

Layer Ensembles: A Single-Pass Uncertainty Estimation in Deep Learning for Segmentation

Kaisar Kushibar^{*1}, Víctor Manuel Campello¹, Lidia Garrucho Moras¹,
Akis Linardos¹, Petia Radeva¹, and Karim Lekadir¹

¹*University of Barcelona, Department of Mathematics and Computer Science, Barcelona, 08007, Spain.*

Abstract

Uncertainty estimation in deep learning has become a leading research field in medical image analysis due to the need for safe utilisation of AI algorithms in clinical practice. Most approaches for uncertainty estimation require sampling the network weights multiple times during testing or training multiple networks. This leads to higher training and testing costs in terms of time and computational resources. In this paper, we propose Layer Ensembles, a novel uncertainty estimation method that uses a single network and requires only a single pass to estimate predictive uncertainty of a network. Moreover, we introduce an image-level uncertainty metric, which is more beneficial for segmentation tasks compared to the commonly used pixel-wise metrics such as entropy and variance. We evaluate our approach on 2D and 3D, binary and multi-class medical image segmentation tasks. Our method shows competitive results with state-of-the-art Deep Ensembles, requiring only a single network and a single pass.

1 Introduction

Despite the success of Deep Learning (DL) methods in medical image analysis, their black-box nature makes it more challenging to gain trust from both clinicians and patients [26]. Modern DL approaches are unreliable when encountered with new situations, where a DL model silently fails or produces an overconfident wrong prediction.

^{*}Corresponding author: kaisar.kushibar@ub.edu

Uncertainty estimation can overcome these common pitfalls, increasing the reliability of models by assessing the certainty of their prediction and alerting their user about potentially erroneous reports.

Several methods in the literature address uncertainty estimation in DL [1]. General approaches include: 1) Monte-Carlo Dropout (MCDropout) [7], which requires several forward passes with enabled dropout layers in the network during test time; 2) Bayesian Neural Networks (BNN) [6] that directly represent network weights as probability distributions; and 3) Deep Ensembles (DE) [13] which combines the outputs of several networks to produce uncertainty estimates. MCDropout and BNN have been argued to be unreliable in real world datasets [14]. Nonetheless, MCDropout is one of the most commonly used methods, often favourable when carefully tuned [8]. Despite their inefficiency in terms of memory and time, evidence showed DE is the most reliable uncertainty estimation method [4, 1].

There have been successful attempts that minimised the cost of training of the original DE. For example, snapshot-ensembles [12] train a single network until convergence, and further train beyond that point, storing the model weights per additional epoch to obtain M different models. In doing so, the training time is reduced drastically. However, multiple networks need to be stored and the testing remains the same as in DE. Additionally, the deep sub-ensembles [23] method uses M segmentation heads on top of a single model. This method is particularly similar to our proposal – Layer Ensembles (LE). However, LE, in contrast to existing methods, exhibits the following benefits:

- Scalable, intuitive, simple to train and test. The number of additional parameters is small compared to BNN approaches that double the number of parameters;
- Single network compared to the state-of-the-art DE;
- Unlike multi-pass BNN and MCDropout approaches, uncertainties can be calculated using a single forward pass, which would greatly benefit real-time applications;
- Produces global (image-level) as well as pixel-wise uncertainty measures;
- Allows estimating difficulty of a target sample for segmentation (example difficulty) that could be used to detect outliers;
- Similar performance to DE regarding accuracy and confidence calibration.

