Beyond Boundary Noise: Aggregated Aleatoric Uncertainty Fails to Capture Presence Ambiguity in 3D Lung Nodule Segmentation

Simon Baur[1], Arne Schernich[1], Ekin Böke[1], Wojciech Samek[1,2,3], Jackie Ma[1]

[1] Fraunhofer Heinrich-Hertz-Institut, 10587 Berlin, Germany
{simon.baur, wojciech.samek, jackie.ma}@hhi.fraunhofer.de
[2] Technische Universität Berlin, 10623 Berlin, Germany
[3] The Berlin Institute for the Foundations of Learning and Data (BIFOLD), 10587 Berlin, Germany

Abstract. Uncertainty estimation is critical for the safe clinical deployment of deep learning in medical image segmentation, with aleatoric uncertainty theoretically designed to capture irreducible data ambiguity. However, whether entropy-based measures reflect clinically meaningful ambiguity, i.e. case-level disagreement about whether a pathology is present at all, remains poorly understood. Contrary to most prior work, which focused on pixel-wise boundary disagreement, we systematically evaluate how well aleatoric uncertainty captures presence ambiguity. Our evaluation spans 3D lung nodule segmentation across four architectures with Monte Carlo dropout and deep ensembles, on LIDC-IDRI and an external validation cohort (LNDb). We find that entropy-based uncertainty maps align with boundary noise and minor drawing variation but carry insufficient discriminative signal for presence ambiguity. In contrast, a lightweight supervised ambiguity head trained on frozen segmentation features substantially outperforms all entropy-aggregation-based baselines across architectures, metrics, and both cohorts, and matches or exceeds methods that explicitly model ambiguity under disagreement supervision (Probabilistic U-Net, Annotator-Confusion 3D-UNet). A qualitative feature-space analysis shows that presence ambiguity is already encoded in the frozen encoder features of pixel-wise-trained networks, only to be discarded by the segmentation output and its entropy aggregation. Our findings expose a fundamental mismatch between the theoretical promise of aleatoric uncertainty and its practical behavior, and suggest that practitioners should not rely on entropy-based uncertainty as a proxy for clinical ambiguity in safety-critical applications.

Keywords: Uncertainty Quantification · Aleatoric Uncertainty · Lung Nodule Segmentation · Annotator Disagreement · Ambiguity Estimation

1 Introduction

Machine-assisted medical decision making increasingly supports clinical workflows such as diagnosis, treatment planning, and disease monitoring (21), where erroneous or overconfident predictions can directly affect patient outcomes (13). Models must therefore provide not only accurate predictions but also reliable uncertainty estimates (3; 19), making uncertainty quantification (UQ) a fundamental requirement for safe clinical deployment (14), with medical image segmentation as a prominent testbed for high-risk applications (7; 12). Most existing work obtains segmentation uncertainty by sampling an approximate predictive distribution at inference, e.g., via Monte Carlo dropout or deep ensembles (5; 11), then processing these samples with measures such as Shannon entropy or voxel-wise variance to produce spatial (aleatoric) uncertainty maps (7). In practice, such maps are consumed by visual inspection: a clinician reading them performs an implicit case-level aggregation, dismissing isolated highlighted voxels and attending instead to regions where uncertainty clusters into salient shapes, with a conspicuous cluster prompting closer re-inspection, follow-up, or a second reading. This pooled reading, rather than any single voxel, is the operative signal at the point of care: to inform such decisions, an uncertainty map must flag the cases that experts themselves would flag. Yet uncertainty maps of pixel-to-pixel trained segmentation models are dominated by low-level labeling variability at object boundaries (Fig.1), driven by partial volume effects or minor inter-annotator inconsistencies (25; 22), and thus tend to emphasize disagreement over boundary regions (9; 26). However, segmentation tasks with potentially absent pathological findings involve a semantically distinct and more clinically meaningful ambiguity, which is case-level inter-annotator disagreement about whether a pathology is present at all. For the reader of an uncertainty map, this failure mode is silent: the map remains visually plausible, highlighting boundary regions as expected, while a case in which experts disagree about the very existence of a finding can appear less uncertain than an unambiguous one (Fig.1). Aleatoric uncertainty maps can thus offer false reassurance precisely where caution is warranted. Despite its relevance, this presence ambiguity remains comparatively understudied, a gap we address in this work. Our contributions can be summarized as follows:

- We systematically evaluate whether entropy-based aleatoric uncertainty scores capture *existential* ambiguity in lung nodule segmentation (case-level disagreement about the *presence* of a segmentation target) and show that they cannot: entropy-based maps reliably highlight object boundaries (low-level pixel-wise annotator noise) but fail to capture whether a pathology is present at all, regardless of whether uncertainty is aggregated over the ground-truth or the predicted mask, revealing a fundamental mismatch between standard UQ methods and clinically relevant ambiguity.
- We show that a lightweight supervised post-hoc ambiguity head, trained on top of frozen segmentation features, recovers case-level annotator disagreement far better than entropy aggregation of MC-dropout and ensemble

predictions, and matches or exceeds methods that explicitly model ambiguity under disagreement supervision (Probabilistic 3D-UNet (10), Annotator-Confusion 3D-UNet (20)), consistently across architectures on both in-distribution (LIDC) (1) and out-of-distribution (LNDb) (18) data.

– We offer an explanation of this phenomenon: a qualitative feature-space analysis (UMAP) (15) shows that networks trained purely on pixel-wise supervision already encode presence ambiguity in their internal representations.

2 Related Work

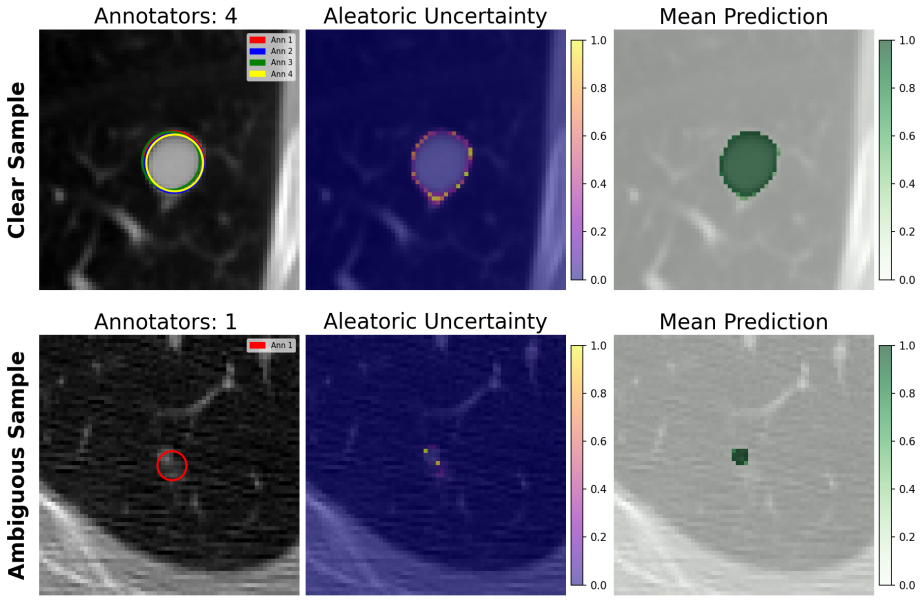

Fig. 1. Visualization of a non-ambiguous (top row) and ambiguous (bottom row) lung nodule patch from the LIDC-IDRI dataset. Each column shows: the CT image with annotator segmentations overlaid as colored circles (left), the aleatoric uncertainty map derived from MC-Dropout (center), and the mean predicted segmentation mask (right). We can clearly see that the ambiguous sample (where annotators disagree on whether a structure is present at all) produces virtually no aleatoric uncertainty signal.

Bayesian Uncertainty Estimation and Disentanglement
Within the supervised learning framework, a probabilistic learner aims to construct a predictive distribution $\hat{p}(\cdot \mid x)$ over outcomes $y \in \mathcal{Y}$ given inputs $x \in \mathcal{X}$. In a Bayesian setting, this corresponds to the posterior predictive distribution, marginalizing over model parameters θ (17; 23). Predictive uncertainty (PU) is

quantified as the Shannon entropy of this distribution:

$$\mathrm{PU}(y \mid x) = H\big(\mathbb{E}_{p(\theta|\mathcal{D})}[p(y \mid x, \theta)]\big)\,, \tag{1}$$

and is commonly decomposed into aleatoric and epistemic components (8):

$$\mathrm{PU}(y \mid x) \;=\; \underbrace{\mathbb{E}_{p(\theta|\mathcal{D})}[H(p(y \mid x, \theta))]}_{\mathrm{AU}(y|x)} \;+\; \underbrace{\mathcal{I}(y, \theta \mid x)}_{\mathrm{EU}(y|x)}. \tag{2}$$

