# LEXU: Learning from Expert Disagreement for Single-Pass Uncertainty Estimation in Medical Image Segmentation

Kudaibergen Abutalip[1], Numan Saeed[1], Fadillah Maani[1], Ikboljon Sobirov[1], Vincent Andrearczyk[2], Adrien Depeursinge[2], and Mohammad Yaqub[1]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
kudaibergen.abutalip@mbzuai.ac.ae
[2] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

**Abstract.** Deploying deep learning (DL) models in medical applications relies on predictive performance and other critical factors, such as conveying trustworthy predictive uncertainty. Uncertainty estimation (UE) methods provide potential solutions for evaluating prediction reliability and improving the model confidence calibration. This paper introduces **L**earning from **EX**pert Disagreement for **U**E (LEXU) for medical image segmentation, a method that leverages the variability in annotations from multiple experts to guide model training. By focusing on regions of disagreement among experts and incorporating multi-rater optimization strategy, LEXU enhances the model's awareness of challenging cases, resulting in better calibration and predictive uncertainty. The method shows a 55% improvement in correlation with expert disagreements at the image level and a 23% improvement at the pixel level, along with competitive segmentation performance compared to state-of-the-art techniques, all while requiring only a single forward pass.

**Keywords:** Uncertainty estimation · Medical image segmentation · Model calibration.

## 1 Introduction

Maximizing the predictive performance of deep learning (DL) models is not the only factor leading to a wide-scale deployment in real-world applications. Particularly in the medical domain, other model properties must be analyzed to ensure clinical adoption and minimize unforeseen consequences. Experts underline the inability of models to convey trustworthy predictive uncertainty as one of the main reasons for their slow and limited adoption in clinical practice [6]. Uncertainty estimation (UE) is gaining attention as a promising solution for evaluating prediction reliability as well as for purposes such as enhancing prediction quality, conducting quality assurance, domain adaptation, and active learning [9,32].

Various UE methods have been proposed to enhance the reliability and safety of predictive systems. For instance, Stochastic Variational Inference [7] estimates

Table 1: Inter-observer variability Dice scores, shown as mean (std). Object 1 and Object 2 refer to {disc, cup} for RIGA and {GTVp, GTVn} for HECKTOR.

| Dataset | RIGA | | HECKTOR | | |
|---|---|---|---|---|---|
| Split/Fold | Val | Test | Fold 1 | Fold 2 | Fold 3 |
| Object 1 | 0.954 (0.012) | 0.956 (0.013) | 0.773 (0.243) | 0.722 (0.250) | 0.835 (0.209) |
| Object 2 | 0.789 (0.094) | 0.796 (0.114) | 0.723 (0.251) | 0.705 (0.263) | 0.720 (0.231) |

the posterior distribution by modeling a Gaussian distribution for each parameter of the network. Monte-Carlo Dropout (MCDO) [8] aggregates outputs of multiple forward passes of the same input with activated dropout layers to approximate the true posterior of the model. Deep Ensembles (DE) [20] consists of multiple networks trained with different initializations. Test-time data augmentation (TTA) [4] uses multiple forward passes but with differently augmented versions of the same input. Studies [24,3] highlight that DE outperforms most methods in robustness and confidence calibration despite their time and memory inefficiency. Although a growing number of studies on UE indicate a promising trajectory, some questions remain unanswered, e.g. the calibration of uncertainty estimates, uncertainty vs fairness, and practical deployment of UE methods.

Generally, low model uncertainty suggests the prediction is accurate. However, there are cases where the model shows low uncertainty, yet multiple experts disagree, indicating the need for a closer examination of the given case. This mismatch between low model uncertainty and high expert disagreement can lead to critical oversights, such as missing early diagnoses of diseases. The disagreement between multiple annotators is measured by the inter- and intra-observer variability analysis, whereby experts often have different opinions and levels of expertise when assigning labels [30,14,18]. To illustrate, Table 1 presents the inter-observer agreement scores for the retinal fundus images used in glaucoma analysis (RIGA) [1] and for the head and neck tumor segmentation (HECKTOR) [2] datasets employed in this study. Although multi-rater label sampling and training strategies exist [13,23,21,10,15,22], the direction of explicitly incorporating this natural uncertainty information into the training process to obtain better calibrated and more reliable model outputs remains underexplored.