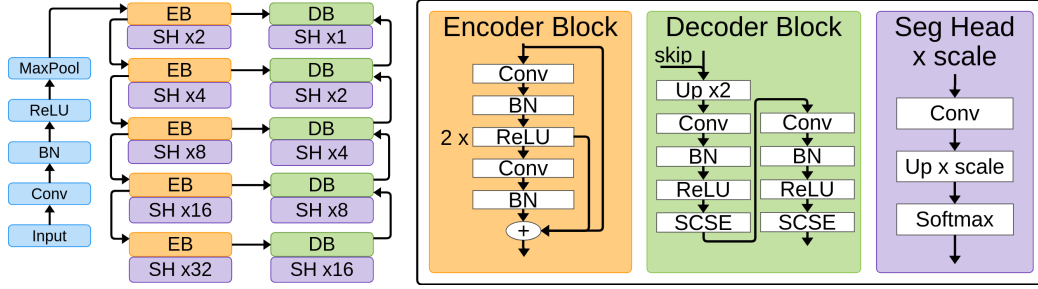


Figure 1: LE built on top of U-Net like architecture. Encoder Block (EB) in orange and Decoder Block (DB) in green have internal structures as depicted in the boxes below with corresponding colours. Ten Segmentation Heads (SH) are attached after each layer output with an up-scaling factor depending on the depth of the layer. SCSE - Squeeze and Excitation attention module. BN - Batch Normalisation.

2 Methodology

Our method is inspired by the state-of-the-art DE [13] for uncertainty estimation as well as a more recent work [3] that estimates example difficulty through prediction depth. In this section, we provide a detailed explanation of how our LE method differs from other works [13, 3], taking the best from both concepts and introducing a novel method for uncertainty estimation in DL. Furthermore, we introduce how LE can be used to obtain a single image-level uncertainty metric that is more useful for segmentation tasks compared to the commonly used pixel-wise variance, entropy, and mutual information (MI) metrics.

2.1 Prediction depth

Prediction Depth (PD) [3] measures example difficulty by training k-NN classifiers using feature maps after each layer. Given a network with N layers, the PD for an input image x is $L \sim [0, N]$ if the k-NN prediction for the L^{th} layer is different to layer at $L - 1$ and the same for all posterior layer predictions. The authors demonstrated that easy samples have small PD, whereas difficult ones have high PD by linking the known phenomena in DL that early layers converge faster [15] and networks learn easy data first [22]. Using PD for estimating example difficulty is appealing, however, it requires training additional classifiers on top of a pre-trained network. Moreover, using the traditional Machine Learning classifiers (e.g. k-NN) for a segmentation task is not trivial.

We extend the idea of PD to a more efficient segmentation method. Instead of k-NN classifiers, we attach a segmentation head after each layer output in the network

as shown in Figure 1. We use a CNN following the U-Net [20] architecture with different modules in the decoder and encoder blocks. Specifically, we use residual connections [10] in the encoder and squeeze-and-excite attention [11] modules in the decoder blocks. Our approach is architecture agnostic and the choice of U-Net was due to its wide use and high performance on different medical image segmentation tasks.

2.2 Ensembles of networks of different depths

DE has been used widely in the literature for predictive uncertainty estimation. The original method assumes a collection of M networks with different initialisation trained with the same data. Then, the outputs of each of these M models can be used to extract uncertainty measurements (e.g. variance). As we have shown in Figure 1, ten segmentation heads were added after each layer. Then, LE is a compound of M sub-networks of different depths. Since each of the segmentation heads is randomly initialised, it is sufficient to cause each of the sub-networks to make partially independent errors [9]. The outputs from each of the segmentation heads can then be combined to produce final segmentation and estimate the uncertainties, similarly to DE. Hence, LE can be considered equivalent to DE, but using only one network model.

2.3 Layer agreement as an image-level uncertainty metric

As we have stated above, LE is a combination of sub-networks of different depths. It can also be viewed as stacked networks where the parameters of a network f_t is shared by f_{t+1} for all $t \in [0, N)$, where N is the total number of outputs. This sequential connection of N sub-networks allows us to observe the progression of segmentation through the outputs of each segmentation head. We can measure the agreement between the adjacent layer outputs – e.g. using the Dice coefficient – to obtain a layer agreement curve. Depending on the network uncertainty, the agreement between layers will be low, especially in the early layers (Figure 2). We propose to use the Area Under Layer Agreement curve (AULA) as an image-level uncertainty metric. In the following sections, we demonstrate that AULA is a good uncertainty measure to detect poor segmentation quality, both in binary and multi-class problems.

