A Real-World Evaluation of Failure Detection for Liver CT Segmentation

Jeddy Bennett[1,2⋆][0009−0003−2546−1524], McKell Woodland[1⋆][0000−0003−0995−3695], Austin Castelo[1][0009−0009−3015−3570], Mais Altaie[1][0009−0000−9298−6644], Ajith Anthony[3][0009−0008−2265−8758], Noreen S. Siddiqi[3][0009−0002−9738−6431], James P. Long[4][0000−0001−5853−5938], and Kristy K. Brock[1][0000−0001−9364−5040]

[1] Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston TX 77030, USA
mewoodland@mdanderson.org
[2] Department of Mathematics, Brigham Young University, Provo UT 84602, USA
[3] Department of Interventional Radiology, The University of Texas MD Anderson Cancer Center, Houston TX 77030, USA
[4] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston TX 77030, USA

Abstract. Deep learning models deployed in clinical imaging frequently encounter distribution shifts, yet most out-of-distribution (OOD) detection methods are evaluated only on controlled research datasets. As a result, it is unclear whether existing approaches can reliably identify segmentation failures that arise in real-world clinical practice. We evaluated six OOD detection methods on a deployed liver CT segmentation model (3D nnU-Net) using internal data from 400 patients and external data from 100 patients collected across nearly 70 sites in 7 countries. One method was Pairwise Surface DSC, a surface-based extension of Pairwise DSC, that we introduced. OOD performance was measured using sensitivity, AUROC, and balanced accuracy, with thresholds determined on an independent cohort of 400 patients using the Youden J statistic. Statistical significance was assessed using McNemar tests and stratified bootstraps ($\alpha = 0.05$) with Benjamini–Hochberg correction. Pairwise Surface DSC was the top-performing method, with perfect sensitivities (1.00), near-perfect AUROCs (0.97 internal; 1.00 external), and the highest balanced accuracies (0.94 internal; 0.88 external; $p < 0.001$). These results show that automated failure detection for liver CT segmentation is clinically feasible and that Pairwise Surface DSC is a promising candidate for deployment. Our code is available at `https://github.com/mckellwoodland/liver_ct_ood_translation`.

Keywords: Out-of-Distribution Detection · Failure Detection · Liver Segmentation.

⋆ Equal contribution

1 Introduction

Liver cancer is the third leading cause of cancer death globally, with around 765,000 deaths and over 855,000 new cases in 2022 [1]. Effective management of liver disease relies on cross-sectional imaging [2], requiring accurate characterization of liver anatomy for volumetry, surgical planning [3], and lesion analysis. Automated liver segmentation is increasingly used to support these tasks, enabling consistent liver volume estimation [4], reducing the time required for manual annotation [5], and contributing to improved clinical endpoints such as ablative margin assessment [6].

Reliable segmentation is critical for patient safety, yet automated models often struggle with input features that were not present during training [7]. Liver segmentation is particularly vulnerable due to significant shape variability, especially in the presence of hepatic conditions [3]. For example, the only liver CT segmentations deemed clinically unacceptable in Anderson et al. [8] were on scans containing stents and ascites. Out-of-distribution (OOD) detection in safety-critical applications aims to identify such model failures [9].

While OOD detection has been extensively studied in theoretical settings [9], its effectiveness in real-world clinical workflows remains largely unexplored. Contemporary research often relies on curated datasets that fail to capture the complexity of clinical practice, and direct assessments of failure detection are rare [10]. We aim to address these gaps with the following contributions:

1. **Real-world OOD detection evaluation.** We assess failure detection for a deployed liver CT segmentation model using real-world clinical data, including validation on imaging collected from 73 distinct sites (8 countries).
2. **Introduction of Pairwise Surface DSC.** We present a surface-based adaptation of Pairwise DSC [10] that provides a contour-quality measure that is more consistent with clinical acceptability [11].

