Improving Uncertainty Estimation with Confidence-aware Training Data — Supplementary Material —

Sergey Korchagin
IITP RAS, Russia
korchagin.sergey.97@gmail.com

Ekaterina Zaychenkova
IITP RAS, Russia
zaichenkova.ee@phystech.edu

Aleksei Khalin
IITP RAS, Russia
alexeyx0418@gmail.com

Aleksandr Yugay
Skoltech, Russia
aleksandr.yugay@skoltech.ru

Alexey Zaytsev
Skoltech, Sber, Russia
a.zaytsev@skoltech.ru

Egor Ershov
IITP RAS, AIRI, Russia
ershov@airi.net

A. Results of uncertainty estimation in semantic segmentation of lung CT scans for individual experts on LIDC dataset

Tab. S1 presents uncertainty estimation results via different methods for each individual expert on LIDC dataset.

UE method		AAC, $\times 10^{-4}$, $\downarrow$			
Epist.	Aleat.	Ex. 1	Ex. 2	Ex. 3	Ex. 4
MCMC [8]		62.6	40.3	82.2	88.7
ABNN [2]		134.9	164.2	143.7	137.5
Ens	–	32.2	29.8	26.8	28.7
Ens	Ens	**25.2**	26.1	**21.6**	**24.5**
Ens	CAE	30.0	**24.4**	29.0	25.2
Ens	Exps	**16.3**	**13.9**	**17.6**	**12.4**
UE method		TRD-95, %, $\downarrow$			
Epist.	Aleat.	Ex. 1	Ex. 2	Ex. 3	Ex. 4
MCMC [8]		2.30	1.75	1.98	2.25
ABNN [2]		5.93	5.79	5.73	5.45
Ens	–	2.26	1.55	1.67	1.88
Ens	Ens	**1.92**	**1.30**	**1.45**	**1.67**
Ens	CAE	2.57	1.77	1.86	1.94
Ens	Exps	**1.47**	**1.15**	**1.30**	**1.28**

Table S1. Quantitative comparison of uncertainty estimation methods for semantic segmentation. AAC stands for Area Above the rejection Curve, normalized with respect to the oracle curve. TRD-95 stands for Throwaway Rate required to attain Dice of 95%. Two best results are in bold, and the best result is underlined for each expert.

As mentioned before, using CAE for aleatoric uncertainty estimation only outperforms ensemble total variance for Expert 2, decreasing AAC by 6.5%. For Expert 4 results for two methods are comparable, and for Experts 1 and 3 CAE performs worse.

Other trends are common for all experts. Ensemble methods outperform MCMC and ABNN, which is expected. ABNN models are the easiest to train (of methods studied), but perform comparatively worse. MCMC provides adequate uncertainty estimation requiring less computational resources than ensembles. Still, for Experts 3 and 4 Total Variance-based methods provide AAC from three to four times smaller than MCMC, respectively, which is significant.

B. Uncertainty estimation in multi-class segmentation of retinal fundus images on RIGA dataset

To further validate the proposed approach to uncertainty estimation in segmentation, we conducted additional experiments on retinal fundus images for glaucoma analysis (RIGA) task presented in [1]. RIGA dataset contains images of retinal fundus annotated by six experienced ophthalmologists. The data is split into three subsets named "MESSIDOR", "Bin Rushed" and "Magrabi" containing 460, 195 and 95 images respectively for a total of 750 images. Each image is of size 256×256 pixels and is supplied with six masks of an optic disc and an optic cup, one for each expert. Ophthalmologists often use the vertical and horizontal cup-to-disk ratios as well as the disk and cup area ratios for diagnosing glaucoma.

In our experiments we mostly followed the setup described in the main paper with the following notable exceptions. To obtain expert labels for multiclass segmentation we first one-hot encoded each pixel with a vector $\tilde{y} \in \mathbb{R}^3$ where $\tilde{y}_i = 1$ if the pixel belongs to class i and 0 otherwise. In our experiments, classes of background,

UE method		AAC, $\times 10^{-3}$, ↓						TRD-95, %, ↓					
Epist.	Aleat.	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6
MCMC [8]		19.72	63.84	36.21	30.61	24.37	31.30	9.30	17.90	7.79	8.03	6.75	7.49
ABNN [2]		74.96	92.45	88.09	51.09	55.45	67.85	13.00	92.70	11.70	12.10	10.80	10.90
Ens	–	10.96	15.92	**10.37**	9.33	12.82	11.30	7.60	11.70	**6.53**	6.53	5.87	6.62
Ens	Ens	12.35	17.54	12.80	12.55	15.81	13.13	7.04	10.80	**6.17**	**6.14**	**5.43**	**6.06**
Ens	CAE	**6.81**	**7.52**	20.49	**5.81**	**4.81**	**6.13**	**5.98**	**8.14**	6.61	**6.31**	**5.25**	**5.63**
Ens	Exps	**7.33**	**9.30**	**8.12**	**9.08**	**5.87**	**6.78**	**6.96**	**9.61**	6.76	7.56	6.04	6.75

