ACKNOWLEDGEMENTS

This work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. This work is supported by the Helmholtz Association Initiative and Networking Fund under the Helmholtz AI platform grant (ALEGRA (ZT-I-PF-5-121)).

REFERENCES

Samuel G. Armato III, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J. R. van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.-Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. 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. ISSN 2473-4209. doi: 10.1118/1.3528204.

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 119–127. Springer, 2019.

Pascal Colling, Lutz Roese-Koerner, Hanno Gottschalk, and Matthias Rottmann. Metabox+: A new region based active learning method for semantic segmentation using priority maps. *arXiv preprint arXiv:2010.01884*, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pp. 715–726. Springer, 2021.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Camila Gonzalez, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. Detecting When Pre-trained nnU-Net Models Fail Silently for Covid-19 Lung Lesion Segmentation. In *Medical Image Computing and Computer Assisted Intervention MICCAI 2021*, pp. 304–314, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87234-2. doi: 10.1007/978-3-030-87234-2_29.

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 318–319, 2020.

Matthew C. Hancock and Jerry F. Magnan. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: Probing the Lung Image Database Consortium dataset with two statistical learning methods. *Journal of Medical Imaging*, 3(4):044504, 2016. ISSN 2329-4302. doi: 10.1117/1.JMI.3.4.044504.

Christopher J. Holder and Muhammad Shafique. Efficient Uncertainty Estimation in Semantic Segmentation via Distillation. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3080–3087. IEEE, 2021. ISBN 978-1-66540-191-3. doi: 10.1109/ICCVW54120.2021.00343.

Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 137–145. Springer, 2019.

Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175 – 193, 1906. doi: 10.1007/BF02418571.

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in neural information processing systems*, 30, 2017.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Benjamin Lambert, Florence Forbes, Senan Doyle, Alan Tucholka, and Michel Dojat. Improving Uncertainty-based Out-of-Distribution Detection for Medical Image Segmentation. *arXiv preprint arXiv:2211.05421*, 2022.

Carsten Tim Lüth, Till J. Bungert, Lukas Klein, and Paul F Jaeger. Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment. In *Thirty-Seventh Conference on Neural Information Processing Systems*, volume 36, 2023.

Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. CEREALS - Cost-Effective REgion-based Active Learning for Semantic Segmentation. *British Machine Vision Conference 2018 (BMVC)*, 2018.

Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.

Raghav Mehta, Angelos Filos, Yarin Gal, and Tal Arbel. Uncertainty Evaluation Metric for Brain Tumour Segmentation. *arXiv preprint arXiv:2005.14262*, 2020.

Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with Illusions about Deep Active Learning. *arXiv preprint arXiv:1912.05361*, 2019.

Sudhanshu Mittal, J. Niemeijer, J. Schäfer, and Thomas Brox. Best practices in active learning for semantic segmentation. In *German Conference on Pattern Recognition (GCPR)*, 2023.

Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu, and Hien Van Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):1–14, 2021.

Miguel Monteiro, Loic Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12756–12767. Curran Associates, Inc., 2020.

Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

Jishnu Mukhoti, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*, 2021.

Will Nash, Liang Zheng, and Nick Birbilis. Deep learning corrosion detection with confidence. *npj Materials degradation*, 6(1):26, 2022. ISSN 2397-2106. doi: 10.1038/s41529-022-00232-6.

Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection. 2018.

Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2931–2940, 2019.

Janis Postels, Mattia Segu, Tao Sun, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. *arXiv preprint arXiv:2107.00649*, 2021.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pp. 102–118. Springer International Publishing, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer International Publishing, 2015. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Luke Whitbread and Mark Jenkinson. Uncertainty Categories in Medical Image Segmentation: A Study of Source-Related Diversity. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 26–35. Springer, 2022.

Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Ge Zhang, Hao Dang, and Yulong Xu. Epistemic and aleatoric uncertainties reduction with rotation variation for medical image segmentation with ConvNets. *SN Applied Sciences*, 4(2):56, February 2022. ISSN 2523-3963, 2523-3971. doi: 10.1007/s42452-022-04936-x.

A Downstream Tasks & Metrics

A.1 Segmentation Performance Assessment

Dice To measure the segmentation performance of the segmentation backbone and prediction models, we used the Dice score which is defined as:

$$\text{Dice}(\hat{y}, y^*) = \frac{2|y^* \cap \hat{y}|}{|y^*| + |\hat{y}|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2}$$

As we have multiple segmentation predictions $\hat{y}$ for most prediction models and multiple reference segmentations y^*, we decided to take the average Dice between each of the N reference segmentations and the mean prediction $\bar{y}$:

$$\text{Dice} = \frac{1}{N} \sum_{i=1}^{N} \text{Dice}(\bar{y}, y_i^*) \tag{3}$$

A.2 Out of Distribution Detection

In our evaluation, we concentrate on image-level OoD-D to facilitate human assessment, as humans typically evaluate complete images rather than individual pixels. Notably, if any part of an image is identified as OoD, it has the potential to impact all predictions, rendering them unreliable.

Area Under the Receiver Operating Characteristics Curve (AUROC) We calculate the Area Under the Receiver Operating Characteristics Curve (AUROC) to determine a method's capability of detecting OoD cases. Thus, we assign each image a label with 1 equal to an image being OoD and 0 equal to an image being i.i.d. We then use the sklearn library for determining the ROC curve[1] with the ground truth input (0 or 1) and the uncertainty scores as target values. We then determine the AUC also using sklearn[2].

A.3 Failure Detection

In our evaluation, we concentrate on image-level FD to facilitate human assessment, as humans typically evaluate complete images rather than individual pixels. To this end, we make use of our performance assessment metric on image-level (Dice) to define a continuous failure label for our utilized FD metrics.

Our motivation behind this approach is that FD based on the Dice is more informative with regard to the performance of the model than on the pixel level, as deciding whether a single image should be assessed by a human in place of an automated decision-making process requires to have an assessment of the quality of the segmentation for an entire image than for single pixels.

We compute our FD metrics twice: first using the i.i.d. test data and then the OoD test data. This approach enables us to assess how effectively failures are detected within the i.i.d. data and, subsequently, how well the uncertainty method detects failures when exposed to OoD data.

Area under the Risk-Coverage-Curve (AURC) The Area under the Risk-Coverage-Curve (AURC) is a metric used in selective classification. The goal hereby is to successfully identify failures by having a low *risk*, i.e. a good classifier performance but also achieve high *coverage*, i.e. select as few cases as possible for manual correction. For calculating the Area under the Risk-Coverage-Curve, we use the implementation following Jaeger et al. (2023). To adapt it for a semantic segmentation predictor f and evaluation dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, we define the *confidence scoring function* (CSF) $g(x_i)$ as the negative uncertainty score. Furthermore, we choose the inverted

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html

Dice score as the risk l associated with a prediction:

$$l(x, y, f) = 1 - \text{Dice}(f(x), y) \tag{4}$$

The risk-coverage curve is obtained by introducing a confidence threshold τ, which leads to the selective risk

$$\text{Risk}(\tau | f, g, D) = \frac{\sum_{i=1}^{N} l(x_i, y_i, f) \cdot \mathbb{I}(g(x_i) \geq \tau)}{\sum_{i=1}^{N} \mathbb{I}(g(x_i) \geq \tau)} \tag{5}$$

and coverage, defined in Jaeger et al. (2023) as the ratio of cases remaining after selection:

$$\text{Coverage}(\tau | g, D) = \frac{\sum_{i=1}^{N} \mathbb{I}(g(x_i) \geq \tau)}{N} \tag{6}$$

The AURC based on a threshold list $\{\tau\}_{t=1}^{T}$ with T values of a CSF that are sorted ascending can then be calculated as Jaeger et al. (2023):

$$\text{AURC}(f, g, D) = \sum_{t=1}^{T} (\text{Coverage}(\tau_t) - \text{Coverage}(\tau_{(t-1)})) \cdot (\text{Risk}(\tau_t) + \text{Risk}(\tau_{t-1}))/2 \tag{7}$$

where we omitted the conditioning on f, g, D on the RHS for clarity.

Excess-AURC (E-AURC) Further, as also analyzed in Jaeger et al. (2023) and originally proposed in Geifman et al. (2019), we use the *excess AURC* (E-AURC) as an evaluation metric that is independent of the segmentation model's performance:

$$\text{E-AURC} = \text{AURC}(f, g, D) - \text{AURC}(f, g^*, D) \tag{8}$$

where the second term corresponds to the optimal AURC. The optimal CSF g^* can be formally obtained, for example, by using an oracle CSF that returns a confidence equal to the negative risk of a particular prediction, $g^*(x) = -l(x, y, f)$. Practically, it ranks the predictions perfectly by their risk (in our case ascending Dice score). Although we are aware of the fact that evaluating a CSF without considering the performance of the model itself harms the meaningful comparison of uncertainty methods (see Jaeger et al. (2023)), we use this as an additional debugging metric which is feasible in our case as the there are no significant outliers in terms of segmentation performance as seen in the Dice score of Table 5.

A.4 Active Learning

In our evaluation, we concentrate on AL performing queries on image-level to facilitate human assessment, as humans typically evaluate complete images rather than individual pixels. The general concept here is that we have a model that is already performing well on an i.i.d. dataset with saturated performance for a specific task which should be adapted to a shifted (OoD) dataset with the same task. Therefore we only measure the performance increase on the OoD test set.

Active Learning Improvement (AL improvement) To assess the AL improvement of the uncertainty methods, we measure the relative performance change between two cycles t_1 and t_2 on the OoD test set:

$$C = \frac{\text{Dice}_{t_2} - \text{Dice}_{t_1}}{\text{Dice}_{t_1}} \tag{9}$$

As we do not want to consider effects of random querying in our evaluation, we subtract the performance change that is reached with random querying from the performance change of the uncertainty method, leading to the following final performance change:

$$C_{\text{final}} = C_{\text{method}} - C_{\text{random}} \tag{10}$$

A.5 Calibration

Our evaluation of the CALIB follows standard protocol is performed with pixel-level ground truth and aggregated to single images requiring therefore no aggregation.

We compute our CALIB metrics twice: first using the i.i.d. test data and then the OoD test data. This approach enables us to assess how well the uncertainty measure is calibrated on i.i.d. data and, subsequently, how well the uncertainty measure is calibrated when exposed to inputs from OoD data.

Average Calibration Error (ACE) The Average Calibration Error (ACE) is introduced in Neumann et al. (2018) and used for segmentation in Jungo et al. (2020). In contrast to the Expected Calibration Error (ECE), which is used e.g. in Gustafsson et al. (2020); Jungo et al. (2020), every bin in the calibration histogram is weighted equally, leading to the following formulation:

$$\text{ACE} = \frac{1}{M} \sum_{m}^{M} |c_m - \text{Acc}_m| \tag{11}$$

Here, M is the number of non-empty bins, c_m is the average confidence in bin m, and Acc_m the respective average accuracy. We apply Platt scaling in order to get confidence scores between 0 and 1. We chose this metric in comparison to the ECE as it avoids overweighting the background pixels which are predominant in our case.

A.6 Ambiguity Modeling

Our evaluation of AM is separated into two main parts: first, whether an uncertainty measure can successfully indicate AU in the correct regions, and second, whether a prediction model is able to produce multiple realistic predictions.

The evaluation is performed using pixel-level ground truth based on single images requiring therefore no aggregation.

We compute our AM metrics twice: first using only the i.i.d. test data and on the OoD test data. This approach enables us to assess how good the uncertainty measures model AU on i.i.d. data and, subsequently, how good the uncertainty measures model AU on the OoD data.

Normalized Cross-Correlation (NCC) We calculate the normalized cross-correlation (NCC) following Hu et al. (2019):

$$\frac{1}{n_p \sigma_a \sigma_b} \sum_{i=1}^{n_p} (a_i - \mu_a) \cdot (b_i - \mu_b) \tag{12}$$

Here, a is the reference uncertainty map, b is the predicted uncertainty map, n_p is the total number of pixels in the uncertainty maps, and μ and σ are mean and standard deviation of the uncertainty maps. The reference uncertainty map is calculated with the pixel variance of a pixel y_i for N different segmentation raters $\{y_i^1, ..., y_i^N\}$:

$$\mathbb{V}_{p(D)}[y_i] = \frac{1}{N} \sum_{j=1}^{N} (y_i^j - \bar{y}_i)^2 \tag{13}$$

where $\bar{y}_i$ is the mean over the segmentation raters $\bar{y}_i = \frac{1}{N} \sum_{j=1}^{N} y_i^j$.