Aleatoric uncertainty (AU) captures irreducible input-dependent ambiguity, such as measurement noise or ambiguous anatomical boundaries, while epistemic uncertainty (EU) reflects reducible lack of knowledge (4). In practice, AU is approximated via the expected entropy of Monte Carlo dropout samples or deep ensembles (5; 11), and its separation from EU (disentanglement) has been studied in general benchmarks (16) and medical image classification (2). However, this decomposition has been called into question: the use of Shannon entropy, conditional entropy, and mutual information as uncertainty proxies, and the assumed additivity of the AU–EU split itself, have been empirically challenged (23; 16; 2), casting doubt on whether it yields the semantically meaningful separation it is commonly assumed to provide.

Aleatoric Uncertainty From Spatial Inter-Annotator Disagreement
A line of work grounds aleatoric uncertainty directly in inter-annotator disagreement, building on probabilistic segmentation models that augment a U-Net with a conditional latent space to sample globally consistent segmentation hypotheses (10). Unlike independent pixel-wise fluctuations, these latent samples correspond to coherent segmentation variants, naturally capturing spatial annotation ambiguity. Subsequent work calibrates this uncertainty by supervising the model against empirical per-pixel label frequencies from multiple annotators via soft-label cross-entropy (20; 6; 27). While an important step beyond unsupervised sampling-based UQ, these methods target voxel-level spatial disagreement rather than the case-level existential ambiguity of whether a pathology is present at all, the setting we address here.

3 Method and Experimental Setup

We conduct all experiments on the LIDC-IDRI (LIDC) dataset (1) (70/15/15 train/val/test split), a large-scale thoracic CT collection with a two-stage annotation protocol in which radiologists first labeled independently and then revised after reviewing their peers' assessments, so that remaining disagreements reflect genuine clinical ambiguity about nodule presence rather than incidental labeling noise. For generalization, we additionally evaluate on the LNDb dataset (18) as an external validation cohort of completely unseen data, with no retraining or supervision.

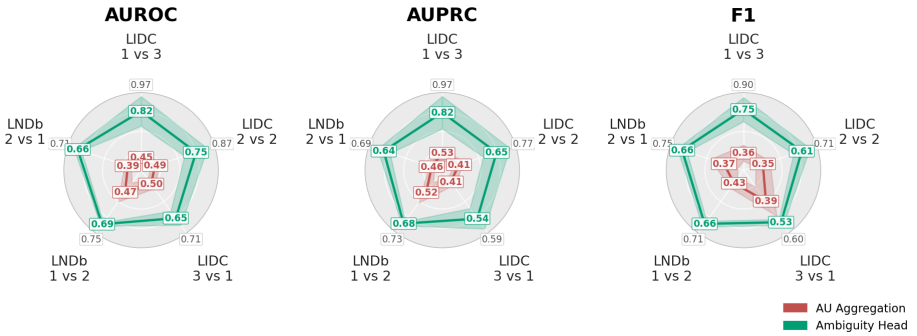

Fig. 2. Presence ambiguity detection across disagreement levels on LIDC and LNDb. Radar plots compare pixel-wise AU aggregation (mean of Ensemble and MC-Dropout) against post-hoc Ambiguity Head, averaged over four architectures. Entropy-based aggregation stays near-random across all disagreement levels, while the lightweight head recovers a highly discriminative signal from frozen segmentation features.

Case-Level Ambiguity Formulation The LIDC and LNDb datasets provide up to K independent radiologist annotations per scan ($K = 4$ and $K = 3$, respectively). Each annotator supplies a binary segmentation mask $m_k \in \{0,1\}^{H\times W\times D}$ indicating the presence of a lung nodule. We define a binary presence indicator for each annotator:

$$\mathbf{1}_k(x) = \begin{cases} 1 & \text{if } |m_k| > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $|m_k|$ denotes the number of foreground voxels in mask m_k. The empirical presence agreement for a case x is then:

$$A(x) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_k(x) \in [0,1]. \tag{4}$$

We define *non-ambiguous cases* as those with unanimous agreement on presence or absence, $A(x) \in \{0,1\}$, and *ambiguous cases* as those in which annotators disagree, $0 < A(x) < 1$. Ambiguity detection is then formulated as a binary classification problem:

$$y_{\text{amb}}(x) = \begin{cases} 1 & \text{if } 0 < A(x) < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The resulting label distinguishes global presence ambiguity from pixel-wise boundary disagreement, covering different partial-disagreement configurations (1-vs-3, 2-vs-2, and 3-vs-1 for LIDC; 1-vs-2 and 2-vs-1 for LNDb).