As models are trained to mimic human annotators for disease detection, there is a need to align the uncertainty estimates with real-world divergences in expert judgment for different scenarios. If we focus on trust and transparency, such an alignment can foster more effective collaboration between clinicians and DL models. When models express uncertainty in situations that mirror human uncertainty, clinicians are more inclined to trust the model's predictions. Moreover, enhancing the ability to handle scenarios deviating from training data promotes robustness and adaptability.

Additionally, the development of new UE methods should emphasize simplicity and efficiency to ensure widespread adoption and accessibility. Most current UE methods require several input passes (e.g., MCDO, TTA) or considerably

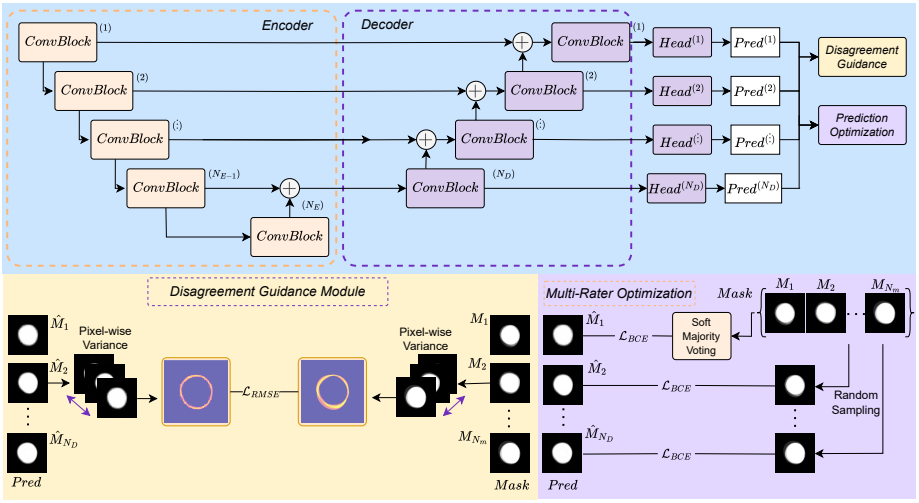



Fig. 1: LEXU employs a U-Net-like architecture. The *Disagreement Guidance Module* (DGM) captures uncertainty by comparing variance heatmaps from the model and annotations. *Multi-Rater Optimization* (MRO) enables segmentation training using multiple expert-provided annotations.

increase the number of parameters, which incurs additional time and financial and environmental costs.

To address these issues, we propose LEXU (**L**earning from **EX**pert Disagreement for **U**E), an expert disagreement-guided uncertainty estimation method for medical image segmentation. We explicitly use variability in ground-truth annotations from several raters to guide the model and develop a multi-rater optimization strategy to incorporate ground-truth segmentation masks from all annotators during training. We validate our results on two distinct ophthalmology and head and neck tumor datasets, which contain images from different modalities, including fundus photography, CT, and PET scans. We show that LEXU produces well-calibrated uncertainty outputs that correlate better with expert disagreements than existing state-of-the-art methods. The method does not explicitly distinguish between aleatoric and epistemic uncertainties, as recent studies [25,17,29] indicate that reliably separating them is often infeasible in practice. Instead, it addresses the challenge of identifying difficult samples where multiple annotators may disagree, which is inherently tied to both types of uncertainty. The main contributions of this study are:

- We develop a novel, simple and intuitive UE method that takes into account **inherent variability in ground-truth masks** by leveraging multiple-annotator datasets.
- We demonstrate an efficient **single forward pass** method to estimate both image and pixel-level uncertainties.

– We offer insights into relevant downstream applications and conduct a comprehensive analysis of method components, addressing recommendations from prior studies [17]. The code is publicly available on github.com.[3]

## 2 Methodology

We propose a method that utilizes variability in annotations from multiple experts to enhance predictive UE, model calibration, and awareness of challenging cases. We build upon Layer Ensembles (LE) [19], a single-pass uncertainty estimation framework for medical image segmentation, which extends the concept of prediction depth (PD) [5]. PD assesses sample complexity and segmentation quality by incorporating multiple segmentation heads at different depths of the network. LEXU leverages this idea by linking PD, segmentation heads, and ground truth variability from multi-annotator datasets.