Generalized Energy Distance (GED) To better assess the capability of the uncertainty methods to model multiple raters, we use the generalized energy distance (GED), which has been used in various other works focusing on AM (Kohl et al. (2018); Monteiro et al. (2020); Hu et al. (2019)):

$$D^2_{\text{GED}}(p, \hat{p}) = 2\mathbb{E}_{y\sim p, \hat{y}\sim\hat{p}}[d(y, \hat{y})] - \mathbb{E}_{y,y'\sim p}[d(y, y')] - \mathbb{E}_{\hat{y},\hat{y}'\sim\hat{p}}[d(\hat{y}, \hat{y}')] \tag{14}$$

Here, $d(y, y')$ is the distance between two reference segmentations, and $d(\hat{y}, \hat{y}')$ is the distance between two predicted segmentation variants. p and $\hat{p}$ are the respective reference and predicted distributions for the segmentations masks. The distance has to satisfy that it increases for more dissimilar masks and further $d(x, y) = 0$ for $x = y$. As we use the Dice as our main evaluation metric, we chose to use $d(x, y) = 1 - \text{Dice}(x, y)$ as distance.

B DATASETS

B.1 TOY DATASET SETUP

B.1.1 DATASET SCENARIOS

As mentioned in Sec. 4.2, we generate three different training and four different testing scenarios for the toy dataset. An overview of the different scenarios, including the number of training and testing cases in each scenario, is shown in Table 1. Each of those settings is targeted at answering a specific question in our separation study, described in Sec. 4.1. In setting 1, we induce AU, thus aiming to answer Q1 and Q2 of the separation study. Setting 2 focuses on EU, and thus aims to answer Q3 and Q4. However, since AU is not induced in setting 2, we hypothesize that the behavior of AU-measures should not be well-predictable, limiting the ability to clearly answer Q4. Therefore, we design setting 3, and provide testing scenarios (a) and (b) where AU is induced during training and (b) where AU is also present in the i.i.d testing data. These aim at understanding the behavior of our uncertainty measures to detect EU with varying degrees of AU present.

Table 1: Number of training and testing cases for the toy dataset. For each scenario, the number of training cases and the number of testing cases is specified. Further, the number of cases with ambiguity / blur is specified in brackets and the number of i.i.d and OoD cases in the testset.

Scenario	Description	# Train (# blur)	# Test	
			# i.i.d (# blur)	# OoD
1	Training models on data with induced AU; testing on i.i.d. data also containing AU	200 (200)	20	
			20 (20)	0
2	Training models on data without ambiguity; testing on i.i.d. data and shifted data	200 (0)	42	
			21 (0)	21
3a	Training models on data with and without blur/ambiguity; testing on i.i.d. data and shifted data without blur	200 (100)	42	
			21 (0)	21
3b	Training models on data with and without blur/ambiguity; testing on i.i.d. data and shifted data without blur and i.i.d data with blur	200 (100)	63	
			42 (21)	21

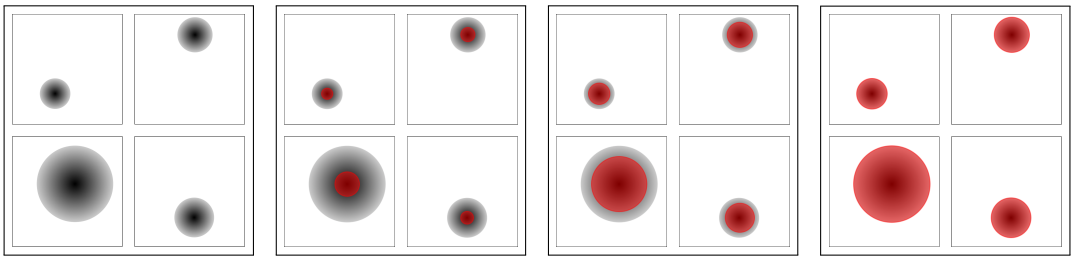
B.1.2 DATA WITH INDUCED ALEATORIC UNCERTAINTY

Figure 4 shows the data scenario that is created with induced aleatoric uncertainty. The input (Figure 4a) shows a sphere that has Gaussian blur to the outside. Due to the blur to the outside, the exact border of the sphere is ambiguous. This ambiguity is modeled by three different reference raters (Figure 4b - Figure 4d). Thereby the segmentation of rater 1 (Figure 4b), that segments the smallest sphere, is 10% the size of the segmentation of rater 3 (Figure 4d). Rater 2 (Figure 4c) lies exactly between those two raters, so the size of its segmentation is 55% the size of rater 3.

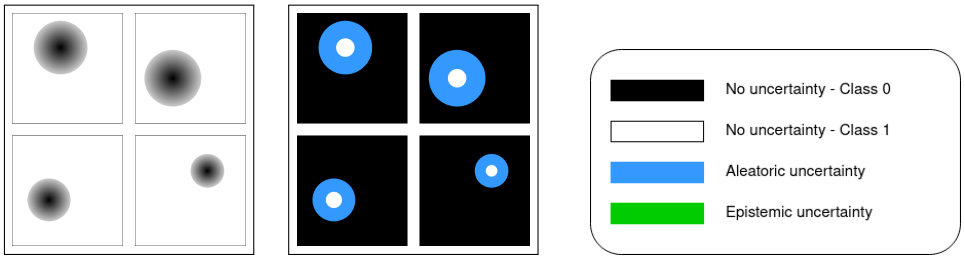
The test set (Figure 4e) is created in the same manner as the training set. The expected uncertainty is shown in Figure 4f with the respective legend in Figure 4g. No epistemic uncertainty should be present in the data when the model converged after training because the test set is created identically to the training set. Instead, only aleatoric uncertainty should be present. This aleatoric uncertainty is expected to be in the ambiguous area of the border of the sphere.

B.1.3 DATA WITH INDUCED EPISTEMIC UNCERTAINTY

Figure 5 shows the data scenario that was created with induced epistemic uncertainty. The input object in the training data (Figure 5a) is a sphere, like in the dataset with aleatoric uncertainty.



(a) Training Data (b) Rater 1 (c) Rater 2 (d) Rater 3



(e) Test Data (f) Expected Uncertainty (g) Legend

Figure 4: Aleatoric data scenario. (a) shows the input images in the training set, which are ambiguous due to Gaussian blur to the outside. (b) - (d) show three different reference ratings that are generated for the input images. (e) shows test images and (f) the expected uncertainty maps. The uncertainty regions are explained in (g).

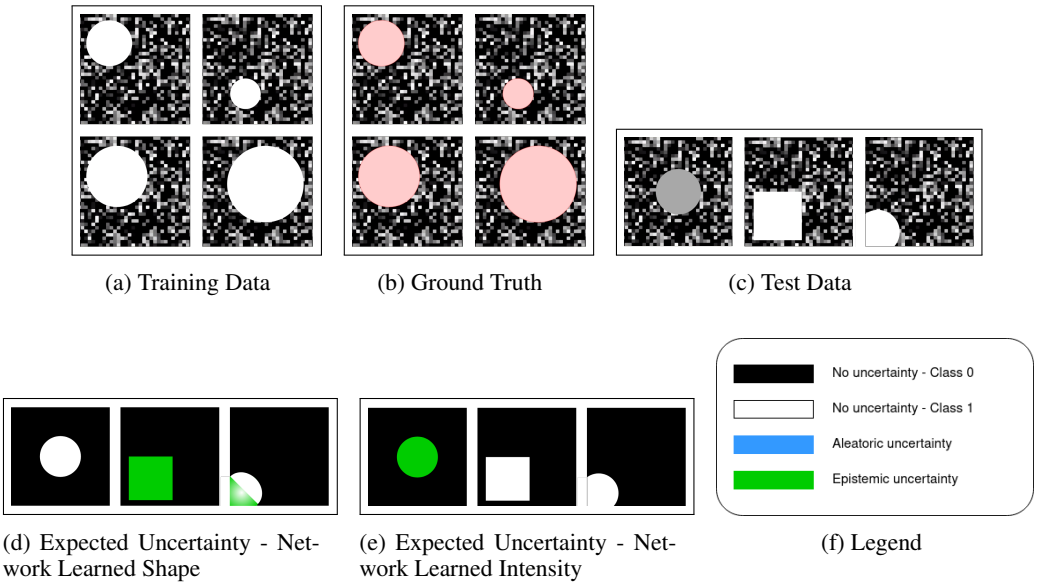
However, for the epistemic data scenario, this sphere has no blur to the outside, to define a clear segmentation boundary for the ground truth segmentation (Figure 5b). As the segmentation problem would be too simple if the background was just black, random noise was added to the background for this case. The test set for this dataset is shown in Figure 5c. It consists of objects of different shapes and colors that were not present in the training data. Some of the objects are still spheres but with varying gray values. Furthermore, there are cubes in the test set and spheres that are partially outside the image while the spheres in the training set were always fully inside the image.

As all segmentations are unique, no aleatoric uncertainty should be present in the data. However, where exactly to expect epistemic uncertainty is not that clear. In some cases, the network might generalize, depending on which features were mainly learned during the training (Geirhos et al. (2020)). For the given toy example it is unknown which training solution the network learned. If it learned to recognize the shape of the object, new shapes should yield a higher uncertainty in the prediction, as shown in Figure 5d. On the other hand, if the network learned the intensity, the epistemic uncertainty might look more like in Figure 5e. It might also be the case that the network learned another decision rule which might result in a different epistemic uncertainty.

B.2 LIDC-IDRI DATASET SETUP

B.2.1 DATASET PREPROCESSING

For preprocessing the dataset, we use the pylidc library (Hancock & Magnan (2016)). With this library, all nodules with size ≥ 3 mm can be queried and clustered, such that each nodule gets assigned up to four raters. We ignore cases that are so close together that they cannot be grouped to one nodule automatically. Further, we calculate a consensus mask which is the union of all raters and ignore cases that have a consensus mask larger than 64 voxels in one direction. We crop patches of size $64 \times 64 \times 64$ with the nodule in the center and all images are resampled to a resolution of $1 \times 1 \times 1$ mm. Also, we only consider nodules in our following analysis that have four reference segmentation masks, which are overall 901 nodules.



(a) Training Data (b) Ground Truth (c) Test Data

(d) Expected Uncertainty - Network Learned Shape (e) Expected Uncertainty - Network Learned Intensity (f) Legend

Figure 5: Epistemic data scenario. (a) shows the input images in the training set. (b) shows the ground truth segmentation. (c) shows test images that differ in various aspects from the training data. (d) and (e) show possible uncertainty maps, depending on what the network learned. The uncertainty regions are explained in (f).

B.2.2 METADATA DISTRIBUTION SHIFT ANALYSIS

Overall, the dataset contains nine different features described in the metadata: *subtlety*, *internal structure*, *calcification*, *sphericity*, *margin*, *lobulation*, *spiculation*, *texture* and *malignancy*. All of these features contain 4-6 different possible categories. Each segmentation rater assigned one of the categories to the metadata features. For inducing distribution shifts, we binarize each metadata feature into two classes (i.i.d. and OoD) instead of the original categories. To not have too much entanglement with aleatoric uncertainty in the distribution shift analysis, we leave out the features *subtlety* and *margin* because if a nodule is subtle, it might be likely that it is not labeled by all raters and if the margin is not sharp, there might be a high variability at the border of the nodule. Further, we do not consider the feature *internal structure*, as it has only one OoD case which makes it unsuitable for a comparison between i.i.d. and OoD cases.

Next, we construct a train/test split to analyze the performance difference of a deterministic U-Net model on the i.i.d. test set and the OoD test set. The way this split is constructed is shown in Figure 6. We first remove all nodules that do not have a majority vote for being i.i.d. or OoD, i.e. when two raters voted for the nodule being i.i.d. and two voted for the nodule being OoD. Next, all patients are identified that have at least one OoD nodule. The OoD nodules of these patients are added to the OoD test set and the i.i.d. nodules of these patients are added to the i.i.d. test set. From the remaining patients that only have i.i.d. nodules, most of the nodules are taken in the i.i.d. training set and some nodules are added to the i.i.d. test set, such that the overall ratio of i.i.d. nodules in the training set and the i.i.d. test set is $80\%/20\%$. The split which cases to include in the training set and the i.i.d. test set is decided by the patient identifier. With the described approach for creating the splits, it is ensured that no patient has nodules in the training- and the test set at the same time.