Table S2. Quantitative comparison of uncertainty estimation methods for optic disc segmentation. AAC stands for Area Above the rejection Curve, normalized with respect to the oracle curve. TRD-95 stands for Throwaway Rate required to attain Dice of 95%. Two best results are in bold, and the best result is underlined for each expert.

UE method		AAC, $\times 10^{-3}$, ↓						TRD-90, %, ↓					
Epist.	Aleat.	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6
MCMC [8]		54.71	64.23	74.31	31.32	79.63	60.82	11.60	54.20	90.10	5.51	98.80	91.10
ABNN [2]		89.56	91.47	175.62	78.31	246.88	189.23	16.00	27.70	29.80	7.36	31.20	23.10
Ens	–	47.75	49.52	203.97	24.45	167.28	164.75	7.82	13.00	11.00	**4.05**	98.80	97.60
Ens	Ens	40.06	52.99	86.33	**21.74**	135.41	107.25	**7.46**	12.00	12.20	**4.31**	96.40	**14.30**
Ens	CAE	**10.69**	**13.38**	**12.96**	**11.56**	**13.78**	**15.76**	**6.05**	**7.92**	**7.60**	4.93	**8.06**	**8.91**
Ens	Exps	**36.64**	**34.26**	**53.55**	50.19	**63.92**	**59.41**	8.17	**11.80**	**10.90**	6.85	**15.10**	14.80

Table S3. Quantitative comparison of uncertainty estimation methods for optic cup segmentation. AAC stands for Area Above the rejection Curve, normalized with respect to the oracle curve. TRD-90 stands for Throwaway Rate required to attain Dice of 90%. Two best results are in bold, and the best result is underlined for each expert.

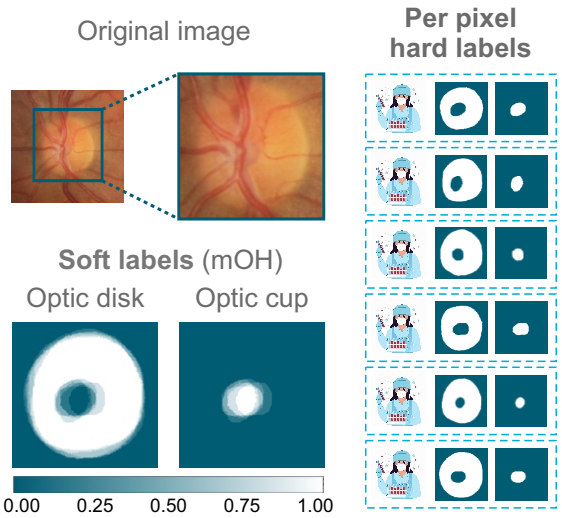

Figure S1. Structure of the RIGA dataset. Each sample is annotated by six experts with two maps for optic disc and cup classes. These maps are averaged by expert to obtain soft labels.

optic disc and optic cup correspond to classes 1, 2 and 3 respectively. Then, we simply averaged one-hot vectors obtained from different experts for each pixel. In other words, $\tilde{y}_{(x_1,x_2)} = \frac{1}{6}\sum_{e=1}^{6} \tilde{y}_{e,(x_1,x_2)} \in \mathbb{R}^3$, where $\tilde{y}_{e,(x_1,x_2)}$ is a one-hot vector for pixel (x_1, x_2) obtained from expert e. Model is trained to predict $\tilde{y}_{(x_1,x_2)}$.

We used "MESSIDOR" and "Magrabi" subsets of data for training and "Bin Rushed" for testing, resulting in an approximately 3:1 division. The training part was further split in 3:1 proportions, and the latter part used for validation.

Uncertainty from multiple classes was combined as described in [6]. To evaluate methods, we plotted rejections curves for both optic dick and cup classes and analyzed them separately. We also computed the throwaway rate to attain Dice of 95% for optic disc segmentation and the throwaway rate to attain Dice of 90% for optic cup segmentation, denoted TRD-95 and TRD-90 respectively.

