gade et al. [5]	Prostate MRI	CV	5-fold CV	✗	✓
Kucybała et al. [16]	Multi-dataset (CT/MRI)	CV	5-fold CV	✗	✓
Molchanova et al. [19]	Multiple sclerosis MRI	DE	DE	✓	✓
Zhao et al. [28]	Cardiac MRI	DE	DE	✓	✓
Zenk et al. [27]	Multi-dataset (CT/MRI)	DE	DE	✓	✓
Buddenkotte et al. [4]	Ovarian/Kidney CT	Ens.	Diff. losses	✓	✓

2 Related Work

Deep ensembles and epistemic uncertainty. Deep ensembles were introduced as a scalable and effective approach to epistemic uncertainty estimation by training multiple networks on the same dataset with different random initializations and stochastic optimization trajectories [17]. Despite their simplicity, deep ensembles have been shown to provide strong uncertainty estimates, often outperforming approximate Bayesian methods such as Monte Carlo dropout in terms of calibration, robustness, and out-of-distribution behavior [21,14,13].

Uncertainty estimation in medical image segmentation. Uncertainty quantification has been widely studied in medical image segmentation, motivated by applications such as quality control, error detection, and human-in-the-loop decision support [11,18]. Ensemble-based methods have consistently demonstrated improved calibration and error awareness compared to single-model predictions [18,8,13]. Recently, there has been an increased emphasis on systematic validation of uncertainty estimates with respect to concrete downstream applications. This is formalized in the ValUES framework by identifying calibration, failure detection, and ambiguity modeling as key uncertainty evaluation scenarios in semantic segmentation [13]. For calibration, spatially resolved evaluation metrics, such as BA-ECE have been proposed to better capture the alignment between uncertainty maps and segmentation errors [26]. Failure detection, on the other hand, is a task that is tackled at the subject-level, motivated by the need to flag whole subjects as failures in clinical deployment [12,11,27]. In ambiguity modeling, the goal is to capture data ambiguity, for example when raters provide different segmentation masks for the same image.

Ensembles in nnU-Net pipelines. By default, the widely adopted state-of-the-art segmentation framework, nnU-Net, trains models in a 5-fold CV in order to estimate in-distribution performance and then re-uses these models as a CV ensemble to boost test set performance [9]. However, as shown in Table 1, several recent studies implicitly or explicitly interpret disagreement among nnU-Net CV models as epistemic uncertainty, often referring to these as DE [11,3,1,15,6,24,23]. Only a limited number of works explicitly adhere to the DE definition by training all ensemble members on identical data with independent random seeds [28,19,27].

3 Problem Statement

We study supervised medical image segmentation with dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an image and y_i its mask. A segmentation network $f(x;\theta)$ predicts $p(y \mid x, \theta)$. We compare two ensemble constructions for uncertainty estimation:

Deep Ensembles (DE). A deep ensemble consists of M independently trained models $\{f(x;\theta_m)\}_{m=1}^{M}$, where each parameter vector θ_m is obtained by optimizing the same objective on the same training dataset $\mathcal{D}$, but with different random initializations and stochastic training [17]. Formally,

$$\theta_m \sim P(\theta \mid \mathcal{D}), \quad m = 1, \ldots, M, \tag{1}$$

where the randomness induced by initialization, data shuffling, and augmentation implicitly samples multiple modes of the weight posterior. The ensemble predictive distribution for a new input x is given by $p(y \mid x, \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^{M} p(y \mid x, \theta_m)$, and the variability among predictions $\{f(x;\theta_m)\}$ reflects the epistemic uncertainty of the model given the full dataset $\mathcal{D}$.