To measure the performance drop between the i.i.d. and the OoD test set, 5 folds are trained for every metadata split with varying seeds between the folds. The mean Dice between the prediction and one random rater and the standard deviation are calculated on the i.i.d. and the OoD test set to measure the performance. The results are shown in Table 2.

After determining the performance drop on each feature, the two features with the highest performance drops are selected to examine in further experiments. These are the texture shift and the malignancy shift. It can be seen from the results that there is a substantial drop between the i.i.d. and OoD performance which confirms the approach for inducing epistemic uncertainty.

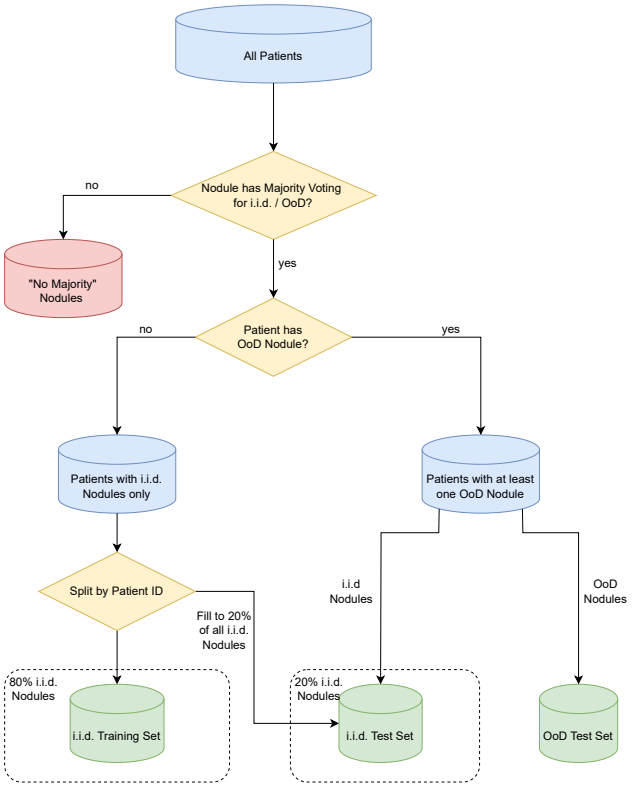


Figure 6: Splits for the LIDC-IDRI shift analysis. Only nodules are considered that have a majority vote for either being i.i.d. or OoD. Furthermore, the splits are created considering the patient ID, so that no nodules of the same patient are in the training set and the test set at the same time. In the end, an i.i.d. training set, an i.i.d. test set, and an OoD test set are created to analyze the shifts between the features.

B.2.3 Setup for evaluation on downstream tasks

To evaluate the performance on the various downstream tasks, the lung nodules are divided into three sets: i.i.d. training set, i.i.d. and OoD test set, and i.i.d. and OoD unlabeled pool. The size of these sets is shown in Table 3. Initially, we only train the model on the i.i.d. training set and assume that its performance on i.i.d. data reaches saturation. After this training, we evaluate the

Table 2: Results for the LIDC-IDRI shift analysis. For each feature, the Dice score on the i.i.d. test set, the Dice score on the OoD test set and the performance drop between i.i.d. and OoD test set are shown. Mean and standard deviation are reported for training with 5 folds, each with a different seed.

Feature	i.i.d. / OoD	Dice i.i.d.	Dice OoD	Performance Drop (%)
Calcification	Absent / Present	0.804 ± 0.0022	0.7669 ± 0.0133	4.6112 ± 1.7699
Sphericity	Round / Linear	0.7934 ± 0.0042	0.7474 ± 0.0106	5.7905 ± 1.191
Lobulation	No Lobulation / Lobulation	0.7887 ± 0.0042	0.7649 ± 0.0046	3.0163 ± 0.6264
Spiculation	No Spiculation / Spiculation	0.7958 ± 0.0022	0.7458 ± 0.0068	6.2865 ± 0.9447
Texture	Solid & Part Solid / Non-solid	0.81 ± 0.0012	0.6081 ± 0.0124	24.9244 ± 1.431
Malignancy	Non-malignant / Malignant	0.7789 ± 0.0051	0.6677 ± 0.0645	14.3093 ± 7.9522

performance of the uncertainty methods on FD, CALIB, and AM. Then, we select samples from the unlabeled pool based on uncertainty rankings. Based on this uncertainty ranking on the unlabeled pool, we determine the performance of a method for detecting OoD samples and add the highest 50% of uncertain samples to the training pool, aiming for improved performance on the OoD test set. With this modified training set, we train another iteration and afterward again measure the test set performance.

Table 3: Size of the different sets in the LIDC dataset for the evaluation on the various downstream tasks.

Split	Train	Val	Test		Unlabeled Pool	
			i.i.d	OoD	i.i.d	OoD
Texture	513	129	167	20	42	20
Malignancy	200	51	105	93	184	92

B.3 GTA5/CITYSCAPES DATASET SETUP

As a further dataset, we use a combination of the GTA5 dataset (Richter et al. (2016)) and the Cityscapes dataset (Cordts et al. (2016)). As mentioned in Sec. 4.2, both datasets contain the same classes, and thus we use the GTA5 dataset as i.i.d. data and the Cityscapes dataset as OoD data.

From the Cityscapes dataset, we use the training set as an unlabeled pool for the AL downstream tasks and the validation set as test set. The scheme for splitting the datasets into training, test set, and unlabeled pool is thereby shown in Figure 7, with the concrete number of images per split. Concretely, we randomly select the same amount of images from the GTA5 dataset to be in the unlabeled pool and further create a $75/25$ training/testing split for the GTA5 dataset.

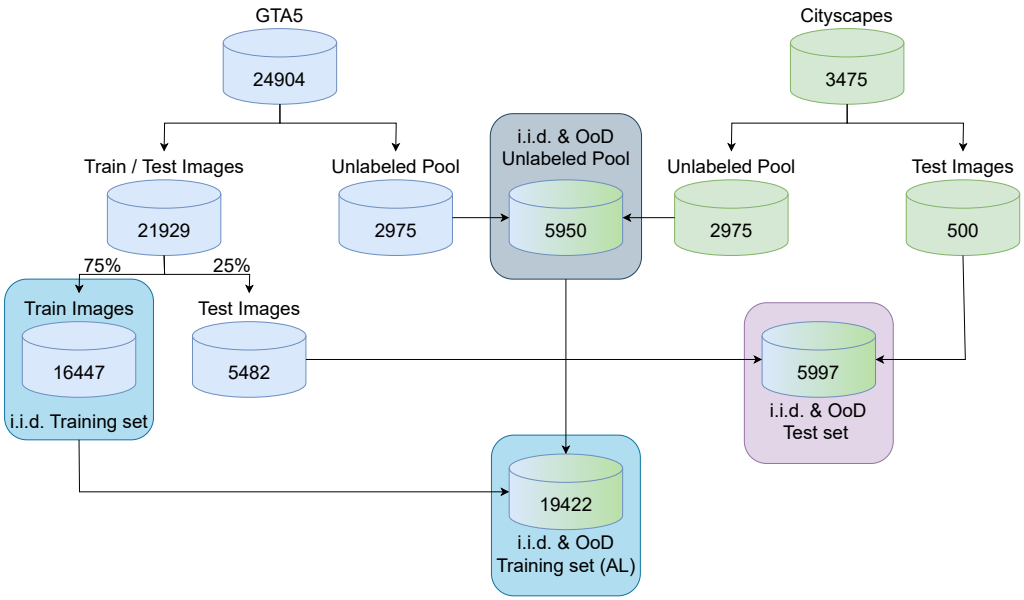


Figure 7: Splits for the GTA5/CS dataset. From the Cityscapes dataset, the training set is used as unlabeled pool and the validation set is used as test set.

The Cityscapes dataset contains up to 30 classes, but only 19 of them are used for validation. As only those 19 classes are contained in the GTA5 dataset, we restrict our analysis to these classes. Further, as mentioned in Sec. 4.2, we perform random class switches for the classes "sidewalk", "person", "car", "vegetation" and "road" with a probability of $\frac{1}{3}$ from <class> to <class 2>. This approach is also applied by Kohl et al. (2018).

All images are first cropped to a size of 1024×1912 and then rescaled to 25% of the size, resulting in an image size of 256×478.

C MODEL IMPLEMENTATION DETAILS

In this section, we describe the implementation details of the model backbones and prediction models. It is important to note that we did not do extensive hyperparameter searches for all the hyperparameters that we state here but rather used standard values if they worked reasonably and if not, we performed sweeps over the hyperparameters to find appropriate settings. We are aware that especially in the implementation of the prediction models, many hyperparameters are involved, and exploring more hyperparameter settings might be an interesting direction for future experiments. All models are trained for 150 epochs.

C.1 SEGMENTATION BACKBONES

U-Net For the toy dataset and the LIDC datasets, we use a 3D U-Net architecture as segmentation backbone. We thereby use an initial filter size of 8 for the toy dataset and 16 for the LIDC datasets and four encoder and four decoder blocks. As loss function, we use a combination of the Dice loss and the cross-entropy loss except for the SSNs as prediction model. For the SSNs, we use the loss function as specified in Monteiro et al. (2020). The Adam optimizer is used with a learning rate of $3e-4$ and a weight decay of $1e-5$. The batch size is set to 8. As augmentations, we apply random flipping and Gaussian noise.

HRNet For the GTA5/CS dataset, we use the HRNet as segmentation backbone, pretrained on ImageNet. As loss function, we use the cross-entropy loss, again, except for the SSNs. For all prediction models except SSNs, SGD is used as optimizer with a learning rate of 0.01, weight decay of $5e-4$, and momentum of 0.9. For the SSNs, RMSprop is used as optimizer with a learning rate of $1e-4$, weight decay of $5e-4$, and momentum of 0.6. The batch size is set to 6. As augmentations, we use random horizontal flipping, rotations, random scaling, random cropping, and Gaussian noise.

C.2 PREDICTION MODELS

Test-time dropout (TTD) For the U-Net, we add dropout after each convolutional block with a probability of $p = 0.5$. For the HRNet, we add dropout at the end of each branch, following Nash et al. (2022). Again, the probability is set to $p = 0.5$. During inference, we perform 10 MC-Dropout forward passes for each input.

Ensemble For the ensemble models, we do not change anything about the models and training schemes themselves but train 5 models with different seeds. During inference, we pass each input image through all 5 models.

Test-time data augmentations (TTA) For the TTA models, we apply the same augmentations as used in training for the 3D U-Net. Thereby, we apply all possible combinations of flipping and Gaussian noise, which result in 16 forward passes per input image (8 possible flipping directions, each with and without noise). For the HRNet, we also apply all possible combinations of random horizontal flipping and Gaussian noise, resulting in 4 forward passes per input image (2 flipping possibilities, each with and without noise).

Stochastic Segmentation Networks (SSNs) For the stochastic segmentation networks, we do 10 forward passes per input image. For the toy dataset and the LIDC datasets, we use a rank of 5 and for the GTA5/CS dataset, we use a rank of 10. As the training behaved more stable when pretraining the mean first, we perform 5 pretraining epochs where we only train the mean before we also train the covariance matrix.

D Uncertainty Measures for Probabilistic Variability Variable Prediction Models

For a probabilistic prediction model $p(Y|x) = \mathbb{E}_{z\sim p(z)}[p(Y|x,z)]$ which predicts the class variable Y given a sample x with an additional variable Z following $p(z)$ which is supposed to capture the variability of the raters/labels (variability variable), we hypothesize that AU und EU can be estimated in a similar fashion as it is done for Bayesian models following

$$\underbrace{H(Y|x)}_{\text{PU}} = \underbrace{\text{MI}(Y,Z)}_{\text{AU (for i.i.d. } x)} + \underbrace{\mathbb{E}_{z\sim Z}[H(Y|z,x)]}_{\text{EU}}. \tag{15}$$

Examples of these methods are the SSNs (Monteiro et al. (2020)), the probabilistic U-Net (Kohl et al. (2018)) or PHiSeg (Baumgartner et al. (2019)) where the prediction model is trained explicitly to learn the variability of the raters.
A more detailed motivation is given in the following two paragraphs and a reason for our observed failure mode is described in the third paragraph.