Case-Level Uncertainty Aggregation Since ambiguity is defined at the case level, voxel-wise uncertainty maps must be aggregated into a scalar score. As

reference region for the aggregation we use the union of the annotator masks, $M(x) = \{v_i : \max_k m_k(v_i) = 1\}$, i.e., a voxel belongs to $M(x)$ if at least one annotator labeled it as foreground. $M(x)$ is therefore well-defined under annotator disagreement and non-empty for every evaluated case, as each case contains at least one non-empty annotation. We provide results for aggregating over the ground truth mask as well as the predicted mask. When aggregating over the predicted mask instead, $\hat{M}(x) = \{v_i : \hat{y}_{v_i} = 1\}$, the prediction may be empty (roughly 5% of cases); these cases are excluded from the predicted-mask evaluation. We aggregate aleatoric uncertainty by computing the mean of $\mathrm{AU}(v_i)$ over all voxels within the ground-truth mask:

$$U_{\mathrm{AU}}(x) = \frac{1}{|M(x)|} \sum_{v_i \in M(x)} \mathrm{AU}(v_i). \tag{6}$$

In our experiments we evaluate how well the aggregated uncertainty score $U_{\mathrm{AU}}(x)$ aligns with the case-level ambiguity label $y_{\mathrm{amb}}(x)$.

Post-hoc Ambiguity Head Our post-hoc ambiguity head trained on frozen segmentation features is defined as follows: let $f_\theta(x)$ denote the feature representation extracted at the architecture bottleneck (the most compressed latent representation, prior to any decoder upsampling) of the frozen, pretrained segmentation backbone. Operating on this representation, we attach a trainable head g_ϕ consisting of attention pooling followed by a 3-layer MLP with ReLU activations and a sigmoid output. We optimize:

$$\phi^* = \arg\min_{\phi} \; \mathcal{L}_{\mathrm{BCE}}(g_\phi(f_\theta(x)),\, y_{\mathrm{amb}}(x))\,, \tag{7}$$

where $\mathcal{L}_{\mathrm{BCE}}$ denotes binary cross-entropy loss. The resulting scalar output $U_{\mathrm{amb}}(x) = g_{\phi^*}(f_\theta(x))$ serves as ambiguity estimate akin to $U_{\mathrm{AU}}(x)$.

Experimental Details Details regarding experimental setting, training and evaluation beyond what is stated above are provided in appendix A.1.

4 Results

The radar plots in Figure 2 compare entropy aggregation (mean of MC-dropout and ensemble, which behave near-identically) against the ambiguity head, averaged over all four architectures and broken down by disagreement configuration. Across both cohorts, entropy aggregation stays close to random at every disagreement level, whereas the ambiguity head is strongly discriminative throughout. Figure 3 aggregates over models and disagreement levels, adds the supervised competitors, and reports every baseline for both aggregation regions: the ambiguity head reaches 0.74 AUROC, 0.67 AUPRC, and 0.63 F1 on LIDC, against