2 Materials and Methods

We conducted a retrospective evaluation of OOD methods for failure detection of a deployed liver CT segmentation model on clinical scans. The study was approved by the The University of Texas MD Anderson Cancer Center Institutional Review Board (PA18-032), with informed consent waived. Our code is available at `https://github.com/mckellwoodland/liver_ct_ood_translation`.

2.1 Data

We queried our internal database for liver-protocol CT studies acquired July to December 2024, resulting in 17,911 scans from 919 patients. We kept scans whose series description specified liver anatomy and removed non-axial, non-diagnostic, and partial-liver scans (1,890 scans remained; 892 patients). 400 unique patients were then separately sampled for validation and testing, selecting one scan per

Table 1. Demographics and imaging attributes, with medians or counts reported. Abbreviations: interquartile range (IQR), manufacturers (mfrs).

Attribute	Validation	Internal Test	External Test
Patients	400	400	100
Female (%)	213 (53)	206 (52)	42 (42)
Age (IQR)	67 (57-73)	66 (57-72)	58 (50-69)
Images	400	400	100
Voxel Size X/Y (IQR)	0.8 (0.7-0.9)	0.78 (0.74-0.82)	0.78 (0.70-0.87)
Slice Thickness (IQR)	3.0 (2.5-5.0)	3.0 (2.5-5.0)	2.0 (1.0-3.0)
Sites (Countries)	3 (1)	4 (2)	69 (7)
Scanner Models (Mfrs)	9 (2)	10 (3)	31 (6)

Table 2. Likert scale for determining AI segmentation performance.

Scale		Description
5	Excellent	No edits are required before clinical use.
4	Good	Minor editing is required (i.e., edits on $<$10% of liver slices).
3	Fair	Major editing is required; edit time $<$ manual contouring.
2	Poor	Major editing is required; edit time $>$ manual contouring.
1	Very Poor	The contour cannot be reasonably edited for clinical use.

patient. Applying the same criteria to our external database from July to August 2025 yielded 384 scans (100 patients), from which one scan per patient was chosen for external validation. Table 1 summarizes demographics and imaging characteristics, demonstrating clear distribution shifts between the validation and external test cohorts. One scan from an external institution was unintentionally included in the internal test dataset; this affected a single case and did not materially impact the reported results. All imaging was retrieved from UT MD Anderson's XNAT platform, which ingests and anonymizes PACS data.

As all scans came from routine clinical workflows, no reference segmentations were available. Three readers assessed model failures: two radiologists (7 and 3 years of experience) and an interventional radiology postdoctoral researcher (3.5 years of experience). For the validation dataset, the radiologist with 7 years of experience labeled segmentations as successes or failures based on whether major edits (edits to $\geq$10% of liver slices) would be required before clinical use. For test datasets, all readers scored segmentations on a five-point Likert scale (Table 2), with scores 1-3 considered failures and 4-5 successes. The final label was determined by majority vote. We reported number of failures, mean Likert ratings, mean pairwise differences between raters, and qualitative observations on failure causes. A two-sided permutation test compared internal and external mean Likert scores (significance level $\alpha = 0.05$).

2.2 Segmentation

The liver CT segmentation model [6, 12] used in this study was deployed via XNAT before the study began. The model uses the 3D full-resolution nnU-Net