Aleatoric uncertainty. Multiple plausible predictions for a sample due to ambiguity or other factors are commonly attributed as AU (Monteiro et al. (2020); Kendall & Gal (2017)) and therefore leads to the assumption that the variability variable Z essentially captures the learned AU of the prediction model. Therefore the mutual information between the class label Y and the variability variable Z given a sample x describes how much information about the AU could be gained by obtaining the class label y.

$$\text{MI}(Y,Z|x) = H(Y|x) - \mathbb{E}_{z\sim Z}[H(Y|x,z)] \tag{16}$$

Knowing the optimal variability variable Z would essentially lead to alleviating the uncertainty. Therefore we hypothesize that this uncertainty measure models AU.

Epistemic uncertainty. Following the notion that there is no reason for a variability variable prediction model ever to be unsure about its prediction on i.i.d. data if it is still dependent on the variability variable $p(Y|x,z)$ [3] Therefore the uncertainty of the classifier $H(Y|x)$ which can not be attributed to the variability variable Z should be novel and previously unseen (by the prediction model). Following this line of reasoning, we hypothesize that the expected entropy of the variability variable models EU.

$$\mathbb{E}_{z\sim Z}[H(Y|x,z)] = H(Y|x) - \text{MI}(Y,Z|x) \tag{17}$$

Failure mode. For the SSNs, we observe in our experiments that the model while still dependent on the variability variable is often uncertain in border regions between two classes but generally does not extend to large regions of the image [4]. This offers an explanation why for the experiments on the LIDC-IDRI dataset, where for most samples the disagreement between raters is present purely in the border regions of the nodule, $\text{MI}(Y,Z|x)$ has the lowest NCC scores (Q1 + Q2).

E Uncertainty Measures for Test-Time Augmentation Models

Given a model using a set label preserving data augmentations during inference which are defined on the input space $\mathcal{T}$ and used the form of a random variable T (support$(T) = \mathcal{T}$) from which samples are drawn from $t \sim T$. The inference using test-time augmentations can can be described as $p(Y|x) = \mathbb{E}_{t\sim T}[p(Y|t,x)] = \mathbb{E}_{t\sim T}[p(Y|t(x))]$. During training the model is optimized on the training set $\mathcal{D}$ with a training objective (usually the cross-entropy loss (CE-Loss)) to be invariant against augmentations in $\mathcal{D}$. Given an optimal model for which the training objective is minimal (e.g. CE-Loss=0), the outputs of the model on the training set $\mathcal{D}$ are fully invariant to all transformations $p(Y|x,t_1) = p(Y|x,t_2) \forall t_1, t_2 \in \mathcal{T}, x \in \mathcal{D}$.

[3] This is essentially designed into the training of the variability variable prediction models. E.g. for the SSNs, this is done so by using the logsumexp of the logarithmic loss Monteiro et al. (2020).

[4] We hypothesize that this behavior arises due to $p(z)$ modeling a Gaussian distribution in logit space.

For this model, we hypothesize that AU und EU can be estimated in a similar fashion as it is done for Bayesian models following

$$\underbrace{H(Y|x)}_{\text{PU}} = \underbrace{\text{MI}(Y,T)}_{\text{EU}} + \underbrace{\mathbb{E}_{t\sim T}[H(Y|t,x)]}_{\text{AU (for i.i.d. } x)}. \tag{18}$$

A more detailed motivation is given in the following two paragraphs.

Aleatoric uncertainty. As our model is perfectly trained on the training set, the model is able to detect previously seen uncertainty over the augmentations similar to a Bayesian model can do so with each set of parameters Mukhoti et al. (2021). Therefore, the expected entropy over the augmentations should give information about the amount of AU in the prediction of a datapoint.

$$\mathbb{E}_{t\sim T}[H(Y|x,t)] \tag{19}$$

Epistemic uncertainty. As our model is invariant to augmentations on the training set it also follows that $\text{MI}(Y,T|x) = 0 \forall x \in \mathcal{D}$ Jensen (1906). If the mutual information between the augmentation variable and predicted label is greater than zero ($\text{MI}(Y,T|\hat{x}) > 0$) for a datapoint $\hat{x} \notin \mathcal{D}$, then this indicates that this datapoint deviates in some form from $\mathcal{D}$. Further, if $\hat{x}$ would be added to the training set and the model retrained, the model would have learned to be invariant against the augmentations for this datapoint. Following this argumentation, this term is therefore reducible by adding previously unseen datapoints. Based on this, we hypothesize that the mutual information between the augmentation variable and the predicted label models EU.

$$\text{MI}(Y,T|x) = H(Y|x) - \mathbb{E}_{t\sim T}[H(Y|x,t)] \tag{20}$$

Implications. Based on these derivations it seems that TTA actually allows the model to estimate EU, rather than improving the estimation of AU. This falls in line with the hypothesis made by Hu et al. (2019) and directly opposes the claims of two prominent papers claiming it models AU (Wang et al. (2019); Ayhan & Berens (2018)).

F Details on the aggregation strategies

F.1 Ablation study: Correlation of image level aggregation and object size

To confirm the hypothesis about the correlation between the object size and the amount of uncertainty, we generated plots to see the connection between those two variables for the LIDC datasets. One of the generated plots is shown in Figure 8. This plot is for a TTD model on the LIDC TEX dataset. In the top row, the aggregated amount of uncertainty compared to the mean size of the predicted segmentation is shown for the epistemic, the aleatoric, and the predictive uncertainty. In the bottom row, the summed uncertainty is divided by the object size. To confirm that the size of the predicted segmentation corresponds to the ground truth segmentation size, the two variables are plotted on the right-hand side. It can be seen, that a positive correlation between the aggregation sum and the object size is given in the top row, but if the aggregation mean is taken in the bottom row, this correlation is not present. This means that the summed uncertainty only correlates with the size of the objects and does not represent the objects' uncertainty independent of the size.

F.2 Selection of threshold for threshold level aggregation

For the threshold level aggregation, we need to determine a threshold where the pixels that are above this are considered as "uncertain". Intuitively, most uncertainty is likely to be at the border of the object and thus correlates with the object size. Therefore, the threshold is calculated with respect to the object sizes in the validation set in the following way: First, the mean foreground ratio α over all predicted segmentations in the validation set is determined:

$$\alpha = \frac{\#\text{voxels foreground pred}}{\#\text{voxels}} \tag{21}$$

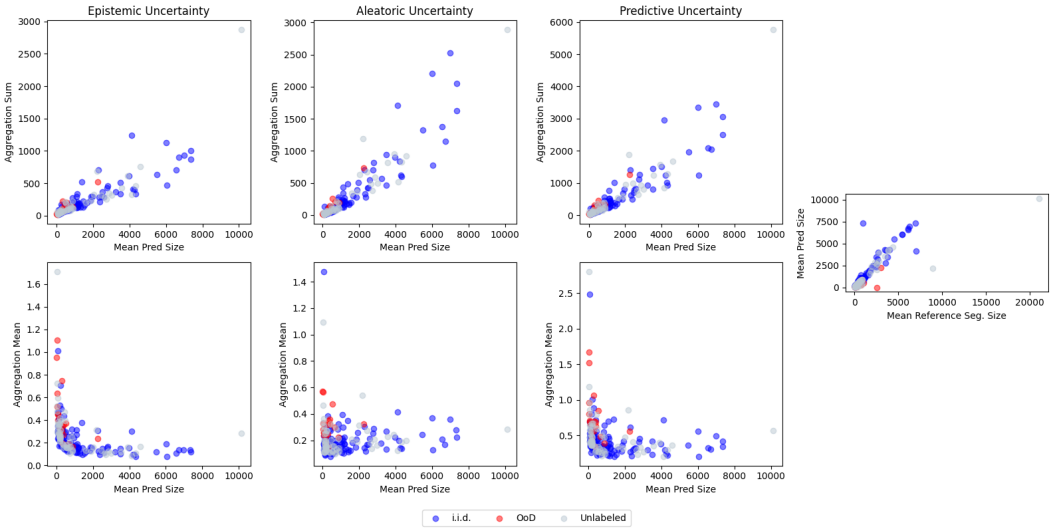


Figure 8: Correlation between object size and uncertainty for image level aggregation. In the top row, all pixels in the uncertainty maps are added up and this aggregation sum is plotted with respect to the mean size of the predicted segmentations. In the bottom row, the aggregation sum is additionally divided by the predicted object size, resulting in the aggregation mean. On the right-hand side, the mean prediction size is plotted with respect to the mean reference segmentation size to see that the size of the predictions roughly corresponds to the reference segmentation sizes of the objects.

With this foreground ratio, the quantile value q is calculated with $q = 1 - \alpha$. This quantile value is applied on the predicted uncertainty maps of the validation set u_{val}, to determine a pixel value of pixels that lie in that quantile Q. This pixel value then serves as a threshold for later predicted images:

$$\text{threshold} = Q(q, u_{\text{val}}) \tag{22}$$

With this method, one threshold per uncertainty modeling method is determined.

G Detailed results of the separation study

G.1 Detailed analysis

Q1 & Q2

Toy dataset. In the toy dataset analysis, AU uncertainty measures have generally higher NCC values compared to EU uncertainty measures, indicating a successful separation of AU and the highlighting of relevant areas (Q1) which is also supported by the qualitative analysis with high uncertainty signals in areas with rater disagreement (see Sec. G.3). Meanwhile, the EU-measures perform worse than PU and AU measures indicating that EU-measures do not measure AU. An exception to this finding are SSNs, where NCC scores are higher for EU-measures compared to other prediction models. This discrepancy may be attributed to the presence of AU at the border regions which is not explained by the variability variable.

LIDC datasets. On the LIDC datasets, EU-measures performance is similar to that of AU-measures. Therefore the approaches seem to model EU in the areas attributed to AU (Q2).
In fact, for SSNs, the AU-measure even shows a lower NCC than the EU-measure, which could be attributed to the SSNs rating the border regions with high EU, which are the regions of disagreement. The qualitative analysis shows that a slightly better indication of AU by AU-measures becomes apparent when there is meaningful inter-rater variability beyond small border regions (see Figure 10). Interestingly, this effect is only noticeable for i.i.d. nodules. For example, the same nodule that

shows a good indication for AU in the i.i.d. test set on the LIDC TEX dataset (Figure 10) shows a poor indication of AU in the OoD test set on the LIDC MAL dataset (Figure 13).

GTA5/Cityscapes dataset. For the GTA5/CS dataset, the NCC scores are generally lower compared to the other datasets. However, for AU-measures, the NCC scores are at least positively correlated, while the EU-measures are mostly even negatively correlated with the AU, showing that they really do not model AU. The only prediction that reaches a high AU with its respective AU-measure are SSNs. This qualitative difference can also be seen in Figure 14: While most prediction models show the highest AU at the borders of the object, SSNs AU-measure highlight the whole ambiguous area.

Q3 & Q4

Toy dataset. In Setting 2, where only EU is present in the data, there is no significant difference in AUROC between AU-measures and EU-measures. It could be assumed that in the absence of learning AU in the training data, every uncertainty measure can be interpreted as EU-measure. However, as soon as AU is introduced into the training data in settings 3a and 3b, EU-measures become a better separator between i.i.d and OoD data. In setting 3b, where AU is present in both the training and test data, the separation of EU becomes beneficial. To address Q3 and Q4 on the toy dataset, it can be seen that the AUROC retrieved with EU-measures is almost always better than random, confirming Q3. The answer to Q4 depends on the amount of AU present in the training and test data.