Cross-Validation (CV) Ensembles. In a K-fold CV ensemble, the dataset is partitioned into disjoint folds $\{\mathcal{D}_k^{\text{val}}\}_{k=1}^K$, with corresponding training subsets $\mathcal{D}_k^{\text{train}} = \mathcal{D} \setminus \mathcal{D}_k^{\text{val}}$. Each model $f(x;\theta_k)$ is trained on a different data subset:

$$\theta_k \sim P(\theta \mid \mathcal{D}_k^{\text{train}}), \quad k = 1, \ldots, K\,. \tag{2}$$

The ensemble predictive distribution is $p(y \mid x) \approx \frac{1}{K}\sum_{k=1}^K p(y \mid x, \theta_k)$, representing a mixture of models conditioned on different training sets. Disagreement between these models arises not only from posterior uncertainty, but also from the fact that each θ_k is optimized with incomplete data $\mathcal{D}_k^{\text{train}}$ (missing $\mathcal{D}_k^{\text{val}}$).

Key Distinction. DE aims to approximate Bayesian model averaging under a fixed dataset. In contrast, CV ensembles aggregate models trained on different and incomplete subsets of the data. Consequently, their disagreement reflects both epistemic uncertainty and variability induced by data subsampling. *A model may therefore appear uncertain due to limited exposure to specific training examples, rather than uncertainty under the full-data distribution.*

4 Experimental Setup

The experiments aim to isolate the effect of ensemble construction on uncertainty estimation. Specifically, we compare CV ensembles, as commonly used in nnU-Net-based studies, with DE that adhere to the original definition from [17].

Datasets. Ensemble uncertainty is evaluated on multi-rater medical image segmentation datasets spanning 2D and 3D modalities, including MRI, CT, and retinal fundus imaging (cf. Table 2). The datasets cover three types of consensus: manual expert delineations, STAPLE [25], and majority voting, thereby reflecting different annotation aggregation paradigms. All tasks provide independent rater annotations, enabling ambiguity analysis. Training uses all available rater annotations by treating each (image, rater) pair as a separate training example. To prevent leakage - i.e., the same image appearing in both training and validation with different rater labels - we generate splits with a K-fold procedure while grouping by image identity. Preprocessing follows the standard nnU-Net pipeline without task-specific modifications, isolating ensemble construction as the primary variable. For out-of-distribution (OOD) evaluation, dataset-specific splits were defined to induce clinically meaningful shifts. For CURVAS, Group C (cysts with contour-altering pathologies; n=23) is held out as OOD. For RIGA, images from the Magrabi Eye Center (n=95) were held out, representing a shift in acquisition center and population. For GoldAtlas, the patients from acquisition site 3 (n=4) were treated as OOD.

Network and Training Configuration. All experiments use nnU-Net (v2.4.1) full-resolution configuration with the `ResEncM` preset and default hyperparameters [9]. Network architecture, training schedule, and data augmentation are automatically determined by the nnU-Net framework for each task. Training is

Table 2. Overview of heterogeneous multirater-datasets used in this study, covering various medical modalities and different methods for generating consensus masks.

Dataset	Modality	Samples	Classes	Raters	Consensus
GoldAtlas [20]	T2w MRI (3D)	19	9	5	Manual (experts)
CURVAS [22]	CT (3D)	89*	4	3	STAPLE
RIGA [2]	Fundus RGB (2D)	749	2	6	Majority vote

*One case was excluded due to annotation spacing inconsistencies from one rater.

run for a fixed number of epochs with a fixed learning-rate schedule (no early stopping and no validation-based learning-rate adaptation). The validation split is used only for reporting, not for model selection; inference uses the final checkpoint for all models. This protocol is identical for CV and DE members, so differences reflect the ensemble construction (shared vs varying training subsets) rather than checkpointing or optimization artifacts.

Ensemble Configurations. We define two ensemble configurations using the same pool of training cases. For the **CV-Ensemble**, 5-fold CV is performed using nnU-Net's default data splits. Each model is trained on 80% of the data and validated on the remaining fold. For the **DE**, $M = 5$ models are trained independently on all available training data, differing only in random initialization and stochastic training effects (data shuffling, augmentation, and optimization). This configuration satisfies the defining criteria of a deep ensemble [17]. Apart from data exposure, all training and inference settings are identical across ensembles.