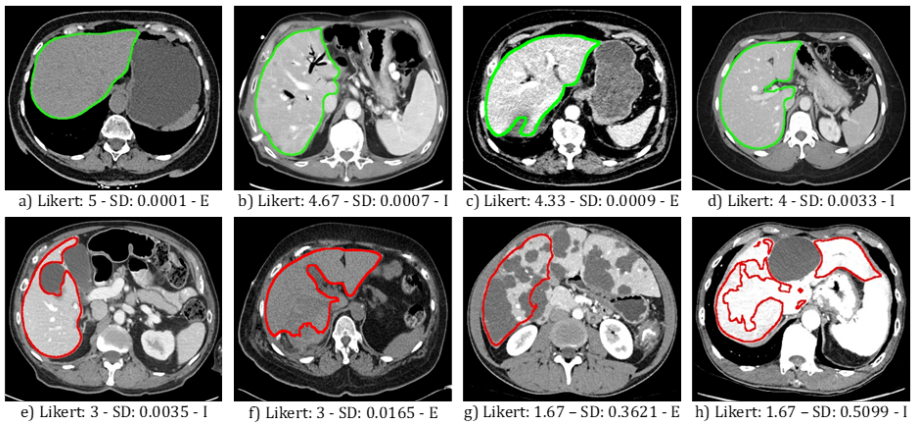

Fig. 1. Segmentations grouped by average Likert scores (L) with the corresponding Pairwise Surface DSC (SD). I - internal image; E - external image. Top row: Likert scores 4–5 (green). Bottom row: Likert scores 1–3 (red).

framework [13] and was trained on 2,840 scans with five-fold cross-validation on five NVIDIA A100 GPUs (1,000 batches per epoch; roughly 4-5 days). During evaluation, it achieved an average Dice similarity coefficient (DSC) of 0.97 across 248 internal and external scans. Inference was performed using cross-validation ensembling, with all ensemble members sharing the same parameters (patch size [48, 244, 192], target spacing [2.5, 0.98, 0.98]) derived from nnU-Net's data fingerprinting and rule-based configuration. Scans were preprocessed using nnU-Net CTNormalization (clip range [-104,191]) and Z-normalization (mean 63.81, SD 53.44), and postprocessing retained only the largest connected component.

2.3 OOD Detection

We evaluated six OOD detection techniques: Maximum Softmax Probability (MSP) [14], Energy Scoring (Energy) [15], Temperature Scaling (Temperature) [16], Deep Ensembling (Ensembling) [17], Pairwise DSC [10], and our proposed Pairwise Surface DSC. For Temperature and Energy, the temperature T was calibrated on the validation dataset over $T \in \{1, 2, 3, 4, 5, 10, 100, 1000\}$, yielding $T=3$ for Temperature and $T=2$ for Energy. For consistency, OOD scores for MSP, Temperature, Pairwise DSC, and Pairwise Surface DSC were inverted (1 - score) so that high values indicated failure. Predictions from the five cross-validation ensemble members were used for Ensembling, Pairwise DSC, and Pairwise Surface DSC. As these predictions were generated during inference, the cost of these methods is limited to lightweight post-processing, making failure detection more practical for clinical deployment than a traditional ensemble.

Pairwise DSC [10] uses the mean DSC across all segmentation pairs for OOD detection. We propose Pairwise Surface DSC, which replaces DSC with surface DSC [18], as surface-based metrics better reflect clinical acceptability [11]. Unlike

Table 3. Performance on internal and external test datasets, with the number of successes (N_S) and failures (N_F) reported. Sensitivity (Sens), specificity (Spec), balanced accuracy (Bal Acc), F1 score, and AUROC are reported with 95% confidence intervals. Evaluated methods are MSP, Energy, Temperature (Temp), Ensembling (Ens), Pairwise DSC (P DSC), and Pairwise Surface DSC (PS DSC) . **Bold** values indicate the best performance for each metric. The † highlights our proposed method.