LIDC datasets. On the LIDC datasets, it is evident that the separation between AU and EU brings particular benefits for TTD, Ensembles, and TTA. Specifically, on LIDC TEX, EU-measures prove to be a more effective separator between i.i.d. and OoD data. Overall, it can be seen that whenever the separation between PU and EU is advantageous, AU as an OoD-detector performs worse than random. Another hypothesis that arises is that the separation of EU appears to be most beneficial in settings where the OoD-detection performance is not yet saturated, such as in the case of LIDC TEX. To summarize the answer for Q3 and Q4 for the LIDC dataset, the AUROC for EU-measures is always better than random, confirming Q3. However, Q4 can only be partially confirmed in the sense that AU is not a good measure whenever separating EU from PU is beneficial.

GTA5/Cityscapes dataset. For the GTA5/CS dataset, most EU-measures significantly outperform the respective AU-measure by means of the AUROC. The only exception is TTD, where the EU-measure even performs worse than the AU-measure. Besides that, the AU-measures are even below random performance for the patch-level aggregation, while for the image-level aggregation, they perform slightly better than random. This indicates in summary with regards to Q3, that EU-measures, except for TTD, capture EU, while for Q4, it can be at least mostly confirmed that AU-measures do not consistently outperform random selection of OoD cases.

G.2 Quantitative results

The detailed quantitative results for the separation study, presented in Sec. 4.4, can be found in Table 4. Table 4a provides insights on answering Q1 and Q2, while Table 4b addresses Q3 and Q4.

Table 4: Quantitative results for the separation study. In order to answer Q1 and Q2 from the separation study, the NCC scores are calculated between the uncertainty maps and the variance of the reference segmentations, shown in Table 4a. To answer Q3 and Q4, the AUROC scores are calculated and reported in Table 4b. Mean results are shown over 3 runs with different seeds for all relevant dataset settings to answer the respective questions. Abbreviations: PM: Prediction model, UM: Uncertainty measure, UT: Modeled uncertainty Type (according to theory), AGG: Aggregation strategy.

Testset	PM	UM	UT	Toy 1	LIDC TEX	LIDC MAL	GTA5/CS
i.i.d	**Determ.**	**MSR**	**PU**	0.68	0.32	0.28	0.51
	TTD	**PE**	**PU**	0.80	0.51	0.48	0.27
		EE	**AU**	0.86	0.52	0.48	0.28
		MI	**EU**	0.47	0.46	0.45	-0.23
	Ensemble	**PE**	**PU**	0.83	0.48	0.43	0.24
		EE	**AU**	0.84	0.49	0.44	0.27
		MI	**EU**	0.51	0.39	0.36	-0.23
	TTA	**PE**	**PU**	0.82	0.46	0.41	0.25
		EE	**AU**	0.82	0.48	0.42	0.26
		MI	**EU**	0.54	0.38	0.35	-0.16
	SSN	**PE**	**PU**	0.96	0.63	0.61	0.56
		MI	**AU**	0.96	0.59	0.55	0.70
		EE	**EU**	0.80	0.64	0.62	0.05
OoD	**Determ.**	**MSR**	**PU**	-	0.20	0.20	0.47
	TTD	**PE**	**PU**	-	0.37	0.36	0.26
		EE	**AU**	-	0.37	0.39	0.26
		MI	**EU**	-	0.33	0.31	-0.13
	Ensemble	**PE**	**PU**	-	0.35	0.33	0.25
		EE	**AU**	-	0.36	0.35	0.30
		MI	**EU**	-	0.30	0.27	-0.06
	TTA	**PE**	**PU**	-	0.32	0.30	0.28
		EE	**AU**	-	0.33	0.33	0.30
		MI	**EU**	-	0.27	0.25	-0.04
	SSN	**PE**	**PU**	-	0.51	0.47	0.37
		MI	**AU**	-	0.47	0.44	0.52
		EE	**EU**	-	0.52	0.47	0.03

(a) NCC scores

PM	UM	UT	AGG	Toy 2	Toy 3a	Toy 3b	LIDC TEX	LIDC MAL	GTA5/CS
Determ.	**MSR**	**PU**	**Patch**	0.84	0.78	0.41	0.46	0.86	0.33
			Thresh	0.73	0.45	0.40	0.52	0.59	-
			Image	-	-	-	-	-	0.70
TTD	**PE**	**PU**	**Patch**	0.83	0.68	0.37	0.46	0.90	0.37
			Thresh	0.48	0.50	0.38	0.61	0.74	-
			Image	-	-	-	-	-	0.68
	EE	**AU**	**Patch**	0.74	0.69	0.36	0.43	0.90	0.37
			Thresh	0.53	0.53	0.29	0.40	0.88	-
			Image	-	-	-	-	-	0.68
	MI	**EU**	**Patch**	0.83	0.61	0.73	0.52	0.88	0.46
			Thresh	0.54	0.43	0.71	0.65	0.60	-
			Image	-	-	-	-	-	0.51
Ensemble	**PE**	**PU**	**Patch**	0.95	0.94	0.50	0.55	0.91	0.33
			Thresh	0.90	0.73	0.69	0.66	0.72	-
			Image	-	-	-	-	-	0.72
	EE	**AU**	**Patch**	0.94	0.83	0.44	0.49	0.89	0.29
			Thresh	0.78	0.19	0.12	0.53	0.53	-
			Image	-	-	-	-	-	0.67
	MI	**EU**	**Patch**	0.95	0.95	0.85	0.65	0.89	0.91
			Thresh	0.91	0.77	0.87	0.72	0.75	-
			Image	-	-	-	-	-	0.90
TTA	**PE**	**PU**	**Patch**	0.95	0.91	0.48	0.51	0.88	0.32
			Thresh	0.93	0.66	0.55	0.60	0.67	-
			Image	-	-	-	-	-	0.70
	EE	**AU**	**Patch**	0.95	0.83	0.44	0.46	0.87	0.29
			Thresh	0.89	0.27	0.16	0.49	0.53	-
			Image	-	-	-	-	-	0.67
	MI	**EU**	**Patch**	0.95	0.94	0.92	0.59	0.86	0.93
			Thresh	0.93	0.71	0.84	0.67	0.70	-
			Image	-	-	-	-	-	0.94
SSN	**PE**	**PU**	**Patch**	0.87	0.76	0.38	0.54	0.84	0.78
			Thresh	0.74	0.65	0.34	0.51	0.72	-
			Image	-	-	-	-	-	0.82
	MI	**AU**	**Patch**	0.68	0.63	0.32	0.54	0.72	0.53
			Thresh	0.67	0.51	0.25	0.49	0.57	-
			Image	-	-	-	-	-	0.55
	EE	**EU**	**Patch**	0.87	0.90	0.43	0.54	0.85	0.78
			Thresh	0.74	0.68	0.78	0.50	0.68	-
			Image	-	-	-	-	-	0.86

(b) AUROC scores

G.3 QUALITATIVE RESULTS

In the following sections, samples are shown for the qualitative analysis to answer Q1 and Q2 from the separation study (see Sec. 4.4).

G.3.1 QUALITATIVE RESULTS FOR THE TOY DATASET

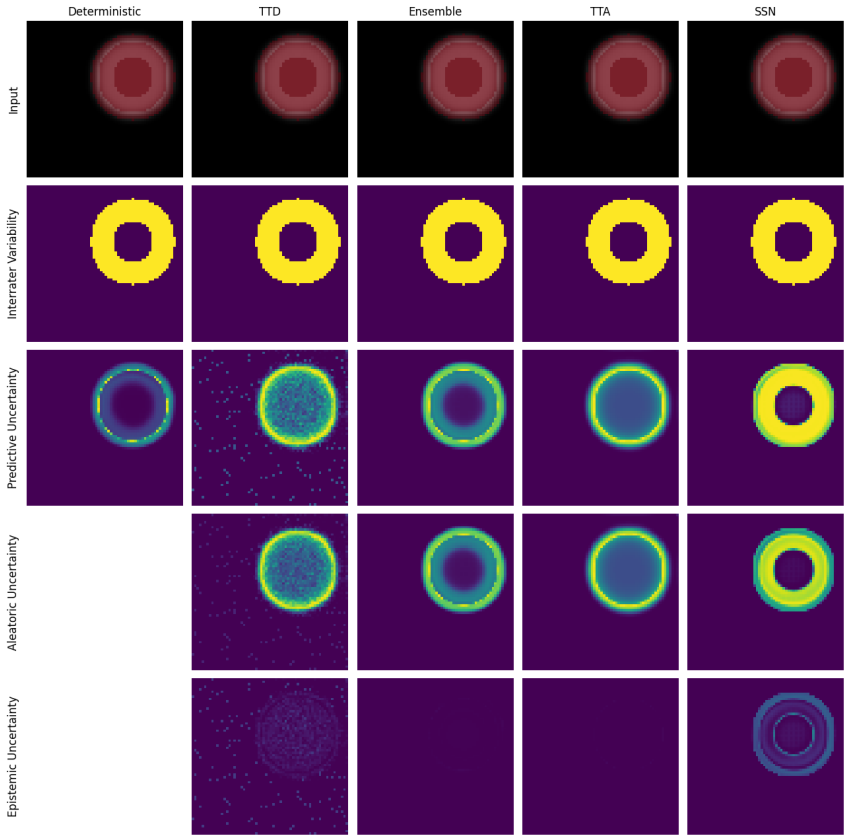


Figure 9: Qualitative results for separating aleatoric and epistemic uncertainty for the toy dataset. The reference segmentations are shown as overlay over the input image. Further, the interrater variability based on the pixel variance is shown. The uncertainty scores per pixel are normalized between 0 and 0.5 for the deterministic model and between 0 and 0.7 for the other prediction models, reflecting the possible range of uncertainty values.

G.3.2 QUALITATIVE RESULTS FOR THE LIDC-IDRI DATASETS

Texture shift i.i.d. example

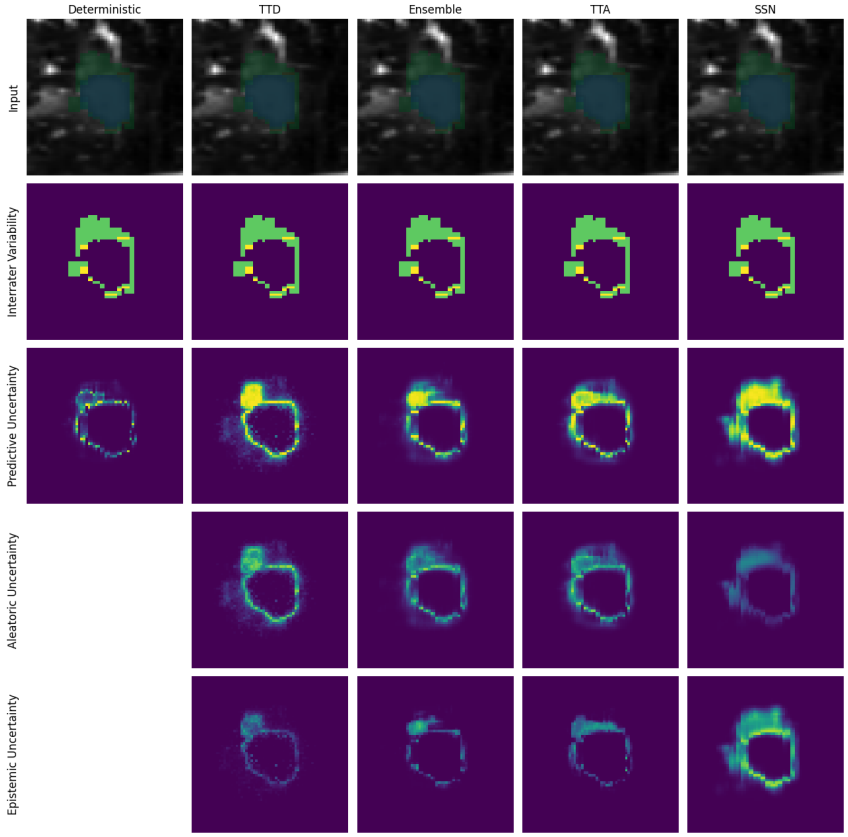


Figure 10: Qualitative results for separating aleatoric and epistemic uncertainty for the LIDC TEX dataset. A case that is part of the i.i.d. test set is shown. The reference segmentations are shown as overlay over the input image. Further, the interrater variability based on the pixel variance is shown. The uncertainty scores per pixel are normalized between 0 and 0.5 for the deterministic model and between 0 and 0.7 for the other prediction models, reflecting the possible range of uncertainty values.