### 2.1 Problem Formulation

Consider a medical image segmentation dataset $\mathcal{D} = \{X^i, \mathbf{M}^i\}_{i=1}^{N}$ consisting of images $X \in \mathbb{R}^{H \times W \times C}$ and corresponding masks $\mathbf{M} = \{M_j\}_{j=1}^{N_m}$ from $N_m$ annotators. Our primary goal is to develop an efficient uncertainty-aware segmentation network $\mathcal{F}$ that enhances predictive uncertainty estimation while maintaining low computational cost. Unlike DE [20] which demands high computational and memory costs, $\mathcal{F}$ requires a single forward pass to generate UE through multiple ($N_p$) segmentation predictions, i.e. $\mathcal{F}(X) = \hat{\mathbf{M}} = \{\hat{M}_j\}_{j=1}^{N_p}$, which are used to generate an uncertainty map. As shown in Figure 1, our model $\mathcal{F}$ consists of an encoder-decoder architecture and multiple segmentation heads.

**Encoder-Decoder.** We adopt the U-Net [28] for our encoder-decoder design. An input image $X$ passes through the encoder with $N_E$ downsampling blocks for extracting hierarchical features. Feature maps after each block are used to feed into subsequent blocks and skip connections from encoder to decoder during upsampling. Each decoder block consists of upsampling, convolutional layers, batch normalization, ReLU, and skip connections.

**Segmentation Heads.** We attach a segmentation head to every decoder output for generating multiple segmentation predictions, resulting in $Np \equiv N_D$. Each segmentation head consists of a convolution layer for transforming its input to the desired segmentation output channels, followed by upsampling to ensure a uniform shape across all predictions. This approach facilitates UE while introducing only a negligible increase in computational overhead and parameters.

### 2.2 Learning from Expert Disagreement for Uncertainty Estimation

LEXU proposes to distill uncertainty information available from multi-rater ground-truth, enabling $\mathcal{F}$ to estimate natural uncertainty provided by experts.

[3] https://github.com/Katalip/grader_soup

As illustrated in Figure 1, LEXU incorporates the *Disagreement Guidance Module* (DGM) to explicitly capture annotation variability, and *Multi-Rater Optimization (MRO)* to enable learning from different expert annotations.

**DGM.** Given a set of masks $\{M_j\}_{j=1}^{N_m}$ for an image $X$, we first compute the pixel-wise variance along the channel axis to obtain the ground-truth variance heatmap $H$. Similarly, we apply the same procedure to generate the model uncertainty heatmap $\hat{H}$ from segmentation predictions $\mathcal{F}(X)$. Then, we apply the RMSE loss between $H$ and $\hat{H}$ to allow the model to learn the inherent uncertainty in the ground-truth masks. In this way, based on the previously defined PD concept [5], each segmentation head can mimic a certain level of expertise, e.g., low-, mid-, and high-level details. Using heatmaps from ground-truth masks, these heads can imitate a high level of disagreement at the pixel level when the sample has many ambiguous regions and, on the other hand, have smaller uncertainty when the image is relatively easier to segment.

**MRO**. Inspired by [12,16], we design MRO to facilitate segmentation training by leveraging annotations from multiple experts to mitigate overconfident predictions and improve model calibration. Let $\hat{M}_1$ be the last segmentation head prediction. We optimize $\hat{M}_1$ to harness all available annotations through a soft majority voting label, denoted as $\mathcal{S}(\mathbf{M})$. For the rest predictions $\{M_j\}_{j=2}^{N_m}$, we **r**andomly **s**ample one of the annotations, denoted as RS($\mathbf{M}$).