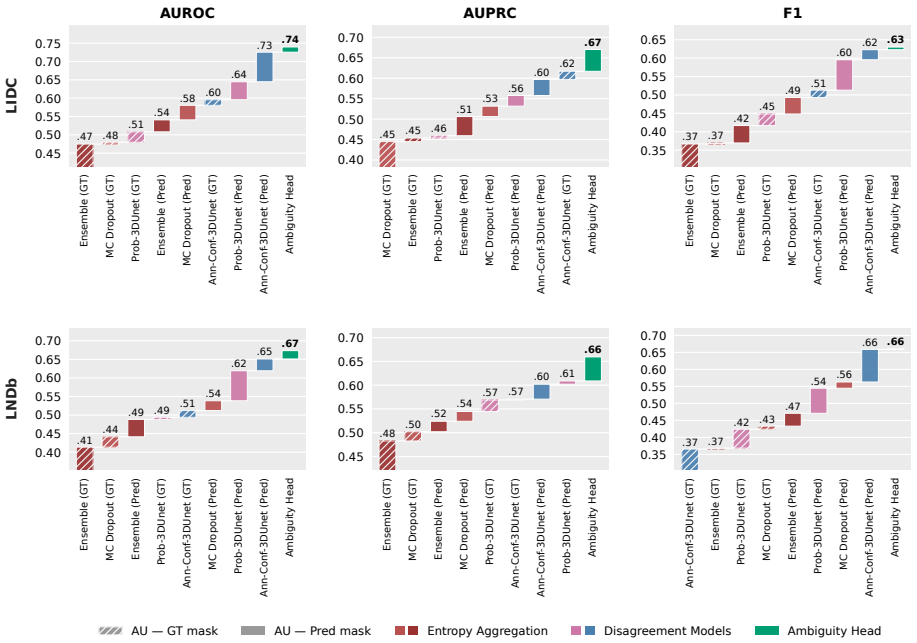

Fig. 3. Sorted performance metrics, averaged across model architectures and ambiguity thresholds. Aleatoric uncertainty (AU) baselines appear twice, differing in the region over which voxel-wise uncertainty is aggregated into a score: the ground-truth mask (hatched) or the predicted mask (solid). Using the predicted mask consistently improves both the unsupervised entropy-aggregation methods and the supervised disagreement-modelling baselines (Prob-3DUnet, Ann-Conf-3DUnet). The Ambiguity Head surpasses all baseline variants across metrics in almost all cases, except for Ann-Conf-3DUnet (prediction mask aggregation) where it performs on par. Notably, this performance also holds in the external validation cohort (LNDb). Standard deviations are omitted here, as averaging over disagreement levels, models, and runs renders them uninformative. Non-aggregated metrics and standard deviations per disagreement level and model are reported in Appendix A.2.

at most 0.48 AUROC for entropy aggregation over the ground-truth mask. Aggregating over the predicted mask instead consistently improves all baselines: entropy aggregation rises to at most 0.58 AUROC, remaining far below the ambiguity head, while the methods that explicitly model ambiguity under disagreement supervision benefit most, with the Annotator-Confusion 3D-UNet improving from 0.60 to 0.73 AUROC and the Probabilistic 3D-UNet from 0.51 to 0.64. The ambiguity head thus remains the strongest method across all metrics, though the margin over the best supervised competitor narrows under predicted-mask aggregation, down to on-par performance on F1 (0.63 vs. 0.62). The same ordering holds on LNDb: 0.67 AUROC for the ambiguity head against at most 0.54 for entropy aggregation and 0.65 for the Annotator-Confusion 3D-UNet, with