Texture shift OoD example

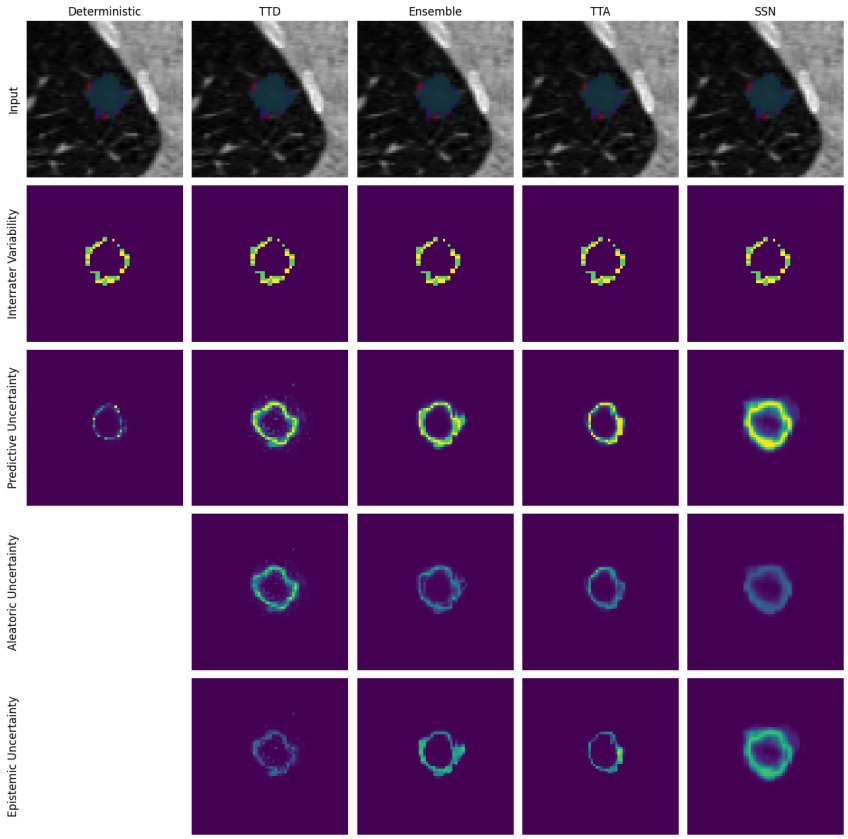


Figure 11: Qualitative results for separating aleatoric and epistemic uncertainty for the LIDC TEX dataset. A case that is part of the OoD test set is shown. The reference segmentations are shown as overlay over the input image. Further, the interrater variability based on the pixel variance is shown. The uncertainty scores per pixel are normalized between 0 and 0.5 for the deterministic model and between 0 and 0.7 for the other prediction models, reflecting the possible range of uncertainty values.

Malignancy shift i.i.d. example

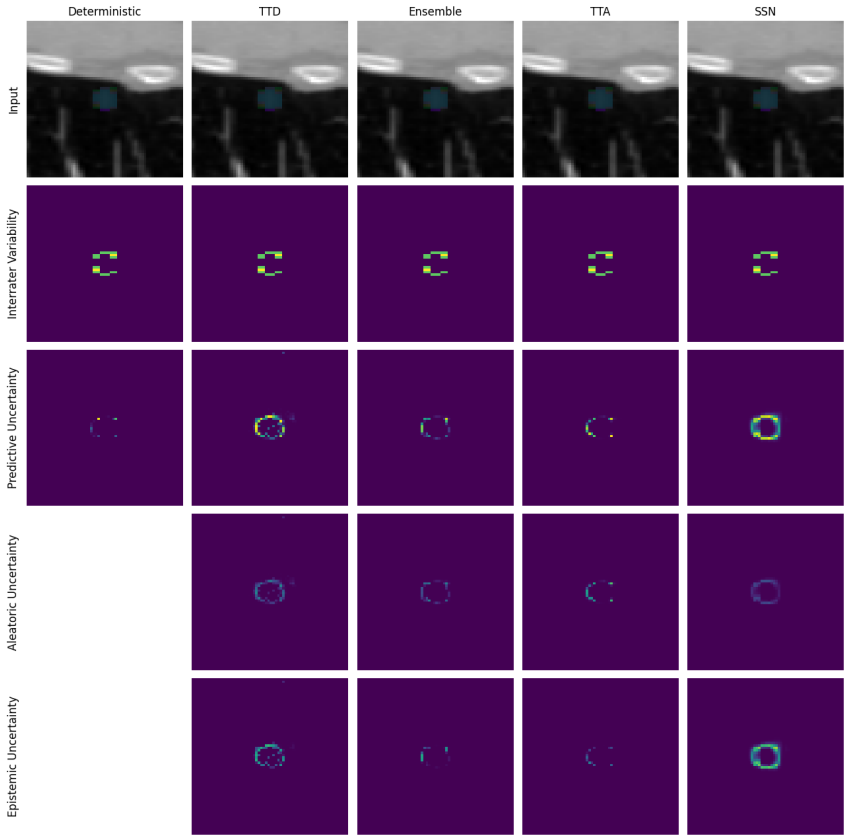


Figure 12: Qualitative results for separating aleatoric and epistemic uncertainty for the LIDC MAL dataset. A case that is part of the i.i.d. test set is shown. The reference segmentations are shown as overlay over the input image. Further, the interrater variability based on the pixel variance is shown. The uncertainty scores per pixel are normalized between 0 and 0.5 for the deterministic model and between 0 and 0.7 for the other prediction models, reflecting the possible range of uncertainty values.

Malignancy shift OoD example

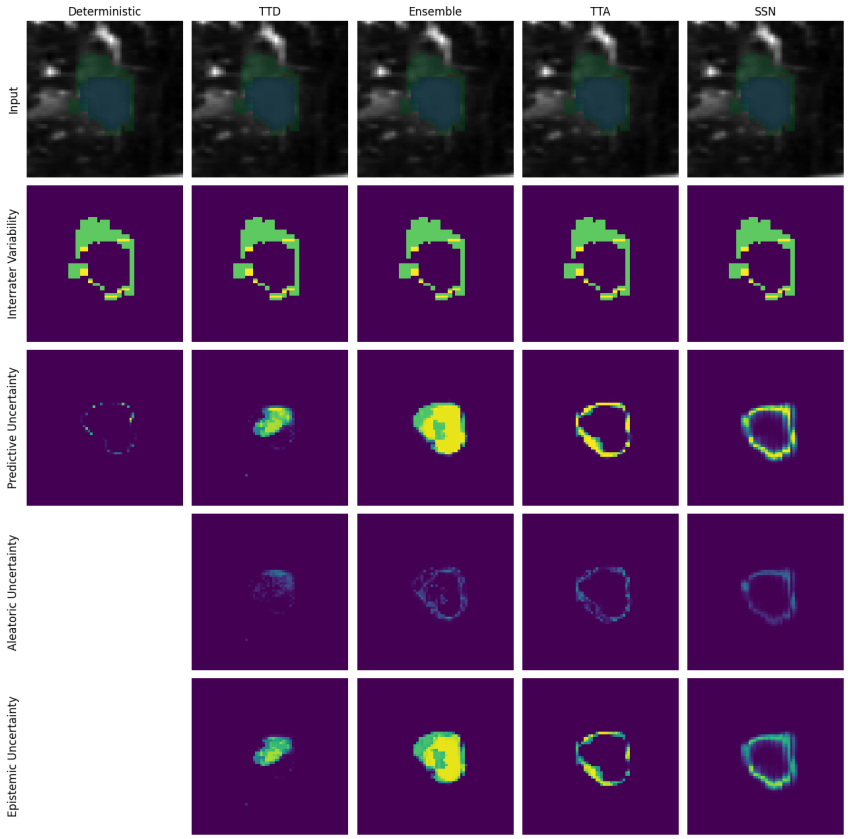


Figure 13: Qualitative results for separating aleatoric and epistemic uncertainty for the LIDC MAL dataset. A case that is part of the OoD test set is shown. The reference segmentations are shown as overlay over the input image. Further, the interrater variability based on the pixel variance is shown. The uncertainty scores per pixel are normalized between 0 and 0.5 for the deterministic model and between 0 and 0.7 for the other prediction models, reflecting the possible range of uncertainty values.

G.3.3 QUALITATIVE RESULTS FOR THE GTA 5 / CITYSCAPES DATASET

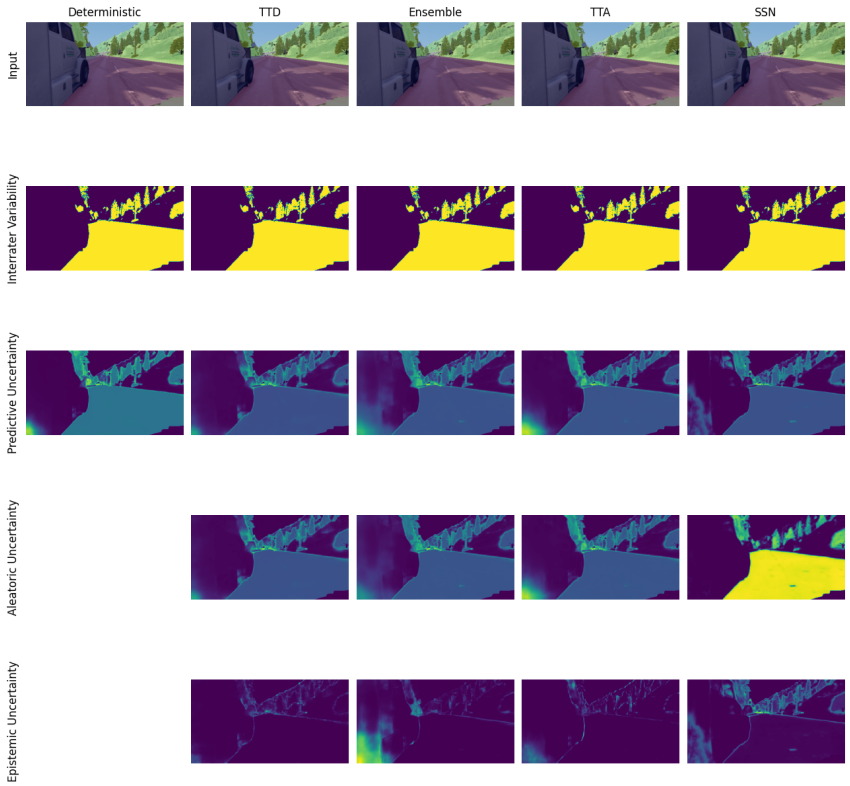


Figure 14: Qualitative results for separating aleatoric and epistemic uncertainty for the GTA 5 / Cityscapes dataset. The reference segmentations are shown as overlay over the input image. Further, the interrater variability based on the pixel variance is shown. The uncertainty scores per pixel are normalized per image.

H DETAILED RESULTS OF THE EVALUATION ON DOWNSTREAM TASKS

The following tables show the detailed results on the downstream tasks as described in Sec. 4.5. For the LIDC datasets, the results are shown in Table 5, while for the GTA5/CS dataset, the results are shown in Table 6.

Table 5: **Evaluation of downstream tasks on the LIDC datasets.** The table shows the segmentation performance by means of the Dice score and evaluation metrics for 5 different downstream tasks, where ↑ depict higher scores are better and ↓ lower scores are better. All scores are multiplyed by 10^2. The color heatmap is normalized per column and per shift, brighter columns imply better scores. For AL, the second cycle was only executed with EU and PU, indicated by empty grey entries for AU. Reported results show the mean and standard deviation over 3 different seeds. Abbreviations: PM: Prediction model, UM: Uncertainty measure, UT: Modeled uncertainty Type (according to theory), AGG: Aggregation strategy.