Obtained rejection curves are presented in Fig. S2 for optic disc segmentation and Fig. S3 for optic cup segmentation. Tab. S2 and Tab. S3 contain results for optic disc and cup segmentation respectively.

The proposed method of combining ensemble epistemic uncertainty and CAE aleatoric uncertainty consistently outperforms other methods, including the method that incorporates ground truth expert annotations, in AAC, which is

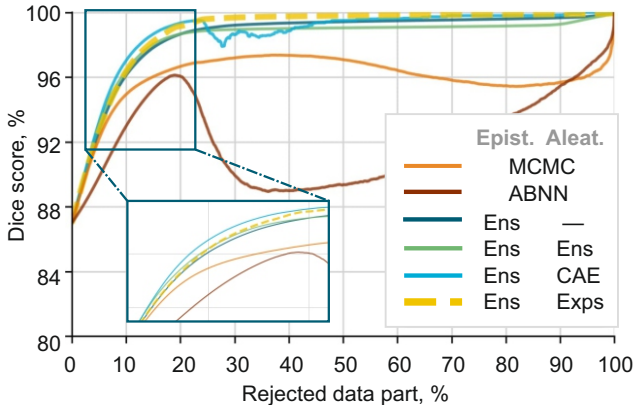

Figure S2. Averaged rejection curves for optic disc segmentation.

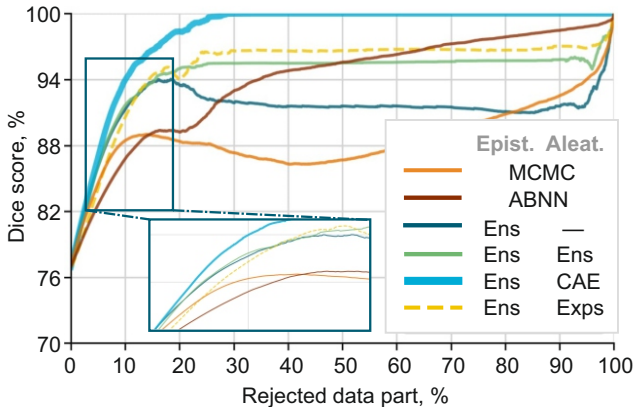

Figure S3. Averaged rejection curves for optic cup segmentation.

around twice as small for the optic disk class and from two to almost five times as small for the optic cup class compared to the next best method for any given expert. TRD-95 for the proposed approach is also the best in most cases. This is in contrast to the results achieved on the LIDC task. We theorize that this is due to the fact that expert annotations are much better aligned on the RIGA task since there are no images where a prediction from any expert is absent.

Using ensemble total variance, compared to using just the epistemic component, generally performs better for optic cup segmentation and slightly worse for the optic disc task. One possible explanation would be that expert assessments for optic disc areas are better aligned than those for optic cup, which in turn provides less diverse CAE predictions.

MCMC provides results comparable to those of ensemble variance methods in optic cup segmentation, but struggles in the optic disc task. ABNN performs close to other methods only in few cases and, in terms of quality, is outperformed by the proposed solution.

C. Experimental setup for baseline training

Classification

The final uncertainty for MCMC [8] and MC Dropout [3] methods is calculated as a variance over ensemble outputs.

Each MCMC ensemble comprises 10 neural networks created by consecutive check-pointing the model at 15-epoch intervals during training. For optimization purposes Stochastic Gradient Langevin Dynamics (SGLD) was used. As a starting point for training, weights of fully trained classification networks were chosen resulting in 10 MCMC networks for each of the 10 classification networks.

MC Dropout builds an ensemble for each classification net by activating dropout layers during testing with 10 random initializations. The dropout probability for inference was set to 0.2, the same as for training.

The final uncertainty for HUQ [7] method was calculated based on DDU [5] epistemic and aleatoric uncertainty decomposition. Epistemic uncertainty for each sample was calculated as density of 10-component PCA of networks' embedding after convolutional layers. Aleatoric uncertainty was taken as cross entropy of the sample. Ranging function was trained on 10% validation set with hyperparameter α equal to 0.5 as if samples were out of distribution.

Segmentation

MCMC was trained in a similar manner to classification but using hyperparameters specified for segmentation.

For ABNN [2], code publicly available at `https://github.com/abtinmU/MakeMe-BNN` was adapted for semantic segmentation. Models were fine-tuned for 50 epochs, and during inference, 5 parameter samples were used for uncertainty estimation for each data sample. A total of 10 models were trained, and results averaged for stability.