Uncertainty Quantification and Evaluation. During inference, voxel-wise softmax probability maps p_m are computed for each ensemble member m. Ensemble predictions are obtained by averaging class probabilities across models, $\bar{p} = \frac{1}{M}\sum_{m=1}^{M} p_m$, and the mean hard prediction is given by $\bar{y}(v) = \arg\max_c \bar{p}_c(v)$. To assess segmentation performance, we compute the Dice score (DSC) between the mean prediction $\bar{y}$ and the consensus segmentation $\bar{y}^*$, i.e., $\mathrm{DSC}(\bar{y}, \bar{y}^*)$. As calibration metrics, the Average Calibration Error (ACE) [11,13] and the Boundary-Aware Expected Calibration Error (BA-ECE) [26] are calculated on a voxel level, where the confidence is defined as $\mathrm{conf}(v) = \max_c \bar{p}_c(v)$. These confidences are then grouped into bins, and the accuracy in a bin B_m is calculated as $acc(B_m) = \frac{1}{|B_m|N}\sum_{v\in B_m}\sum_{n=1}^{N}[\,\bar{y}(v) = y_n^*(v)\,]$. Ambiguity modeling is evaluated using the Normalized Cross-Correlation (NCC) between the predictive entropy map $H(\bar{p})$ and the rater variance map [13], and the Generalized Energy Distance (GED) between the individual ensemble predictions $\{y_m\}_{m=1}^{M}$ and the individual rater annotations $\{y_n^*\}_{n=1}^{N}$ [13]. Failure detection is performed at a sample level following [27], where cases are ranked by inter-model disagreement, $u = \mathbb{E}_{i\neq j}[1 - \mathrm{DSC}(y_i, y_j)]$, and the risk is defined as $r = 1 - \mathrm{DSC}(\bar{y}, \bar{y}^*)$. Non-parametric bootstrap confidence intervals (CI) are used to compare uncertainty and agreement measures between CV ensembles and DE (see Table 3).

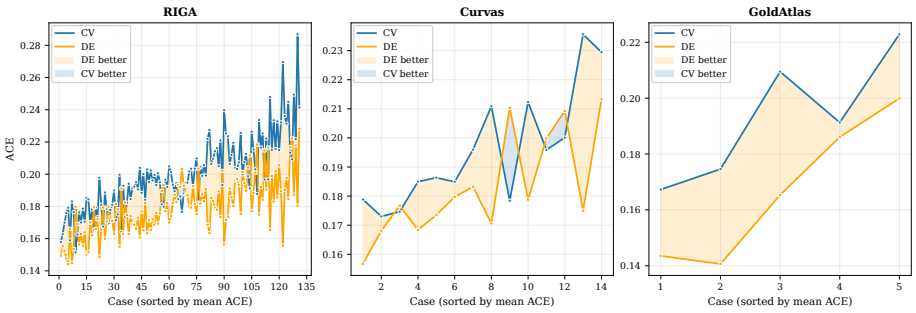

Fig. 1. Per-case Average Calibration Error (ACE) on the in-distribution test set for Cross-Validation (CV) and Deep Ensemble (DE) methods across all datasets.

5 Results

Table 3 summarizes performance and uncertainty across datasets, while Figs. 1–2 visualize calibration and failure detection behavior, respectively. Segmentation accuracy remains essentially unchanged: DE achieve DSC comparable to CV ensembles across all datasets. In contrast, uncertainty properties differ systematically. DE consistently improve voxel-wise calibration, reducing ACE across datasets (Figure 1) and yielding lower BA-ECE. At the sample level, DE provide stronger failure detection, reflected by lower AURC and reduced risk at matched coverage in referral curves (Figure 2). Interestingly, CV ensembles better align with inter-rater ambiguity on CURVAS and RIGA (higher NCC and/or lower GED), suggesting that fold-induced data variability can partially mimic annotation ambiguity. Under distribution shift, differences between CV and DE are generally modest. While DE show slight advantages in some settings, neither ensemble type consistently dominates across datasets, indicating comparable robustness under the considered OOD splits. Overall, these findings indicate that the appropriate ensemble construction depends on the downstream task: calibrated epistemic uncertainty for reliability-oriented decision-making favors DE, whereas ambiguity modeling may benefit from CV ensembles (Table 3).