Internal Test Set ($N_S = 387$, $N_F = 13$)					
Method	**Sens ↑**	**Spec ↑**	**Bal Acc ↑**	**F1 ↑**	**AUROC ↑**
MSP	0.54 (0.25–0.80)	0.80 (0.76–0.84)	0.67 (0.52–0.80)	0.14 (0.05–0.24)	0.64 (0.45–0.80)
Energy	0.31 (0.07–0.58)	0.85 (0.81–0.88)	0.58 (0.45–0.72)	0.11 (0.02–0.20)	0.55 (0.35–0.75)
Temp	0.15 (0.00–0.38)	**0.96** (0.94–0.98)	0.56 (0.48–0.67)	0.14 (0.00–0.32)	0.56 (0.36–0.76)
Ens	0.77 (0.50–1.00)	0.63 (0.58–0.67)	0.70 (0.56–0.81)	0.12 (0.05–0.19)	0.78 (0.59–0.93)
P DSC	0.85 (0.61–1.00)	0.77 (0.72–0.81)	0.81 (0.69–0.89)	0.19 (0.10–0.29)	0.91 (0.83–0.97)
PS DSC†	**1.00** (1.00–1.00)	0.88 (0.85–0.91)	**0.94** (0.92–0.96)	**0.36** (0.22–0.50)	**0.97** (0.94–0.99)
External Test Set ($N_S = 95$, $N_F = 5$)					
Method	**Sens ↑**	**Spec ↑**	**Bal Acc ↑**	**F1 ↑**	**AUROC ↑**
MSP	0.40 (0.00–1.00)	0.52 (0.41–0.62)	0.46 (0.24–0.76)	0.08 (0.00–0.19)	0.60 (0.31–0.95)
Energy	0.00 (0.00–0.00)	0.60 (0.49–0.70)	0.30 (0.25–0.35)	0.00 (0.00–0.00)	0.37 (0.16–0.56)
Temp	0.20 (0.00–0.67)	**0.83** (0.75–0.90)	0.52 (0.38–0.79)	0.09 (0.00–0.29)	0.47 (0.19–0.81)
Ens	**1.00** (1.00–1.00)	0.54 (0.44–0.64)	0.77 (0.72–0.82)	0.19 (0.04–0.33)	0.97 (0.93–1.00)
P DSC	**1.00** (1.00–1.00)	0.76 (0.66–0.84)	**0.88** (0.83–0.92)	0.30 (0.08–0.50)	**1.00** (1.00–1.00)
PS DSC†	**1.00** (1.00–1.00)	0.77 (0.68–0.85)	**0.88** (0.84–0.93)	**0.31** (0.08–0.52)	**1.00** (0.98–1.00)

volumetric DSC, which can remain high despite boundary errors, surface DSC emphasizes boundary agreement by measuring the fraction of surface points within a tolerance (2 mm in our study).

Statistical Analyses. OOD detection performance was measured with specificity, sensitivity, balanced accuracy, F1 score, and the threshold-agnostic area under the receiver operating characteristic curve (AUROC). The utilized binary thresholds were determined by maximizing Youden's J statistic (sensitivity minus false positive rate) [19] on the validation dataset, with failures as the positive class. Nonparametric bootstrapping with 2,000 replicates (400 internal and 100 external scans sampled with replacement) was used to obtain 95% confidence intervals. The significance of performance gains was assessed using McNemar's test on failures for sensitivity and paired, stratified bootstraps for balanced accuracy and AUROC, using Benjamini-Hochberg (BH)-corrected two-sided p-values with significance level $\alpha = 0.05$. Training-based methods (Ensembling, Pairwise DSC, Pairwise Surface DSC) were also compared with output-based methods (MSP, Temperature, Energy) using the same bootstrap tests.

Uncertainty Maps. Interpretability is crucial to OOD, allowing model designers and users to identify image regions linked to failure predictions. MSP and Ensembling offer voxel-wise uncertainty estimates that can provide this interpretability, which we demonstrate by overlaying the estimates on a segmented scan.

3 Results

3.1 Segmentation

The validation, internal test, and external test datasets had 52 (13%), 13 (3%), and 5 (5%) failures, respectively. While the overall failure rates were lower in test data, the rates for the radiologist with seven years of experience remained consistent (10% internal, 12% external). Mean (±SD) Likert ratings of 4.47 (±0.42) and 4.49 (±0.46) for internal and external data demonstrated no significant difference ($p = 0.625$). The mean pairwise difference between raters was 0.6 Likert points, indicating generally small disagreements.