**Final Loss Function**. Overall, LEXU's loss can be defined as follows:

$$\mathcal{L} = \alpha \cdot \left( \mathcal{L}_{BCE}(\mathcal{S}(\mathbf{M}), \hat{M}_1) + \sum_{j=2}^{N_p} \mathcal{L}_{BCE}(\mathrm{RS}(\mathbf{M}), \hat{M}_j) \right) + \beta \cdot \mathcal{L}_{RMSE}(H, \hat{H}), \quad (1)$$

where $\mathcal{L}_{BCE}$ is the segmentation loss, and $\mathcal{L}_{RMSE}$ measures the discrepancy between the ground-truth and predicted variance heatmaps. $\alpha, \beta \in \mathbb{R}$.

## 3 Experimental Details

**Datasets. RIGA** benchmark [1] is a public dataset for retinal cup and disc segmentation, containing 750 color fundus images from three databases: 460 from MESSIDOR, 195 from BinRushed, and 95 from Magrabia. Six glaucoma experts manually labeled the segmentation masks. Following [23,13], we used BinRushed and MESSIDOR for training, while Magrabia was reserved for testing. All images are normalized between 0 and 1 and resized to 256×256.

A subset of **HECKTOR 2022** data [2] consists of 44 cases with multiple annotations, each with registered 3D CT and PET scans of head and neck (H&N) region. Annotations of gross tumor volumes of the primary tumors (GTVp) and lymph nodes (GTVn) come from 10 different experts. On average, 3 annotations are available for an image (ranges from 2 to 4 per patient). We perform preprocessing steps similar to [26] and convert 3D volumes to 2D axial slices.

**Comparison and Evaluation Metrics.** Motivated by [19], we conduct a comprehensive evaluation of correlation analysis and segmentation performance. We measure image-level uncertainty correlations using Spearman's rank correlation

Table 2: Correlation analysis and segmentation performance comparison on the RIGA and HECKTOR datasets.

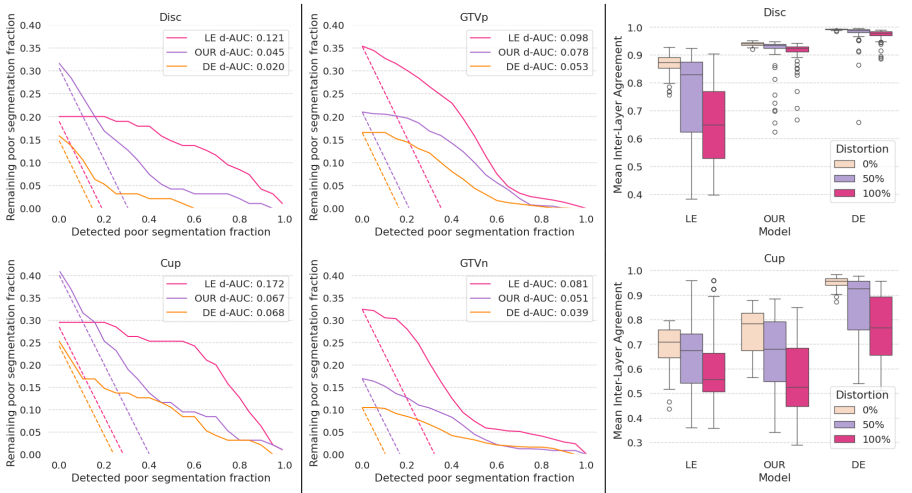
| | | Correlation Analysis | | | | | | Segmentation Scores | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SR↑ | SR↑ | DC↑ | DC↑ | NCC↑ | NCC↑ | Dice↑ | Dice↑ | NLL↓ |
| RIGA | | Disc | Cup | Disc | Cup | Disc | Cup | Disc | Cup | |
| | DE | 0.472 | 0.349 | 0.444 | 0.399 | 0.644 | 0.607 | **0.977** | **0.874** | 0.219 |
| | LE | -0.630 | -0.517 | 0.601 | 0.516 | 0.470 | 0.473 | 0.974 | 0.869 | 0.174 |
| | LEXU | **0.703** | **0.689** | **0.685** | **0.651** | **0.766** | **0.735** | 0.970 | 0.856 | **0.163** |
| HECKTOR | | GTVp | GTVn | GTVp | GTVn | GTVp | GTVn | GTVp | GTVn | |
| | DE | 0.184 | 0.158 | 0.349 | 0.245 | 0.314 | 0.348 | **0.829** | **0.784** | 0.512 |
| | LE | 0.810 | 0.620 | 0.766 | 0.575 | 0.242 | 0.245 | 0.766 | 0.691 | 0.197 |
| | LEXU | **0.904** | **0.816** | **0.862** | **0.799** | **0.456** | **0.492** | 0.788 | 0.726 | **0.131** |