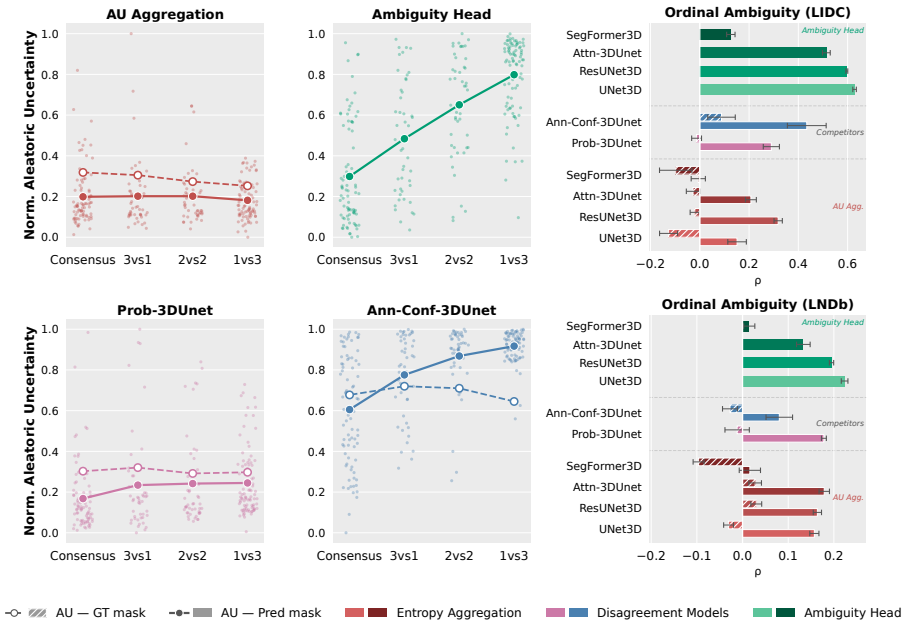

Fig. 4. Ordinal correlation between uncertainty scores and ambiguity severity. Scatter panels (left, mid) show an exemplary distribution of uncertainty scores (UNet-3D). The summary (right) reports Spearman correlations (ρ) across both cohorts, all backbones, and all methods. The Ambiguity Head stratifies the full spectrum of disagreement severity rather than only binary ambiguity, whereas pixel-wise AU aggregation and the supervised baselines track severity only weakly or inconsistently.

both reaching 0.66 F1. Beyond binary detection, Figure 4 measures whether each score ranks cases by disagreement severity via the Spearman correlation with the ordinal agreement tier. The ambiguity head increases monotonically across tiers and attains the strongest correlation across both cohorts and all backbones. Entropy-based AU aggregation over the ground-truth mask is near-zero and flat; switching to the predicted mask recovers a moderate correlation (up to $\rho = 0.43$ for the Annotator-Confusion 3D-UNet on LIDC), which nevertheless stays well below the ambiguity head (0.63), while on LNDb all correlations are compressed and the ambiguity head retains a smaller lead (0.22 vs. at most 0.18). The ambiguity head's discriminative power scales modestly with segmentation quality, though this is confounded with backbone family, as the three UNet variants yield more informative heads than the transformer-based SegFormer3D, hinting that the optimal feature-extraction point may be architecture-dependent. All backbones reach segmentation performance in line with the literature (24; 27): 0.80–0.84 Dice for the UNet variants, 0.78 for SegFormer3D, and 0.70 for Annotator-Confusion 3D-UNet. Per-architecture and per-disagreement level metrics for both aggregation regions are reported in Ap-

pendix A.2. Finally, the UMAP projection in Figure 5 offers insight into why the ambiguity head succeeds: consensus and ambiguous cases form clearly separable clusters across architectures, indicating that presence ambiguity is already encoded in the encoder representations of pixel-wise trained networks.

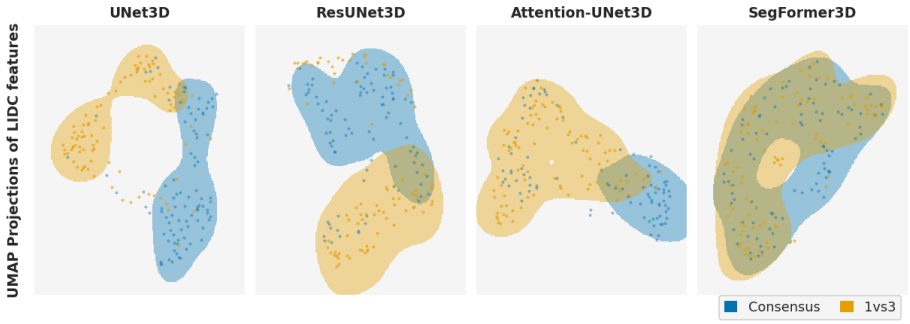

Fig. 5. UMAP projection of frozen model features (1vs3). The clustering indicates presence ambiguity is already well-represented in the latent space, i.e. Ambiguity Head does not learn from scratch but recovers a signal the standard pipeline fails to transport.

5 Conclusion

Across four segmentation architectures and two datasets (one an external validation cohort), we showed that entropy-based aleatoric uncertainty is ill-suited for detecting case-level presence ambiguity for lung nodule segmentation: under standard training, spatial entropy maps capture boundary noise and minor drawing variation but carry near-zero signal for whether a pathology is present at all. This gap between the theoretical promise of aleatoric uncertainty and its practical behavior is critical for clinical deployment, where misleading estimates can directly influence diagnostic decisions, and documenting it is the central contribution of this work. The ambiguity head should be read in this light: not as a proposed competitor that improves ambiguity detection over baseline methods, but as a deliberately supervision-advantaged upper bound that makes the failure quantifiable. It demonstrates that presence ambiguity is readily recoverable from the frozen features of pixel-wise trained networks (i.e. the signal being present, but discarded by entropy aggregation on the pixel-wise segmentation output) and that even methods trained with explicit disagreement supervision recover it only partially. We therefore recommend against using entropy-based uncertainty as a proxy for clinical ambiguity when supervision ground truths are available. As this work mainly aims to highlight the underexplored problem of presence ambiguity, the ambiguity head is only a simple proof of concept; future work should explore richer ways to incorporate presence ambiguity during training

and extend the analysis to other structures and modalities, even though case-level ambiguity with multiple annotator labels remains rare in public datasets, itself a limitation the community should address.