The most common failure mode was underestimating the left hepatic lobe due to liver hypertrophy, with additional failures in cases with multiple cysts or metastases, portal venous embolization, and distended stomachs. Representative examples are shown in Figure 1, including large cysts (subfigures 1e and 1h), adrenal metastases (subfigure 1f), and extensive metastases (subfigure 1g).

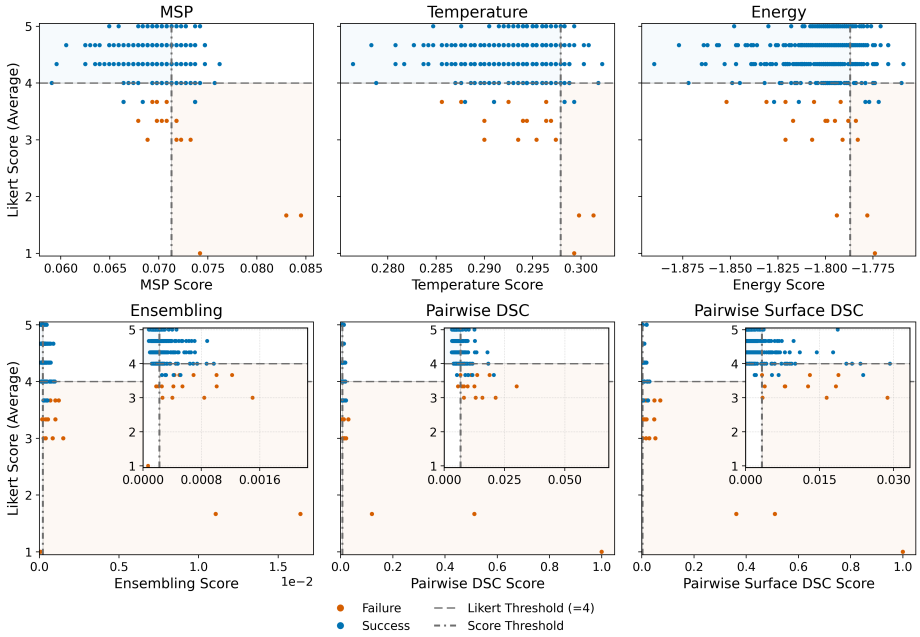

Fig. 2. OOD scores versus Likert ratings. The horizontal line marks the success-failure split used before majority voting. Points represent individual scans, colored by true label, with training-based methods shown with an inset zoom near the score thresholds.

Table 4. BH-adjusted two-sided p-values and power for comparisons of Pairwise Surface DSC (PS DSC) versus each comparator, and for training- vs. output-based method groups, on the combined 500 test scans.

Comparison	**Sensitivity**		**Balanced Accuracy**		**AUROC**	
	p-value	Power	p-value	Power	p-value	Power
PS DSC vs MSP	$\mathbf{p = 0.005}$	1.00	$\mathbf{p < 0.001}$	1.00	$\mathbf{p < 0.001}$	1.00
PS DSC vs Energy	$\mathbf{p < 0.001}$	1.00	$\mathbf{p < 0.001}$	1.00	$\mathbf{p < 0.001}$	1.00
PS DSC vs Temperature	$\mathbf{p < 0.001}$	1.00	$\mathbf{p < 0.001}$	1.00	$\mathbf{p < 0.001}$	1.00
PS DSC vs Ensembling	$p = 0.268$	0.00	$\mathbf{p < 0.001}$	1.00	$\mathbf{p = 0.011}$	0.53
PS DSC vs Pairwise DSC	$p = 0.500$	0.00	$\mathbf{p < 0.001}$	0.75	$\mathbf{p = 0.032}$	0.34
Training vs Output	$\mathbf{p < 0.001}$	0.97	$\mathbf{p < 0.001}$	1.00	$\mathbf{p < 0.001}$	0.99

3.2 OOD Detection

Table 3 summarizes failure-detection performance on the test datasets, while Figure 2 shows the relationship between Likert ratings and detection scores. Pairwise Surface DSC exhibited the best overall performance. Internally, Pairwise Surface DSC achieved perfect sensitivity and the highest balanced accuracy, F1 score, and AUROC. Externally, training-based methods performed similarly, with perfect sensitivity, near-perfect AUROC, and high balanced accuracy.