Table 3: Effect of heatmap loss with different $\beta$ values on the RIGA dataset.

| | Correlation Analysis | | | | Segmentation Scores | | |
|---|---|---|---|---|---|---|---|
| | SR↑ | SR↑ | DC↑ | DC↑ | Dice↑ | Dice↑ | NLL↓ |
| | Disc | Cup | Disc | Cup | Disc | Cup | |
| $\beta = 0$ | -0.594 | -0.524 | 0.572 | 0.526 | 0.974 | 0.874 | 0.173 |
| $\beta = 3$ | 0.690 | 0.707 | 0.665 | 0.657 | 0.973 | 0.854 | 0.166 |
| $\beta = 5$ | 0.712 | 0.673 | 0.680 | 0.643 | 0.971 | 0.855 | 0.163 |
| $\beta = 10$ | 0.661 | 0.606 | 0.680 | 0.634 | 0.965 | 0.818 | 0.161 |

(SR) and distance correlation (DC) [31] between variance sums in heatmaps from models and ground-truth masks. For pixel-level assessment, we compute the average normalized cross-correlation (NCC) between heatmap pairs for the test set. The Negative Log-Likelihood metric (NLL) evaluates network confidence calibration, penalizing small uncertainty for incorrect predictions. Segmentation performance is assessed using the soft Dice metric with soft majority voting labels as ground-truth masks. For the RIGA dataset, we report the results of three runs for LEXU and LE and a single run for DE. With HECKTOR, we report patient-based 3-fold cross-validation results.

**Implementation Details.** We use an ImageNet-pretrained ResNet50 [11] as the encoder, resulting in $N_E = 6$ and $N_D = 5$. The models are trained for 200 epochs with a *bs* of 16 on RIGA and 120 epochs with a *bs* of 32 on HECKTOR; chosen empirically. We use a learning rate of $5e^{-5}$. We employ five networks in DE, and skip the first five segmentation heads for LE as suggested by [19]. For LEXU, we set $\alpha = 1$, while $\beta$ is set to 5 for RIGA and 2.5 for HECKTOR.




(a) RIGA (b) HECKTOR (c) RIGA

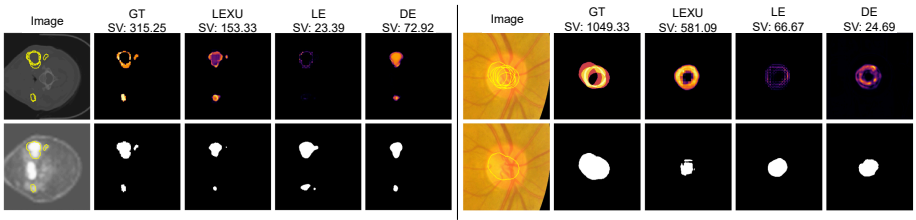
Fig. 2: **(a)**–**(b)** Segmentation quality control results. d-AUC (↓): the difference between the area under the main curve and the ideal line. Dashed lines indicate ideal lines. **(c)** Example difficulty estimation results, showing the distribution of layer agreements and model agreements at 0%, 50%, and 100% distorted images.

## 4 Results and Discussion

### 4.1 Correlation Analysis and Segmentation Performance

Table 2 summarizes the correlation and segmentation performance. In the RIGA dataset, DE shows a moderate correlation with the SR value reaching 0.472 and 0.349 for the disc and cup, respectively, whereas LE indicates a negative correlation. LEXU shows a stronger correlation with SR of 0.703±0.02 and 0.689±0.012 for the disc and cup respectively. This trend is reflected in the DC scores as well. While DE and LE models show some correlation, LEXU shows much higher DC values for both the disc and cup, with 0.685±0.020 and 0.651±0.008. At the pixel level, NCC values for LEXU are also the highest, with 0.766±0.005 and 0.735±0.003 for disc and cup, respectively. Although LEXU has a slight drop in segmentation performance compared to the best-performing DE model, it requires 5x fewer parameters. In addition, LEXU has the lowest NLL value of 0.163±0.005, while DE and LE models tend to be overconfident in their predictions. This suggests LEXU is less prone to overconfidence and making predictions at somewhat ambiguous regions in the image. We confirm this behavior in the qualitative examples (Figures 3) and by examining the $\beta$ coefficient (Table 3).