6 Discussion and Conclusion

CV ensembles are often used for uncertainty estimation, not least in nnU-Net pipelines, because they are readily available and improve segmentation accuracy. Yet, our analysis shows that they are not interchangeable with DE for uncertainty estimation: since fold members are trained on different data subsets, their disagreement reflects both epistemic uncertainty and sensitivity to data exposure. This distinction is critical in reliability-oriented workflows that threshold or rank uncertainty for referral or failure detection. Empirically, DE achieve comparable segmentation performance while consistently improving calibration and failure detection. In contrast, CV ensembles better match inter-rater

Table 3. Comparison of CV vs DE performance. ID rows (CV, DE) report mean and 95% percentile bootstrap CIs over test cases. Paired bootstrap (B=10,000) is performed on per-case differences (DE−CV); for AURC, curves are recomputed per resample. The OOD row reports only the paired OOD Δ effect size as DE improvement $\Delta = (DE{-}CV)\times s$ with CI, where s=+1 if higher is better and s=−1 if lower is better. Stars indicate significance of the paired DE−CV comparison within the corresponding row ($^{*}p < 0.05$, $^{\dagger}p < 0.001$). Within each dataset, the better method (per metric direction) is **bolded**. Gray rows indicate the overall better-performing method according to this metric. SEG: Segmentation; CAL: Calibration; AM: Ambiguity Modeling; FD: Failure Detection. **All reported values are multiplied by 100 for compactness.**

Task	Metric	Setting	GoldAtlas	CURVAS	RIGA
SEG	DSC ↑	CV	84.6 (74.6, 88.4)	93.5 (91.5, 94.6)	93.1 (92.6, 93.6)
		DE	**85.2** (74.3, 88.9)	**93.6** (91.4, 94.8)	**93.2**† (92.7, 93.7)
		OOD Δ	+3.3 (-0.3, 8.8)	+0.4* (0.1, 0.7)	+0.1 (-0.6, 0.8)
CAL	ACE ↓	CV	19.3 (17.5, 21.1)	19.6 (18.7, 20.8)	19.9 (19.5, 20.3)
		DE	**16.7**† (14.7, 18.7)	**18.3*** (17.5, 19.3)	**17.9**† (17.7, 18.2)
		OOD Δ	+2.5 (-0.6, 6.0)	+1.5† (0.8, 2.2)	+0.8 (-0.4, 2.0)
	BA-ECE ↓	CV	7.1 (5.7, 8.8)	6.8 (6.1, 7.8)	21.5 (20.9, 22.0)
		DE	**6.4**† (5.2, 8.5)	**6.4**† (5.8, 7.4)	**20.6**† (20.1, 21.2)
		OOD Δ	+1.6 (0.6, 3.2)	+0.4† (0.3, 0.4)	+0.7† (0.4, 1.1)
AM	NCC ↑	CV	54.4 (51.6, 57.0)	**50.3**† (48.3, 53.1)	**73.7**† (72.5, 74.6)
		DE	**54.5** (51.1, 57.6)	49.2 (47.3, 51.8)	72.9 (71.8, 73.8)
		OOD Δ	+1.4 (-0.6, 3.3)	-0.7† (-0.9, -0.5)	-0.2 (-0.9, 0.5)
	GED ↓	CV	**16.2** (12.9, 21.2)	**8.8** (7.2, 11.3)	**8.2**† (7.7, 8.9)
		DE	16.8 (11.8, 29.9)	9.2 (7.5, 12.1)	8.6 (8.1, 9.3)
		OOD Δ	+2.7 (0.4, 6.0)	-0.1 (-0.4, 0.3)	-0.9 (-2.6, 0.7)
FD	AURC ↓	CV	9.7 (8.9, 13.8)	4.6 (3.9, 5.4)	5.7 (5.2, 6.2)
		DE	**9.3** (8.7, 13.9)	**4.4** (3.8, 5.2)	**5.5** (5.1, 5.9)
		OOD Δ	+2.6 (-0.7, 8.1)	+0.2 (0.0, 0.6)	-4.3 (-11.7, 0.3)