					Seg. Performance		OoD-D	Failure Detection				AL	Calibration		Ambiguity Modeling			
Shift	PM	UM	UT	AGG	Dice i.i.d. ↑	Dice OoD ↑	AUROC ↑	AURC i.i.d. ↓	AURC OoD ↓	EAURC i.i.d. ↓	EAURC OoD ↓	AL Improv. OoD ↑	ACE i.i.d. ↓	ACE OoD ↓	NCC i.i.d. ↑	NCC OoD ↑	GED i.i.d. ↓	GED OoD ↓
Texture-Shift	Determ.	**MSR**	**PU**	**Patch**	81.42±0.28	64.05±0.1	46.43±2.28	19.18±0.24	41.27±1.19	6.13±0.09	18.64±1.27	-0.78±1.93	29.92±0.7	32.58±0.99	31.55±0.26	20.09±1.57	22.18±0.53	52.06±0.18
				Thresh	81.42±0.28	64.05±0.1	52.02±11.21	17.7±1.02	41.82±1.54	4.65±0.82	19.19±1.18	-2.57±3.49	29.92±0.7	32.58±0.99	31.55±0.26	20.09±1.57	22.18±0.53	52.06±0.18
	TTD	**PE**	**PU**	**Patch**	81.1±0.14	63.9±1.43	46.35±1.77	19.97±0.09	42.52±2.12	6.52±0.16	19.05±1.18	1.06±1.0	19.47±0.67	21.98±1.04	51.17±0.13	36.58±1.73	16.83±0.26	40.84±2.01
				Thresh	81.1±0.14	63.9±1.43	60.95±2.3	17.35±0.51	37.97±1.82	3.9±0.42	14.5±1.02	0.9±0.76	19.47±0.67	21.98±1.04	51.17±0.13	36.58±1.73	16.83±0.26	40.84±2.01
		EE	**AU**	**Patch**	81.1±0.14	63.9±1.43	43.41±1.38	20.25±0.04	41.65±3.65	6.8±0.13	18.19±2.78		26.95±0.27	26.86±1.66	51.74±0.26	37.21±2.13	16.83±0.26	40.84±2.01
				Thresh	81.1±0.14	63.9±1.43	40.4±1.68	20.03±0.25	42.24±2.22	6.57±0.17	18.77±1.28		26.95±0.27	26.86±1.66	51.74±0.26	37.21±2.13	16.83±0.26	40.84±2.01
		MI	**EU**	**Patch**	81.1±0.14	63.9±1.43	51.55±0.9	19.51±0.12	42.15±1.92	6.06±0.09	18.69±1.01	2.05±2.57	25.42±0.12	27.82±1.35	46.45±0.35	33.33±1.41	16.83±0.26	40.84±2.01
				Thresh	81.1±0.14	63.9±1.43	65.2±1.38	16.86±0.28	36.3±2.34	3.41±0.18	12.83±1.78	3.24±1.92	25.42±0.12	27.82±1.35	46.45±0.35	33.33±1.41	16.83±0.26	40.84±2.01
	Ensemble	**PE**	**PU**	**Patch**	82.34±0.18	64.36±0.85	54.72±2.55	18.24±0.29	37.06±4.43	5.76±0.2	15.05±3.67	1.02±3.04	19.96±0.91	24.7±0.54	47.82±0.55	34.62±1.03	16.05±0.35	40.37±1.09
				Thresh	82.34±0.18	64.36±0.85	66.47±5.2	16.07±0.39	38.38±2.84	3.59±0.37	16.37±2.06	0.56±2.53	19.96±0.91	24.7±0.54	47.82±0.55	34.62±1.03	16.05±0.35	40.37±1.09
		EE	**AU**	**Patch**	82.34±0.18	64.36±0.85	48.57±0.74	18.53±0.26	37.49±5.1	6.05±0.17	15.47±4.41		27.21±0.28	29.18±1.92	49.31±0.46	35.52±0.82	16.05±0.35	40.37±1.09
				Thresh	82.34±0.18	64.36±0.85	52.66±4.31	16.7±0.35	41.57±2.44	4.22±0.31	19.55±1.97		27.21±0.28	29.18±1.92	49.31±0.46	35.52±0.82	16.05±0.35	40.37±1.09
		MI	**EU**	**Patch**	82.34±0.18	64.36±0.85	65.4±3.71	17.55±0.29	38.15±5.64	5.06±0.22	16.13±4.92	0.96±3.48	28.71±0.38	31.35±1.58	39.05±0.8	30.1±0.84	16.05±0.35	40.37±1.09
				Thresh	82.34±0.18	64.36±0.85	72.1±5.05	15.58±0.32	36.76±2.92	3.1±0.3	14.75±2.13	-0.75±1.84	28.71±0.38	31.35±1.58	39.05±0.8	30.1±0.84	16.05±0.35	40.37±1.09
	TTA	**PE**	**PU**	**Patch**	81.83±0.3	65.08±0.45	50.95±1.4	18.57±0.25	36.28±1.96	5.76±0.12	14.38±1.92	0.45±1.97	22.78±1.01	26.07±2.37	46.04±0.31	31.51±0.22	17.37±0.6	41.49±0.9
				Thresh	81.83±0.3	65.08±0.45	59.8±10.85	16.77±0.52	39.21±0.91	3.97±0.38	17.31±0.78	-0.25±1.44	22.78±1.01	26.07±2.37	46.04±0.31	31.51±0.22	17.37±0.6	41.49±0.9
		EE	**AU**	**Patch**	81.83±0.3	65.08±0.45	46.23±1.89	18.96±0.09	39.39±0.5	6.15±0.13	17.49±0.62		26.45±0.15	28.32±1.87	47.57±0.47	33.48±0.33	17.37±0.6	41.49±0.9
				Thresh	81.83±0.3	65.08±0.45	48.93±11.85	17.6±0.77	40.74±1.09	4.79±0.65	18.84±0.96		26.45±0.15	28.32±1.87	47.57±0.47	33.48±0.33	17.37±0.6	41.49±0.9
		MI	**EU**	**Patch**	81.83±0.3	65.08±0.45	59.09±2.68	17.52±0.43	33.53±3.36	4.71±0.28	11.63±3.29	1.52±0.43	28.95±0.59	30.61±0.39	38.08±0.46	26.68±0.36	17.37±0.6	41.49±0.9
				Thresh	81.83±0.3	65.08±0.45	67.38±7.09	16.31±0.57	36.14±2.59	3.51±0.42	14.23±2.47	0.3±2.32	28.95±0.59	30.61±0.39	38.08±0.46	26.68±0.36	17.37±0.6	41.49±0.9
	SSN	**PE**	**PU**	**Patch**	81.28±0.16	59.51±1.79	53.57±3.02	19.6±0.42	40.79±6.31	6.41±0.39	16.2±5.62	-5.11±8.76	9.68±0.35	17.87±2.1	63.49±0.41	51.22±0.11	13.57±0.3	35.69±2.43
				Thresh	81.28±0.16	59.51±1.79	50.52±5.24	18.47±0.04	49.89±4.56	5.28±0.02	25.3±3.72	-7.24±8.79	9.68±0.35	17.87±2.1	63.49±0.41	51.22±0.11	13.57±0.3	35.69±2.43
		MI	**AU**	**Patch**	81.28±0.16	59.51±1.79	54.48±6.47	19.31±0.76	37.86±2.83	6.11±0.73	13.26±2.12		27.71±0.29	27.98±1.09	59.2±0.63	47.33±0.54	13.57±0.3	35.69±2.43
				Thresh	81.28±0.16	59.51±1.79	49.4±6.31	18.49±0.6	36.24±3.38	5.3±0.56	11.65±2.09		27.71±0.29	27.98±1.09	59.2±0.63	47.33±0.54	13.57±0.3	35.69±2.43
		EE	**EU**	**Patch**	81.28±0.16	59.51±1.79	54.13±1.43	19.54±0.33	41.57±6.22	6.34±0.3	16.98±5.69	-0.73±9.19	17.16±0.31	22.56±0.4	63.61±0.47	51.8±0.23	13.57±0.3	35.69±2.43
				Thresh	81.28±0.16	59.51±1.79	50.44±6.67	18.61±0.28	50.56±3.84	5.42±0.33	25.97±3.06	-5.7±10.94	17.16±0.31	22.56±0.4	63.61±0.47	51.8±0.23	13.57±0.3	35.69±2.43
Malignancy-Shift	Determ.	**MSR**	**PU**	**Patch**	78.97±0.29	65.03±3.36	86.44±2.33	19.79±0.16	36.1±9.33	4.77±0.27	17.87±8.27	0.79±8.0	32.17±0.96	28.41±1.45	28.39±1.24	20.15±1.78	26.42±0.63	54.57±5.45
				Thresh	78.97±0.29	65.03±3.36	58.94±4.42	21.86±0.19	48.62±8.06	6.84±0.38	30.39±6.94	-1.65±7.86	32.17±0.96	28.41±1.45	28.39±1.24	20.15±1.78	26.42±0.63	54.57±5.45
	TTD	**PE**	**PU**	**Patch**	78.6±0.41	67.84±1.62	90.13±0.07	20.6±0.31	22.93±3.61	4.85±0.34	5.39±4.15	2.27±4.26	21.09±0.63	19.13±0.82	47.88±0.27	35.6±1.53	19.88±0.74	39.09±2.89
				Thresh	78.6±0.41	67.84±1.62	73.98±7.04	22.08±0.69	25.97±3.92	6.34±0.67	8.43±4.47	2.68±4.48	21.09±0.63	19.13±0.82	47.88±0.27	35.6±1.53	19.88±0.74	39.09±2.89
		EE	**AU**	**Patch**	78.6±0.41	67.84±1.62	90.5±0.56	20.55±0.35	27.35±5.72	4.81±0.43	9.81±6.14		28.34±0.17	24.24±0.63	48.22±0.48	38.51±1.22	19.88±0.74	39.09±2.89
				Thresh	78.6±0.41	67.84±1.62	87.62±1.42	22.29±0.75	39.51±19.5	6.54±0.71	21.97±2.45		28.34±0.17	24.24±0.63	48.22±0.48	38.51±1.22	19.88±0.74	39.09±2.89
		MI	**EU**	**Patch**	78.6±0.41	67.84±1.62	88.12±1.29	20.6±0.34	21.96±2.13	4.85±0.36	4.43±2.68	1.84±4.39	26.6±0.38	24.43±0.37	44.55±0.08	31.08±0.89	19.88±0.74	39.09±2.89
				Thresh	78.6±0.41	67.84±1.62	59.61±14.74	21.62±0.58	24.2±2.91	5.82±0.58	6.66±3.4	1.55±4.4	26.6±0.38	24.43±0.37	44.55±0.08	31.08±0.89	19.88±0.74	39.09±2.89
	Ensemble	**PE**	**PU**	**Patch**	79.36±0.13	65.34±0.65	91.31±0.14	19.12±0.1	24.77±2.49	4.41±0.06	6.98±2.71	2.15±2.77	22.93±0.44	21.13±0.25	42.85±0.15	32.62±0.58	20.51±0.28	36.79±1.42
				Thresh	79.36±0.13	65.34±0.65	72.07±0.52	19.75±0.17	28.75±2.94	5.04±0.18	10.97±3.16	0.85±2.98	22.93±0.44	21.13±0.25	42.85±0.15	32.62±0.58	20.51±0.28	36.79±1.42
		EE	**AU**	**Patch**	79.36±0.13	65.34±0.65	89.39±0.45	19.56±0.14	37.06±2.1	4.85±0.11	19.28±2.33		29.48±0.47	26.4±0.83	44.23±0.39	35.38±0.73	20.51±0.28	36.79±1.42
				Thresh	79.36±0.13	65.34±0.65	53.44±2.59	22.37±0.29	52.64±1.34	7.66±0.32	34.85±1.41		29.48±0.47	26.4±0.83	44.23±0.39	35.38±0.73	20.51±0.28	36.79±1.42
		MI	**EU**	**Patch**	79.36±0.13	65.34±0.65	89.42±0.73	18.75±0.01	24.43±1.94	4.04±0.04	6.65±2.17	1.94±2.62	30.72±0.07	27.43±0.31	36.3±0.17	26.97±0.66	20.51±0.28	36.79±1.42
				Thresh	79.36±0.13	65.34±0.65	75.02±0.96	18.95±0.36	26.68±2.47	4.24±0.39	8.9±2.67	0.92±2.93	30.72±0.07	27.43±0.31	36.3±0.17	26.97±0.66	20.51±0.28	36.79±1.42
	TTA	**PE**	**PU**	**Patch**	79.08±0.12	65.2±3.68	88.41±1.02	19.24±0.14	28.94±7.15	4.39±0.23	11.07±6.19	0.17±8.74	25.13±0.34	21.54±0.25	40.89±0.61	30.39±2.45	21.85±0.26	42.45±4.5
				Thresh	79.08±0.12	65.2±3.68	66.79±2.58	20.34±0.74	35.83±7.94	5.5±0.42	17.97±7.13	0.92±8.45	25.13±0.34	21.54±0.25	40.89±0.61	30.39±2.45	21.85±0.26	42.45±4.5
		EE	**AU**	**Patch**	79.08±0.12	65.2±3.68	86.63±1.94	19.64±0.11	35.61±9.85	4.8±0.39	17.75±8.61		29.23±0.79	25.37±0.71	42.46±0.47	32.91±2.11	21.85±0.26	42.45±4.5
				Thresh	79.08±0.12	65.2±3.68	52.66±5.18	22.31±0.31	51.72±8.34	7.46±0.23	33.86±7.08		29.23±0.79	25.37±0.71	42.46±0.47	32.91±2.11	21.85±0.26	42.45±4.5
		MI	**EU**	**Patch**	79.08±0.12	65.2±3.68	86.32±2.14	18.75±0.19	28.06±7.07	3.9±0.32	10.19±6.25	0.87±8.57	30.86±0.3	26.51±0.73	34.99±1.08	25.15±2.15	21.85±0.26	42.45±4.5
				Thresh	79.08±0.12	65.2±3.68	70.06±3.16	19.26±0.76	31.25±7.73	4.42±0.38	13.39±6.89	0.49±9.23	30.86±0.3	26.51±0.73	34.99±1.08	25.15±2.15	21.85±0.26	42.45±4.5
	SSN	**PE**	**PU**	**Patch**	79.05±0.24	67.79±1.75	83.59±0.6	20.42±0.68	23.09±1.64	5.57±0.44	5.7±1.4	1.53±5.39	11.62±0.24	13.59±0.32	61.14±1.11	47.48±0.03	16.79±0.3	36.2±2.02
				Thresh	79.05±0.24	67.79±1.75	71.97±1.26	20.74±0.15	32.93±5.1	5.38±0.35	20.1±5.0	0.62±4.99	11.62±0.24	13.59±0.32	61.14±1.11	47.48±0.03	16.79±0.3	36.2±2.02
		MI	**AU**	**Patch**	79.05±0.24	67.79±1.75	71.68±3.1	20.23±0.98	26.71±2.76	5.38±0.79	9.32±2.43		30.31±0.67	25.93±0.37	55.42±1.71	44.16±0.26	16.79±0.3	36.2±2.02
				Thresh	79.05±0.24	67.79±1.75	56.56±8.8	20.63±1.3	31.75±3.27	5.72±1.15	14.36±3.14		30.31±0.67	25.93±0.37	55.42±1.71	44.16±0.26	16.79±0.3	36.2±2.02
		EE	**EU**	**Patch**	79.05±0.24	67.79±1.75	85.3±1.15	20.48±0.49	23.74±1.79	5.62±0.28	6.35±1.63	1.23±5.06	17.67±0.62	19.66±0.28	61.82±0.68	47.3±0.08	16.79±0.3	36.2±2.02
				Thresh	79.05±0.24	67.79±1.75	67.74±4.7	20.56±0.41	37.92±5.94	5.71±0.69	20.53±5.53	0.48±5.55	17.67±0.62	19.66±0.28	61.82±0.68	47.3±0.08	16.79±0.3	36.2±2.02