3 Materials and Implementation

We evaluate our proposal on two active medical image segmentation tasks: 1) Breast mass for binary 2D; and 2) Cardiac MRI for multi-class 3D segmentation. We assess

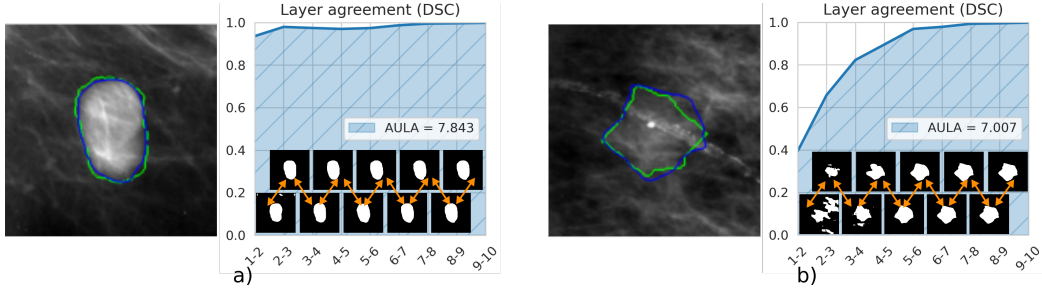


Figure 2: Layer Agreement curve. a) A high contrast lesion: large AULA and low uncertainty. b) A low contrast lesion and calcification pathology is present: small AULA and higher uncertainty. Arrows represent the correspondence between layers 1 and 2, 2 and 3, etc. DSC – Dice Similarity Coefficient. Green contours are ground truths.

LE in terms of segmentation accuracy, segmentation quality control using AULA metric, and example difficulty estimation using PD. For all the experiments, except for the example difficulty, LE is compared against the state-of-the-art DE approach and also a plain network without uncertainty estimation (referred as Plain).

3.1 Datasets

We use two publicly available datasets for the selected segmentation problems. Breast Cancer Digital Repository (BCDR) [16] contains 886 MedioLateral Oblique (MLO) and CranioCaudal (CC) view mammogram images of 394 patients with manual segmentation masks for masses. Original images have a matrix size of 3328×4084 or 2560×3328 pixels (unknown resolution). We crop and re-sample all masses to patches of 256×256 pixels with masses centred in the middle, as done in common practice [19]. We randomly split the BCDR dataset into train (576), validation (134), and test (176) sets so that images from the same patient are always in the same set.

For cardiac segmentation, the M&Ms challenge (MnM) [5] dataset is utilised. We use the same split as in the original challenge – 175 training, 40 validation, and 160 testing. All the images come annotated at the End-Diastolic (ED) and End-Systolic (ES) phases for the Left Ventricle (LV), MYOcardium (MYO), and Right Ventricle (RV) heart structures in the short-axis view. In our experiments, both time-points are evaluated together. All MRI scans are kept in their original in-plane resolution varying from isotropic $0.85mm$ to $1.45mm$ and slice-thickness varying from $0.92mm$ to $10mm$. We crop the images to $128 \times 128 \times 10$ dimensions so that the heart structures are centred.

3.2 Training

The same training routine is used for all the experiments, with only exception in batch-size: 10 for breast mass and 1 for cardiac structure segmentation. The network is trained for 200 epochs using the Adam optimiser to minimise the generalised Dice [21] and Cross-Entropy (CE) losses for breast mass and cardiac structure segmentation, respectively. For the multi-class segmentation, CE is weighted by 0.1 for background and 0.3 for each cardiac structure. An initial learning rate of 0.001 is set with a decay by a factor of 0.5 when the validation loss reaches a plateau. Common data augmentations are applied including random flip, rotation, and random swap of mini-patches of size 10×10 . Images are normalised to have zero-mean and unit standard deviation. A single NVIDIA GeForce RTX 2080 GPU with 8GB of memory is used. The source code with dependencies, training, and evaluation is publicly available¹.