Statistical Analyses. On the validation dataset, the selected thresholds and their Youden J statistics were: MSP 0.0713 (0.27), Energy -1.7870 (0.14), Temperature 0.2979 (0.13), Ensembling 0.0002 (0.64), P DSC 0.0068 (0.59), and Pairwise Surface DSC 0.0033 (0.64). Table 4 reports the p-values and statistical powers for all comparisons. Pairwise Surface DSC showed significantly higher balanced accuracy and AUROC than all other methods. For sensitivity, Pairwise Surface DSC outperformed the output-based methods but not the training-based ones. Overall, the training-based methods had better performances than the output-based methods across all metrics.

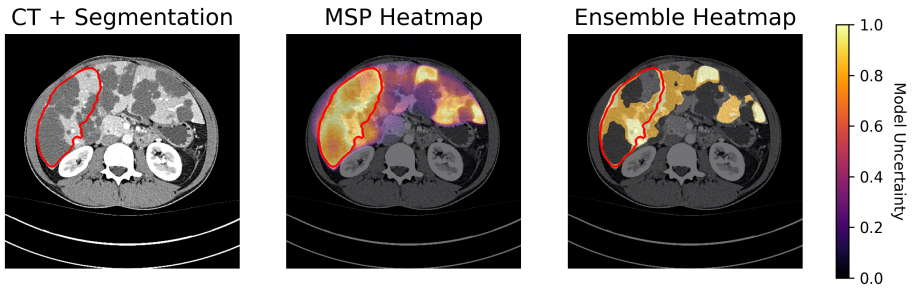

Fig. 3. Spatial uncertainty heatmaps for MSP and Ensembling on the slice of Figure 1g show high uncertainty in yellow and low uncertainty in purple/black, with red contours indicating the predicted segmentation.

Uncertainty Maps. Figure 3 shows voxel-wise uncertainty estimates from MSP and Ensembling for the scan in Figure 1g, which failed due to extensive metastases. Both methods highlight uncertainty in areas missed by the segmentation algorithm.

4 Discussion and Conclusion

Most prior work on OOD detection in medical imaging defines OOD through predefined attributes (e.g., controlled perturbations [20], disease types [21], and domain shifts [22]), rarely applying it to the detection of actual model failures, which is the clinically relevant goal. In contrast, our study assesses failure detection for a deployed liver CT segmentation model with clinical data from 73 sites across 8 different countries. We found training-based approaches to be the top performers, consistent with related literature [7, 17]. The improvement of our proposed Pairwise Surface DSC over Pairwise DSC supports evidence that surface-based metrics more effectively indicate clinical acceptability [11].

Calibrating thresholds and evaluating OOD methods in a clinical context requires access to both failures and successes from the target deployment distribution. Identifying failure cases manually is time-consuming and impractical at scale, a challenge exacerbated by the rarity of failures for well-performing models. Furthermore, when a model is updated, threshold performance may decline, necessitating repeated failure identification to recalibrate thresholds. We hope this work motivates further research into threshold selection and OOD evaluation strategies in low-failure settings.

Several aspects of this work's study design warrant consideration. While expert Likert scoring is more aligned with clinical practice than fixed ground-truth segmentations, it leaves the definition of failure partially subjective. Because the number of failure cases in the test dataset was limited (n=18), statistical power was reported to contextualize each finding. Finally, as this study focuses solely on the liver, future work should evaluate Pairwise Surface DSC on smaller and more complex structures such as tumors.