D. On calibration and train data selection in presence of soft labels

In our main work we proposed using soft labels to train better-calibrated neural networks and ensembles. A logical question then is whether or not these labels are essential, or other methods of calibration would provide comparable results. To answer this, we also incorporated a common approach for model calibration called calibrators.

Formally, a calibrator is a map $\mu : [0, 1] \to [0, 1]$ trained to predict $\mu(s) = \mathbb{E}[Y \mid f(X) = s]$ for some model $f(\cdot)$, where random variables X, Y denote respectively the features and label of a uniformly randomly drawn instance from the dataset. Calibrators are normally trained on a validation subset of data, unseen by the model during training. In our experiments, we utilized isotonic, logistic and Beta-calibration [4] for models trained on hard labels. Calibra-

Calibration method	ECE, ↓
No calibration	0.0245 ± 0.0048
Isotonic	0.0264 ± 0.0051
Logistic	0.0257 ± 0.0038
Beta [4]	$\mathbf{0.0241 \pm 0.0048}$

Table S4. Results of calibrator application for classification models. Best result is in bold.

tion results are presented in Tab. S4.

Surprisingly enough, calibrators fail to provide significant improvement to ECE (with two out of three methods making it worse). To explain this phenomenon, it was noticed that soft labels are distributed differently on train and test parts of data. In order for calibration to work, train and test domains should be similar. We then theorized that expert assessments can help guide train data selection to produce models with better calibration even without using their labels directly for training.

To create an expert-guided train/test split we asigned each sample a value from 0 to 4 based on its expert assessments y_e according to formula

$$\tilde{y} = \text{round}\left[\frac{4}{6}\sum_{e=1}^{6} y_e\right]. \tag{1}$$

Samples were then grouped by their binary label $y \in \{0, 1\}$ and their expert value $\tilde{y} \in \{0, 1, 2, 3, 4\}$, with a total of 10 groups forming. The data was split in a way that these groups are included in equal proportions in train and test subsets. After that new models and calibrators were trained on this data using only hard labels. Calibration results are presented in Tab. S5.

Calibration method	ECE, ↓
No calibration	0.0167 ± 0.0055
Isotonic	0.0141 ± 0.0046
Logistic	0.0125 ± 0.0058
Beta [4]	$\mathbf{0.0121 \pm 0.0047}$

Table S5. Results of calibrator application for classification models on expert-guided data splits. Best result is in bold.

Not only did the calibration error of models without calibrators decrease by 32.9% (compared to random data splits), but using expert-guided data splits also enabled calibrator training to provide meaningful results. The best calibration method is shown to be Beta-calibration, which reduces ECE by 27.2%. Using both expert-guided splits and Beta-calibration provides a **twofold** improvement. This demonstrates that utilizing expert assessments can help train better-calibrated models even when not used directly.

E. Sensitivity analysis of experts' markup

Classification dataset contains confidence assessment of six experts for each sample. In order to evaluate their performance we calculated UE by each possible combination E of N experts where N varies from one up to six:

$$UE = \frac{1}{|E|}\sum_{i \in E} p_{\theta_i}(x)\left(1 - p_{\theta_i}(x)\right), \tag{2}$$

In the Tab. S6 top-5 best combinations of experts are presented sorted by AAC. The values of AAC were averaged over 10 classification networks.

Combination, E	AAC, $\times 10^{-4}$, ↓
{0;2;4;5}	65
{0;2;3;4;5}	65
{0;2;5}	66
{0;1;2;3;4;5}	66
{0;2;3;5}	67

Table S6. AAC for UE by different combinations of experts. Each number in brackets refers to each individual expert.

Each combination shows little difference in quality of UE between each other. With the correct set of experts one will be able to obtain the same quality of uncertainty estimation with only **three** experts instead of **six**. This observation can significantly simplify markup process of the new dataset.

References

[1] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the riga dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, pages 55–62. SPIE, 2018. 1

[2] Gianni Franchi, Olivier Laurent, Maxence Leguéry, Andrei Bursuc, Andrea Pilzer, and Angela Yao. Make me a bnn: A simple strategy for estimating bayesian uncertainty from pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12194–12204, 2024. 1, 2, 3

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 3

[4] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017. 3, 4

[5] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty:

A new simple baseline. In *Proceedings of the IEEE/CVF CVPR*, pages 24384–24394, 2023. 3

[6] Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise aleatoric and epistemic uncertainty quantification. *arXiv preprint arXiv:2406.02354*, 2024. 2

[7] Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, 2023. 3

[8] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, pages 681–688, 2011. 1, 2, 3