The HECKTOR dataset shows a similar trend in performance, both in correlation and segmentation analyses. The DE model captures a low level of correlation in the GTVp and GTVn in SR and DC metrics. While LE shows reasonable scores, LEXU achieves the highest correlation scores at both image and pixel




(a) HECKTOR (GTVn) (b) RIGA (Cup)

Fig. 3: **Top row:** Input image with contours of all masks, ground-truth variance heatmap, variance heatmaps from LEXU, LE, DE. **Bottom row:** Input image with soft majority voting mask's contour (threshold 0.5), corresponding ground-truth mask, predicted masks from LEXU, LE, DE. SV: sum of variances.

levels. DE has the highest dice scores, however LEXU shows much better calibration, achieving the lowest NLL.

### 4.2 Segmentation Quality Control

We evaluate methods for segmentation quality control [27] on both datasets. We choose dice thresholds of 0.97 and 0.85 for disc and cup, and 0.65 and 0.55 for GTVp and GTVn, respectively, to mark model outputs as poor segmentation masks. The variance metric is used for all methods to detect these marked masks. Figure 2a shows the results for disc and cup structures, while the results for GTVp and GTVn are available in Figure 2b. DE shows the best overall performance, with the lowest remaining poor segmentation fraction at all quantile thresholds for both cup and GTVn. LEXU performs similarly to DE, especially for the cup. However, it has a slightly higher remaining poor segmentation fraction than DE for GTVn at higher variance thresholds. LE has the worst performance overall, with the highest remaining poor segmentation fractions. Overall, the results show that LEXU is effective for segmentation quality control, performing better than LE and similarly to DE.

### 4.3 Example Difficulty Estimation

We also evaluate the methods for example difficulty estimation (Figure 2c) by applying Gaussian noise, blurring, hue, saturation, and value changes on the RIGA dataset. We use inter-layer agreement scores for LEXU, LE, and agreement scores between models for DE to detect perturbed images. Overall, the agreement scores decrease for all methods as the distortion level (proportion of images with augmentations) increases. This indicates that all methods are effective at detecting challenging samples as the applied distortions become more severe, which is more apparent with the cup as it has a more complex structure.

## 5 Conclusion

We proposed LEXU, a simple approach that leverages expert disagreements to guide the model during training for improved predictive uncertainty quantification without additional costs. Our extensive evaluations of the RIGA and HECKTOR datasets have shown that LEXU is effective at detecting difficult cases in which multiple annotators may disagree, and also maintains robust segmentation performance. Future work could refine DGM by exploring alternative loss functions, such as KL divergence, and expand LEXU to other medical imaging tasks. Overall, the alignment of model uncertainty with expert variability is a significant step towards fostering trust and transparency in clinical settings, which is crucial for the adoption of deep learning models in medical practice.