Our work has multiple practical applications, the most notable being the detection of AI failures in clinical settings, which enhances safety and mitigates automation bias for underrepresented patient groups. Additionally, OOD-based prioritization may aid in model sharing and in enabling large retrospective studies utilizing segmentation by identifying cases needing manual review. Future efforts should adapt these techniques to other organs, modalities, and multiclass scenarios. Overall, our findings indicate that automated failure detection for liver CT segmentation is feasible, and that Pairwise Surface DSC shows promise for deployment.

Acknowledgments. Research in this publication was supported by resources of the Image Guided Cancer Therapy Research Program at the University of Texas MD Anderson Cancer Center, the Tumor Measurement Initiative through the MD Anderson Strategic Initiative Development Program (STRIDE), and the National Cancer Institute of the National Institutes of Health under award numbers P30CA016672,

1R01CA221971, R01CA235564, and P01CA261669. N.S.S. is supported by the National Institutes of Health Image-Guided Cancer Therapies T32 Training Program Fellowship Grant (T32CA261856).

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

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **74**(3), 229–263 (2024). https://doi.org/10.3322/caac.21834
2. Hazhirkarzar, B., Khoshpouri, P., Shaghaghi, M., Ghasabeh, M.A., Pawlik, T.M., Kamel, I.R.: Current state of the art imaging approaches for colorectal liver metastasis. Hepatobiliary Surgery and Nutrition **9**(1) (2019)
3. Shoup, M., Gonen, M., D'Angelica, M., Jarnagin, W.R., DeMatteo, R.P., Schwartz, L.H., Tuorto, S., Blumgart, L.H., Fong, Y.: Volumetric analysis predicts hepatic dysfunction in patients undergoing major liver resection. Journal of Gastrointestinal Surgery **7**(3), 325–330 (2003). https://doi.org/10.1016/S1091-255X(02)00370-0
4. Wesdorp, N.J., Zeeuw, J.M., Postma, S.C.J., Roor, J., van Waesberghe, J.H.T.M., van den Bergh, J.E., Nota, I.M., Moos, S., Kemna, R., Vadakkumpadan, F., Ambrozic, C., van Dieren, S., van Amerongen, M.J., Chapelle, T., Engelbrecht, M.R.W., Gerhards, M.F., Grunhagen, D., van Gulik, T.M., Hermans, J.J., de Jong, K.P., Klaase, J.M., Liem, M.S.L., van Lienden, K.P., Molenaar, I.Q., Kazemier, G.: Deep learning models for automatic tumor segmentation and total tumor volume assessment in patients with colorectal liver metastases. European Radiology Experimental **7**(75) (2023). https://doi.org/10.1186/s41747-023-00383-4
5. Gotra, A., Sivakumaran, L., Chartrand, G., Vu, K., Vandenbroucke-Menu, F., Kauffmann, C., Kadoury, S., Gallix, B., de Guise, J.A., Tang, A.: Liver segmentation: indications, techniques and future directions. Insights into Imaging **8**, 377–392 (2017). https://doi.org/10.1007/s13244-017-0558-1
6. Odisio, B.C., Albuquerque, J., Lin, Y.M., Anderson, B.M., O'Connor, C.S., Rigaud, B., Briones-Dimayuga, M., Jones, A.K., Fellman, B.M., Huang, S.Y., Kuban, J., Metwalli, Z.A., Sheth, R., Habibollahi, P., Patel, M., Shah, K.Y., Cox, V.L., Kang, H.C., Morris, V.K., Kopetz, S., Javle, M.M., Kaseb, A., Tzeng, C.W., Cao, H.T., Newhook, T., Chun, Y.S., Vauthey, J.N., Gupta, S., Paolucci, I., Brock, K.K.: Software-based versus visual assessment of the minimal ablative margin in patients with liver tumours undergoing percutaneous thermal ablation (cover-all): a randomised phase 2 trial. The Lancet Gastroenterology & Hepatology **10**(5), 442–451 (2025). https://doi.org/10.1016/S2468-1253(25)00024-X
7. Woodland, M., Patel, N., Castelo, A., Al Taie, M., Eltaher, M., Yung, J.P., Netherton, T.J., Calderone, T.L., Sanchez, J.I., Cleere, D.W., Elsaiey, A., Gupta, N., Victor, D., Beretta, L., Patel, A.B., Brock, K.K.: Dimensionality reduction and nearest neighbors for improving out-of-distribution detection in medical image segmentation. Machine Learning for Biomedical Imaging **2**, 2006–2052 (2024). https://doi.org/10.59275/j.melba.2024-g93a
8. Anderson, B.M., Lin, E.Y., Cardenas, C.E., Odisio, B.C., Koay, E.J., Brock, K.K.: Automated contouring of contrast and noncontrast computed tomography liver images with fully convolutional networks. Advances in Radiation Oncology **6**(1), 100464 (2021). https://doi.org/10.1016/j.adro.2020.04.023

9. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024). https://doi.org/10.1007/s11263-024-02117-4
10. 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). https://doi.org/10.1016/j.media.2024.103392
11. Baroudi, H., Brock, K.K., Cao, W., Chen, X., Chung, C., Court, L.E., El Basha, M.D., Farhat, M., Gay, S., Gronberg, M.P., Gupta, A.C., Hernandez, S., Huang, K., Jaffray, D.A., Lim, R., Marquez, B., Nealon, K., Netherton, T.J., Nguyen, C.M., Reber, B., Rhee, D.J., Salazar, R.M., Shanker, M.D., Sjogreen, C., Woodland, M., Yang, J., Yu, C., Zhao, Y.: Automated contouring and planning in radiation therapy: What is 'clinically acceptable'? Diagnostics **13**(4) (2023). https://doi.org/10.3390/diagnostics13040667
12. Castelo, A., O'Connor, C., Gupta, A.C., Anderson, B.M., Woodland, M., Altaie, M., Koay, E.J., Odisio, B.C., Tang, T.T., Brock, K.K.: Quality versus quantity of training datasets for artificial intelligence–based whole liver segmentation. medRxiv (2026). https://doi.org/10.64898/2026.02.17.26346486, `https://www.medrxiv.org/content/early/2026/02/18/2026.02.17.26346486`
13. 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). https://doi.org/10.1038/s41592-020-01008-z
14. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
15. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 21464–21475. Curran Associates, Inc. (2020)
16. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (06–11 Aug 2017)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
18. Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S.A., Garie, B., McQuinlan, Y., Ireland, S., Hampton, K., Fuller, K., Montgomery, H., Rees, G., Suleyman, M., Back, T., Hughes, C.O., Ledsam, J.R., Ronneberger, O.: Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. J Med Internet Res **23**(7), e26151 (Jul 2021). https://doi.org/10.2196/26151
19. Fluss, R., Faraggi, D., Reiser, B.: Estimation of the youden index and its associated cutoff point. Biometrical Journal **47**(4), 458–472 (2005). https://doi.org/10.1002/bimj.200410135

20. Huijben, E.M., Amirrajab, S., Pluim, J.P.: Enhancing reconstruction-based out-of-distribution detection in brain mri with model and metric ensembles. Computer Methods and Programs in Biomedicine **272**, 109045 (2025). https://doi.org/10.1016/j.cmpb.2025.109045
21. Linmans, J., van der Laak, J., Litjens, G.: Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In: Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C. (eds.) Proceedings of the Third Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 121, pp. 465–478. PMLR (06–08 Jul 2020)
22. Zhang, O., Delbrouck, J.B., Rubin, D.L.: Out of distribution detection for medical images. In: Sudre, C.H., Licandro, R., Baumgartner, C., Melbourne, A., Dalca, A., Hutter, J., Tanno, R., Abaci Turk, E., Van Leemput, K., Torrents Barrena, J., Wells, W.M., Macgowan, C. (eds.) Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. pp. 102–111. Springer International Publishing, Cham (2021)