ambiguity on some datasets, suggesting that fold-induced data variability can mimic annotation uncertainty. Ensemble construction should therefore align with the intended downstream task: reliability-sensitive decision-making favors DE, whereas ambiguity modeling may benefit from CV ensembles. In practice, DE also incurs extra cost, as members must be trained in addition to the CV models typically produced during parameter optimization. We recommend that future studies (i) explicitly report the ensemble construction used for uncertainty estimation and (ii) select DE or CV ensembles based on the application. To facilitate this, we release a minimal nnU-Net change that enables deep ensembles support as described in [17]. Table 4 summarizes our task-dependent recommendations.

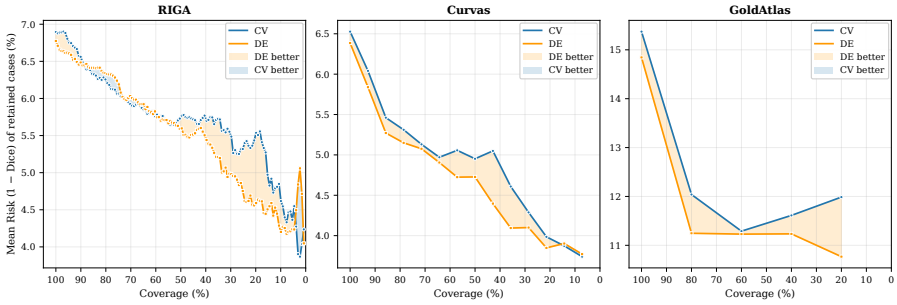

Fig. 2. Referral curves showing the mean risk (1−Dice) of in-distribution retained cases as a function of coverage for CV and DE methods across all datasets.

Table 4. Our ensemble construction recommendations for different uncertainty tasks.

Task	Deep Ensemble	CV Ensemble
Calibration	✓	
Ambiguity modeling		✓
Distribution shift robustness	✓	✓
Failure detection	✓	

Acknowledgments. This work of the ITI HealthTech, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm, was partially supported by IdEx Unistra (ANR-10-IDEX-0002) and SFRI (STRAT'US project, ANR-20-SFRI-0012) under the framework of the French Investments for the Future Program. T.K. research visit to the Medical Image Computing team at DKFZ was supported by a CNRS GdR IASIS grant and Erasmus funding. This work is also supported by the Helmholtz Association under the joint research program "HIDSS4Health - Helmholtz Information and Data Science School for Health".

Disclosure of interests. The authors have no competing interests to declare.

References

1. van Aalst, J.E., Maruccio, F.C., Simoẽs, R., Janssen, T.M., Wolterink, J.M., van Ooijen, P.M.A., Brouwer, C.L.: Reliability of uncertainty quantification methods for deep learning auto-segmentation in head and neck organs at risk. Physics in Medicine & Biology **70**(20), 205023 (2025)
2. Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., Lakshminarayanan, V.: Retinal fundus images for glaucoma analysis: the riga dataset (2018)
3. Alves, N., Bosma, J.S., Venkadesh, K.V., Jacobs, C., Saghir, Z., de Rooij, M., Hermans, J., Huisman, H.: Prediction variability to identify reduced ai performance in cancer diagnosis at mri and ct. Radiology **308**(3), e230275 (2023)