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

## References


1. Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., Lakshminarayanan, V.: Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. Int. Ophthalmol. **37**(3), 701–717 (Jun 2017)
2. Andrearczyk, V., Oreiller, V., Boughdad, S., Le Rest, C.C., Tankyevych, O., Elhalawani, H., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., Hatt, M., Depeursinge, A.: Automatic head and neck tumor segmentation and outcome prediction relying on fdg-pet/ct images: Findings from the second edition of the hecktor challenge. Medical Image Analysis **90**, 102972 (2023)
3. Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.: Pitfalls of in-domain uncertainty estimation and ensembling in deep learning (2021), `https://arxiv.org/abs/2002.06470`
4. Ayhan, M.S., Berens, P.: Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks (2018), `https://api.semanticscholar.org/CorpusID:13998356`
5. Baldock, R.J.N., Maennel, H., Neyshabur, B.: Deep learning through the lens of example difficulty. Advances in Neural Information Processing Systems **34** (2021)
6. Banerji, C.R., Chakraborti, T., Harbron, C., MacArthur, B.D.: Clinical ai tools must convey predictive uncertainty for each individual patient. Nature medicine **29**(12), 2996–2998 (2023)
7. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning (2015)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd ICML. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016)
9. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(1), 1513–1589 (Oct 2023)
10. Gros, C., Lemay, A., Cohen-Adad, J.: Softseg: Advantages of soft versus binary training for image segmentation. Medical image analysis **71**, 102038 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on CVPR (2016)
12. Jensen, M.H., Jørgensen, D.R., Jalaboi, R., Hansen, M.E., Olsen, M.A.: Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: MICCAI. pp. 540–548. Springer (2019)
13. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning calibrated medical image segmentation via multi-rater agreement modeling. Proceedings of the IEEE/CVF Conference on CVPR **3**(1), 12341–12351 (2021)
14. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Inter-observer variability of manual contour delineation of structures in ct. European Radiology **29**(3), 1391–1399 (2019)
15. Joskowicz, L., Cohen, D., Caplan, N., Sosna, J.: Automatic segmentation variability estimation with segmentation priors. Medical Image Analysis **50**, 54–64 (2018)
16. Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: MICCAI. pp. 682–690. Springer (2018)

17. Kahl, K.C., Lüth, C.T., Zenk, M., Maier-Hein, K., Jaeger, P.F.: Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. arXiv preprint arXiv:2401.08501 (2024)
18. Kumar, S., Giubilato, A., Morgan, W., Jitskaia, L., Barry, C., Bulsara, M., Constable, I.J., Yogesan, K.: Glaucoma screening: analysis of conventional and telemedicine-friendly devices. Clinical & Experimental Ophthalmology **35**(3), 237–243 (2007)
19. Kushibar, K., Campello, V., Garrucho, L., Linardos, A., Radeva, P., Lekadir, K.: Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation. In: MICCAI. pp. 514–524. Springer (2022)
20. 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)
21. Lemay, A., Gros, C., Karthik, E., Cohen-Adad, J.: Label fusion and training methods for reliable representation of inter-rater uncertainty. Machine Learning for Biomedical Imaging **1**, 1–27 (01 2023)
22. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation. Medical image analysis **92** (2021)
23. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Transformer-based annotation bias-aware medical image segmentation. In: MICCAI. Springer (2023)
24. Mehrtens, H.A., Kurz, A., Bucher, T.C., Brinker, T.J.: Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. Medical image analysis **89**, 102914 (2023)
25. Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in NeurIPS. vol. 37, pp. 50972–51038. Curran Associates, Inc. (2024)
26. Myronenko, A., Siddiquee, M.M.R., Yang, D., He, Y., Xu, D.: Automated head and neck tumor segmentation from 3d pet/ct (2022)
27. Ng, M., Guo, F., Biswas, L., Petersen, S.E., Piechnik, S.K., Neubauer, S., Wright, G.: Estimating uncertainty in neural networks for cardiac mri segmentation: A benchmark study. IEEE Transactions on Biomedical Engineering **69**(1), 1–23 (2022)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI. pp. 234–241. Springer International Publishing, Cham (2015)
29. Roshanzamir, P., Rivaz, H., Ahn, J., Mirza, H., Naghdi, N., Anstruther, M., Batti'e, M.C., Fortin, M., Xiao, Y.: How inter-rater variability relates to aleatoric and epistemic uncertainty: a case study with deep learning-based paraspinal muscle segmentation. In: UNSURE@MICCAI (2023)
30. Schaekermann, M., Beaton, G., Habib, M., Lim, A., Larson, K., Law, E.: Understanding expert disagreement in medical data analysis through structured adjudication. Proceedings of the ACM on Human-Computer Interaction **3**(1), 1–23 (2019)
31. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. The Annals of Statistics **35**(6), 2769 – 2794 (2007)
32. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. Meta-Radiology **1**(1), 100003 (2023)