3.3 Evaluation

Testing is done using the weights that give the best validation loss during training. The final segmentation masks are obtained by averaging the outputs of individual networks in DE ($M = 5$). For LE, we tried both averaging the sub-network outputs and using the well-known Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm [24] that uses weighted voting. Both results were similar and for brevity we present only the version using STAPLE.

We evaluate LE and DE using the common uncertainty metrics in the literature – the pixel-wise variance, entropy, and MI, and they are summed for all pixels/voxels in cases where an image-level uncertainty is required. The AULA metric is used for LE. The network confidence calibration is evaluated using the Negative Log-Likelihood metric (NLL). It is a standard measure of a probabilistic model’s quality that penalises wrong predictions that have small uncertainty [18]. Note that AULA can also be calculated by skipping some of the initial segmentation heads.

4 Results

4.1 Segmentation performance and confidence calibration

Table 1 compares the segmentation performance of LE with DE and Plain models in terms of Dice Similarity Coefficient (DSC) and Modified Hausdorff Distance (MHD). Two-sided paired t-test is used to measure statistically significant differences. In

¹Github link will be presented soon.

Table 1: Segmentation and confidence calibration performance for Plain U-Net, DE, and LE on BCDR and MnM datasets. The values for DSC, MHD, and NLL are given as ‘mean(std)’. \uparrow - higher is better, \downarrow - lower is better. Best values are in bold. Statistically significant differences compared to LE are indicated by ‘*’.

BCDR – breast mass segmentation				MnM – all structures combined		
Method	DSC \uparrow	MHD \downarrow	NLL \downarrow	DSC \uparrow	MHD \downarrow	NLL \downarrow
Plain	*0.865(0.09)	*1.429(1.72)	*2.312(1.35)	0.900(0.11)	1.061(2.69)	0.182(0.41)
DE	0.870(0.09)	1.373(1.76)	*0.615(0.54)	*0.896(0.13)	1.465(4.86)	*0.157(0.33)
LE	0.872(0.084)	1.317(1.692)	0.306(0.25)	0.903(0.10)	1.302(5.31)	0.173(0.37)

MnM – Structure-wise DSC \uparrow			MnM – Structure-wise MHD \downarrow			
Method	LV	MYO	RV	LV	MYO	RV
Plain	0.882(0.13)	*0.804(0.12)	*0.826(0.15)	1.313(3.63)	1.303(2.79)	2.884(10.89)
DE	0.885(0.14)	*0.804(0.13)	0.829(0.16)	1.536(5.6)	1.500(3.98)	2.113(5.67)
LE	0.883(0.13)	0.809(0.11)	0.832(0.14)	1.525(5.86)	1.529(5.50)	2.525(8.92)

breast mass segmentation, LE performs similarly to DE and Plain model for both DSC and MHD metrics. The NLL of LE, however, is significantly better compared to others ($p < 0.001$). For cardiac structure segmentation, the combined DSCs for all methods are similar and MHD of Plain is slightly better. NLL of DE (0.157 ± 0.33) is significantly better than ours (0.173 ± 0.37) ($p < 0.05$), however, LE can achieve an NLL of 0.140 ± 0.23 by skipping less layers without compromising segmentation performance (see Figure 3, right). In our experiments, skipping the first three and five outputs gave the best results in terms of correlation between uncertainty metrics and segmentation performance for breast mass and cardiac structure tasks, respectively (see Table 2). Skipping all but the last segmentation head in LE is equivalent to the Plain network. In terms of structure-wise DSC, all methods are similar for all the structures. Plain method has a slightly better MHD compared to DE and LE for the LV and MYO structures ($p > 0.05$), and DE is better for the RV structure compared to LE $p > 0.05$.