Table 6: **Evaluation of downstream tasks on the GTA 5 / Cityscapes dataset.** The table shows the segmentation performance by means of the Dice score and evaluation metrics for 5 different downstream tasks, where ↑ depict higher scores are better and ↓ lower scores are better. All scores are multiplyed by 10^2. The color heatmap is normalized per column, brighter columns imply better scores. For AL, the second cycle was only executed with EU and PU, indicated by empty grey entries for AU. Reported results show the mean and standard deviation over 3 different seeds. Abbreviations: PM: Prediction model, UM: Uncertainty measure, UT: Modeled uncertainty Type (according to theory), AGG: Aggregation strategy.

				Seg. Performance		OoD-D	Failure Detection				AL	Calibration		Ambiguity Modeling			
PM	UM	UT	AGG	Dice i.i.d. ↑	Dice OoD ↑	AUROC ↑	AURC i.i.d. ↓	AURC OoD ↓	EAURC i.i.d. ↓	EAURC OoD ↓	AL Improv. OoD ↑	ACE i.i.d. ↓	ACE OoD ↓	NCC i.i.d. ↑	NCC OoD ↑	GED i.i.d. ↓	GED OoD ↓
Determ.	**MSR**	**PU**	**Patch**	71.73±0.13	58.37±0.51	33.43±1.16	26.52±0.23	39.24±0.75	7.28±0.15	8.15±0.72	-0.5±1.78	13.09±0.15	17.21±0.31	50.61±0.43	46.81±0.61	36.09±0.26	58.87±0.83
			Image	71.73±0.13	58.37±0.51	69.99±0.58	24.2±0.17	36.85±0.33	4.96±0.04	5.76±0.24	1.44±1.73	13.09±0.15	17.21±0.31	50.61±0.43	46.81±0.61	36.09±0.26	58.87±0.83
TTD	**PE**	**PU**	**Patch**	71.63±0.14	58.62±0.91	36.67±1.42	26.77±0.2	39.46±0.57	7.43±0.06	8.43±0.35	-0.29±2.72	15.14±0.11	15.07±0.07	26.68±0.29	25.57±3.5	34.03±0.28	56.08±1.61
			Image	71.63±0.14	58.62±0.91	67.8±0.77	24.51±0.21	36.14±0.3	5.16±0.06	5.12±0.09	0.94±2.81	15.14±0.11	15.07±0.07	26.68±0.29	25.57±3.5	34.03±0.28	56.08±1.61
	EE	**AU**	**Patch**	71.63±0.14	58.62±0.91	36.83±1.35	26.76±0.21	39.45±0.61	7.41±0.07	8.43±0.38		15.08±0.11	15.18±0.08	27.67±0.29	26.39±3.46	34.03±0.28	56.08±1.61
			Image	71.63±0.14	58.62±0.91	68.03±0.75	24.5±0.21	36.15±0.3	5.15±0.06	5.13±0.09		15.08±0.11	15.18±0.08	27.67±0.29	26.39±3.46	34.03±0.28	56.08±1.61
	MI	**EU**	**Patch**	71.63±0.14	58.62±0.91	46.3±4.04	26.88±0.11	39.18±0.6	7.54±0.05	8.15±0.36	0.03±2.86	17.56±0.13	17.82±0.5	-23.17±0.19	-12.85±2.61	34.03±0.28	56.08±1.61
			Image	71.63±0.14	58.62±0.91	51.08±2.53	26.05±0.24	36.61±0.38	6.71±0.09	5.58±0.16	0.45±2.82	17.56±0.13	17.82±0.5	-23.17±0.19	-12.85±2.61	34.03±0.28	56.08±1.61
Ensemble	**PE**	**PU**	**Patch**	71.92±0.14	59.36±0.46	32.72±0.57	26.4±0.13	38.33±0.63	7.42±0.07	7.98±0.56	-0.65±1.67	15.64±0.12	15.61±0.22	23.77±0.11	25.37±0.98	31.92±0.22	50.12±0.36
			Image	71.92±0.14	59.36±0.46	72.06±0.8	24.02±0.12	35.51±0.83	5.05±0.01	5.16±0.17	1.17±1.73	15.64±0.12	15.61±0.22	23.77±0.11	25.37±0.98	31.92±0.22	50.12±0.36
	EE	**AU**	**Patch**	71.92±0.14	59.36±0.46	28.69±0.35	26.5±0.15	38.79±0.42	7.53±0.08	8.44±0.41		15.57±0.11	15.31±0.21	26.74±0.14	29.59±0.9	31.92±0.22	50.12±0.36
			Image	71.92±0.14	59.36±0.46	66.81±0.94	24.05±0.11	35.54±0.78	5.08±0.01	5.19±0.22		15.57±0.11	15.31±0.21	26.74±0.14	29.59±0.9	31.92±0.22	50.12±0.36
	MI	**EU**	**Patch**	71.92±0.14	59.36±0.46	90.63±0.29	26.69±0.04	38.57±1.15	7.72±0.13	8.22±0.61	1.77±1.79	21.33±0.05	17.43±0.28	-22.55±0.08	-6.06±0.79	31.92±0.22	50.12±0.36
			Image	71.92±0.14	59.36±0.46	90.05±0.29	25.68±0.09	36.51±1.29	6.71±0.08	6.16±0.63	1.85±1.8	21.33±0.06	17.43±0.28	-22.55±0.08	-6.06±0.79	31.92±0.22	50.12±0.36
TTA	**PE**	**PU**	**Patch**	71.82±0.13	58.39±0.47	31.66±0.71	26.54±0.26	39.44±0.94	7.38±0.15	8.39±1.07	-0.61±1.71	15.37±0.14	15.42±0.31	24.84±0.23	28.14±1.28	33.52±0.26	54.06±0.72
			Image	71.82±0.13	58.39±0.47	69.81±0.32	24.27±0.18	36.44±0.16	5.11±0.05	5.39±0.05	1.44±1.78	15.37±0.14	15.42±0.31	24.84±0.23	28.14±1.28	33.52±0.26	54.06±0.72
	EE	**AU**	**Patch**	71.82±0.13	58.39±0.47	28.85±0.8	26.55±0.27	39.39±0.98	7.39±0.17	8.35±1.12		15.3±0.14	15.32±0.33	26.29±0.24	30.47±1.19	33.52±0.26	54.06±0.72
			Image	71.82±0.13	58.39±0.47	66.61±0.31	24.28±0.18	36.42±0.16	5.12±0.05	5.37±0.05		15.3±0.14	15.32±0.33	26.29±0.24	30.47±1.19	33.52±0.26	54.06±0.72
	MI	**EU**	**Patch**	71.82±0.13	58.39±0.47	92.75±0.11	27.1±0.09	39.65±0.47	7.95±0.08	8.61±0.41	2.12±1.77	22.04±0.39	19.1±0.32	-16.25±0.14	-4.45±1.2	33.52±0.26	54.06±0.72
			Image	71.82±0.13	58.39±0.47	94.32±0.33	26.12±0.05	37.57±0.41	6.96±0.08	6.53±0.54	2.46±1.77	22.04±0.39	19.1±0.32	-16.25±0.14	-4.45±1.2	33.52±0.26	54.06±0.72
SSN	**PE**	**PU**	**Patch**	65.49±0.13	51.9±0.44	77.95±1.28	33.85±0.35	46.53±0.89	10.14±0.29	9.91±0.55	2.01±2.21	16.1±0.13	17.84±0.41	55.97±0.09	37.26±0.95	25.6±0.03	45.26±1.1
			Image	65.49±0.13	51.9±0.44	82.0±1.1	30.07±0.05	43.05±0.95	6.36±0.14	6.43±0.41	3.21±2.25	16.1±0.13	17.84±0.41	55.97±0.09	37.26±0.95	25.6±0.03	45.26±1.1
	MI	**AU**	**Patch**	65.49±0.13	51.9±0.44	53.14±3.37	32.57±0.31	46.28±1.29	8.86±0.4	9.65±1.0		21.6±0.14	25.8±0.42	70.34±0.36	51.8±0.44	25.6±0.03	45.26±1.1
			Image	65.49±0.13	51.9±0.44	55.21±1.37	31.0±0.09	46.07±0.67	7.29±0.13	9.44±0.51		21.6±0.14	25.8±0.42	70.34±0.36	51.8±0.44	25.6±0.03	45.26±1.1
	EE	**EU**	**Patch**	65.49±0.13	51.9±0.44	78.12±1.74	33.87±0.42	46.86±0.86	10.16±0.37	10.23±0.39	2.44±2.17	10.49±0.31	11.12±0.45	4.57±1.07	2.95±0.9	25.6±0.03	45.26±1.1
			Image	65.49±0.13	51.9±0.44	86.2±1.08	32.58±0.43	43.62±1.1	8.87±0.41	7.0±0.56	2.37±1.91	10.49±0.31	11.12±0.45	4.57±1.07	2.95±0.9	25.6±0.03	45.26±1.1