Acknowledgments. This work was supported by the Senate of Berlin and the European Commision's Digital Europe Programme (DIGITAL) as grant TEF-Health (101100700).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)

[2] Baur, S., Samek, W., Ma, J.: Benchmarking uncertainty and its disentanglement in multi-label chest x-ray classification. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 193–203. Springer (2025)

[3] Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence **1**(1), 20–23 (2019)

[4] Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In: International conference on machine learning. pp. 1184–1193. PMLR (2018)

[5] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)

[6] Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 137–145. Springer (2019)

[7] 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 **97**, 103223 (2024)

[8] Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning **110**(3), 457–506 (2021)

[9] Islam, M., Glocker, B.: Spatially varying label smoothing: Capturing uncertainty from expert annotations. In: international conference on information processing in medical imaging. pp. 677–688. Springer (2021)

[10] Kohl, S.A.A., Romera-Paredes, B., Meyer, C., Fauw, J.D., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.M.A., Rezende, D.J., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. ArXiv **abs/1806.05034** (2018)

[11] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)

[12] Li, S., Yuan, M., Dai, X., Zhang, C.: Evaluation of uncertainty estimation methods in medical image segmentation: Exploring the usage of uncertainty in clinical deployment. Computerized Medical Imaging and Graphics **124**, 102574 (2025)

[13] Li, Y., Yi, X., Fu, J., Yang, Y., Duan, C., Wang, J.: Reducing misdiagnosis in ai-driven medical diagnostics: A multidimensional framework for technical, ethical, and policy solutions. Frontiers in Medicine **12**, 1594450 (2025)

[14] López, L., Elsharief, S., Jorf, D.A., Darwish, F., Ma, C., Shamout, F.E.: Uncertainty quantification for machine learning in healthcare: a survey. arXiv preprint arXiv:2505.02874 (2025)

[15] McInnes, L., Healy, J.: Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv **abs/1802.03426** (2018), https://api.semanticscholar.org/CorpusID:3641284

[16] Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. Advances in neural information processing systems **37**, 50972–51038 (2024)

[17] Murphy, K.P.: Probabilistic machine learning: an introduction. MIT press (2022)

[18] Pedrosa, J., Aresta, G., Ferreira, C., Rodrigues, M., Leitão, P., Carvalho, A.S., Rebelo, J., Negrão, E., Ramos, I., Cunha, A., et al.: Lndb: a lung nodule database on computed tomography. arXiv preprint arXiv:1911.08434 (2019)

[19] Seoni, S., Jahmunah, V., Salvi, M., Barua, P.D., Molinari, F., Acharya, U.R.: Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). Computers in Biology and Medicine **165**, 107441 (2023)

[20] Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11236–11245 (2019)

[21] Thomas, K.S., Edpuganti, S., Puthooran, D.M., Thomas, A., Joy, A., Latheef, S.: Artificial intelligence in modern clinical practice. Medicine International **6**(1), 5 (2025)

[22] Tohka, J.: Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review. World journal of radiology **6 11**, 855–64 (2014)

[23] Wimmer, L., Sale, Y., Hofman, P., Bischl, B., Hüllermeier, E.: Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: Uncertainty in artificial intelligence. pp. 2282–2292. PMLR (2023)

[24] Xi, Y., Xu, C., Ye, F., Yuan, M., Ye, C., Jiang, L., Huang, Y., Zhang, J., Liu, M., Liu, X., et al.: Coreformer high fidelity pulmonary nodule segmentation with structural core priors and geodesic implicit fields. npj Digital Medicine (2025)

[25] Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., Antani, S.K.: Assessing inter-annotator agreement for medical image segmentation. IEEE Access **11**, 21300–21312 (2023)

[26] Ye, A., Chen, Q.Z., Zhang, A.: Confidence contours: Uncertainty-aware annotation for medical semantic segmentation. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 11, pp. 186–197 (2023)

[27] Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.: Disentangling human error from ground truth in segmentation of medical images. Advances in Neural Information Processing Systems **33**, 15750–15762 (2020)