The ranking across all metrics in Table 1 are: LE (1.58), DE (2.08), and Plain (2.33), showing that on average LE is better. Overall, segmentation performance of all three are similar and LE has a better confidence calibration.

4.2 Segmentation quality control

We evaluate our uncertainty estimation proposal for segmentation quality control, similarly to [17], and compare it to the state-of-the-art DE. We use the proposed AULA uncertainty metric to detect poor segmentation masks and the variance metric for DE. Figure 3 shows the fraction of remaining images with poor segmentation after a fraction of poor quality segmentation images are flagged for manual correction. We consider DSCs below 0.90 as poor quality for both segmentation tasks. We set the

Table 2: Spearman’s correlation of segmentation metrics with uncertainty metrics for breast mass and cardiac segmentation tasks. Absolute highest values are shown in bold.

	BCDR				MnM			
	Entropy	MI	Variance	AULA	Entropy	MI	Variance	AULA
DE-DSC	-0.783	-0.785	-0.785	N/A	-0.323	-0.433	-0.377	N/A
LE-DSC	0.615	0.597	0.620	0.785	0.221	0.207	0.203	0.649
DE-MHD	0.762	0.764	0.763	N/A	0.401	0.499	0.447	N/A
LE-MHD	-0.594	-0.575	-0.598	-0.730	-0.309	-0.313	-0.300	-0.571

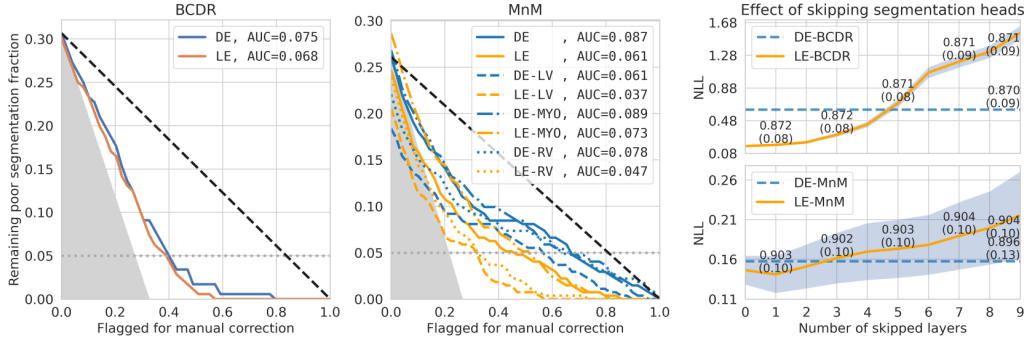


Figure 3: **Segmentation quality control** for DE and LE. The following are averaged indicators for: random flagging (dashed black); remaining 5% of poor segmentations (dotted grey); and ideal line (grey shaded area). **The effect of skipping initial segmentation head outputs** on model calibration. Numbers on top of the lines represent DSC in ‘mean(std)’ format. Shaded areas are standard deviations for NLL.

threshold for the cardiac structure following the inter-operator agreement identified in [2], and use the same value on the threshold for masses. As proposed in [17], the areas under these curves can be used to compare different methods. It can be seen that LE and DE are similar in terms of detecting poor quality segmentations, with LE achieving slightly better AUC for all the cases – mass, combined and structure-wise cardiac segmentations. Table 2 supports this statement by confirming high correlation between AULA and segmentation metrics. In BCDR, both are somewhat close to the averaged ideal line. For the cardiac structure segmentation, all the curves take a steep decline, initially being also close to the averaged ideal line indicating that severe cases are detected faster. Moreover, as can be seen in Figure 4, DE’s uncertainty maps are overconfident, while LE manages to highlight the difficult areas. We believe that having such meaningful heatmaps is more helpful for the clinicians (e.g. for manual correction). More visual examples including entropy and MI are given in Appendix.