4. Buddenkotte, T., Escudero Sanchez, L., Crispin-Ortuzar, M., Woitek, R., McCague, C., Brenton, J.D., Öktem, O., Sala, E., Rundo, L.: Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. Computers in Biology and Medicine **163**, 107096 (2023)
5. Gade, M., Nguyen, K.M., Gedde, S., Fernandez-Quilez, A.: Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based mri prostate segmentation. Insights into Imaging **15**(1), 286 (2024)
6. Gotkowski, K., Gonzalez, C., Kaltenborn, I., Fischbach, R., Bucher, A., Mukhopadhyay, A.: i3Deep: Efficient 3d interactive segmentation with the nnu-net. In: Proc. of the International Conference on Medical Imaging with Deep Learning (MIDL). vol. 172, pp. 1–16 (2022)
7. Göttlich, H.C., Korfiatis, P., Gregory, A.V., Kline, T.L.: AI in the loop: Functionalizing fold performance disagreement to monitor automated medical image segmentation workflows. Frontiers in Radiology **3**, 1223294 (2023)
8. Huang, L., Ruan, S., Xing, Y., Feng, M.: A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. Medical Image Analysis (2023)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
10. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
11. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience **14**, 282 (2020)
12. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 48–56. Springer (2019)
13. Kahl, K.C., Lüth, C.T., Zenk, M., Maier-Hein, K.H., Jaeger, P.F.: Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. In: International Conference on Learning Representations (ICLR) (2024)
14. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in Neural Information Processing Systems **30** (2017)
15. Khalili, N., Spronck, J., Ciompi, F., van der Laak, J., Litjens, G.: Uncertainty-guided annotation enhances segmentation with the human-in-the-loop. arXiv preprint arXiv:2404.07208 (2024)
16. Kucybała, I., Rozynek, M., Krupa, K., Matusik, P., Jarczewski, J., Tabor, Z.: Evaluating uncertainty quantification in medical image segmentation: A multi-dataset, multi-algorithm study. Applied Sciences **14**(21), 10020 (2024)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
18. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging **40**(12), 3436–3447 (2021)
19. Molchanova, N., Gordaliza, P.M., Cagol, A., Ocampo-Pineda, M., Lu, P.J., Weigel, M., Chen, X., Beck, E.S., Tsagkas, H., Reich, D.S., Stölting, A., Maggi, P., Ribes, D., Depeursinge, A., Granziera, C., Müller, H., Bach Cuadra, M.: Explainability of

ai uncertainty: Application to multiple sclerosis lesion segmentation on mri. arXiv preprint arXiv:2504.04814 (2025)
20. Nyholm, T., Svensson, S., Andersson, S., Jonsson, J., Sohlin, M., Gustafsson, C., Kjellén, E., Söderström, K., Albertsson, P., Blomqvist, L., Zackrisson, B., Olsson, L.E., Gunnlaugsson, A.: Mr and ct data with multiobserver delineations of organs in the pelvic area–part of the gold atlas project. Medical Physics **45**(3), 1295–1300 (2018)
21. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems **32** (2019)
22. Riera-Marín, M., Kleiß, J.M., Aubanell, A., Antolín, A.: Curvas dataset: Calibration and uncertainty for multirater volume assessment in multi-organ segmentation (2024)
23. Rosas-Gonzalez, S., Birgui-Sekou, T., Hidane, M., Tauber, C.: Asymmetric ensemble of asymmetric u-net models for brain tumor segmentation with uncertainty estimation. Frontiers in Neurology **12**, 609646 (2021)
24. Schwab, M., Haltmeier, M., Mayr, A.: Disagreement-driven uncertainty quantification in late gadolinium enhancement cardiac mri. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 24–33. Springer (2025)
25. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (2004)
26. Zeevi, T., Lieffrig, E.V., Staib, L.H., Onofrey, J.A.: Spatially-aware evaluation of segmentation uncertainty. In: CVPR Workshop on Uncertainty Quantification for Computer Vision (2025), arXiv:2506.16589
27. Zenk, M., Zimmerer, D., Isensee, F., Traub, J., Norajitra, T., Jäger, P.F., Maier-Hein, K.: Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation. Medical Image Analysis **101**, 103392 (2025)
28. Zhao, Y., Yang, C., Schweidtmann, A., Tao, Q.: Efficient bayesian uncertainty estimation for nnu-net. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 13438, pp. 535–544. Springer (2022)