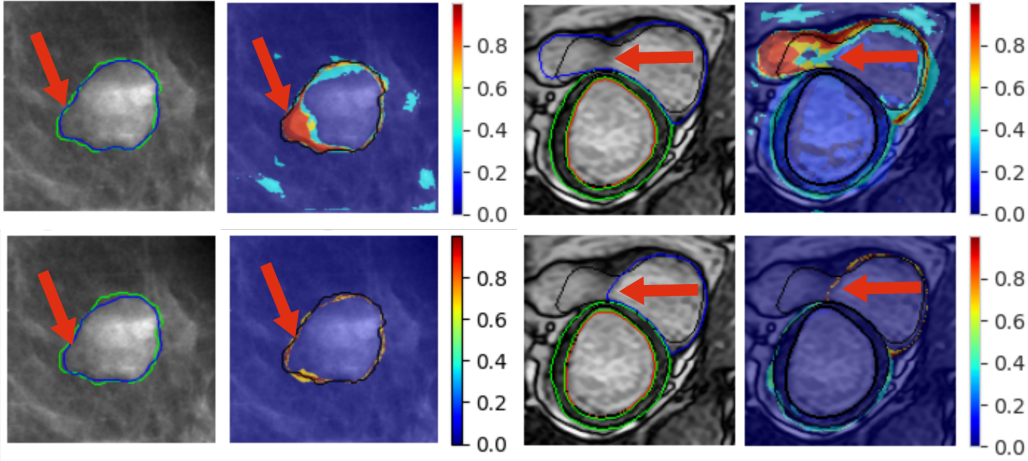


Figure 4: Examples of visual uncertainty heatmaps based on variance for high uncertainty areas (red arrows) using LE (top) and DE (bottom) for breast mass and cardiac structure segmentation. Black and green contours correspond to ground truth.

4.3 Example difficulty estimation

We evaluate example difficulty estimation using PD by perturbing the proportion of images in the test set. We added random Gaussian noise to MnM dataset and used Random Convolutions [25] in BCDR as the model was robust to noise in mammogram images. Examples of perturbed images are provided in Appendix. Then, for a given sample, PD is the largest L corresponding to one of the N segmentation heads in a network, where the agreement between L and $L - 1$ is smaller than a threshold that is the same as in segmentation quality control. In this sense, PD represents the minimum number of layers after which the network reaches to a consensus segmentation. Figure 5 shows how the distribution of PD shifts towards higher values as

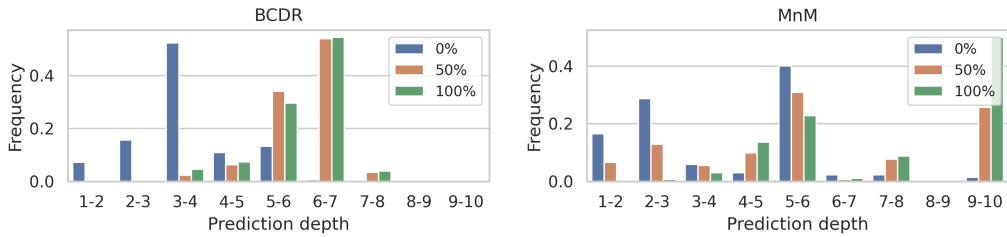


Figure 5: PD distribution with 0%, 50%, and 100% of the images corrupted by Gaussian noise – MnM $\mathcal{N}(0.3, 0.7)$, BCDR Random Convolutions with kernel-size (37, 37). Layer agreement threshold is 0.90 in terms of DSC for both datasets.

the number of corrupted images increases in the test set for both BCDR and MnM datasets. Overall, cardiac structure segmentation is more difficult than breast mass segmentation while the latter is more robust to noise. This demonstrates how PD can be used to evaluate example difficulty for the segmentation task and detect outliers.

5 Conclusions

We proposed a novel uncertainty estimation approach that exhibits competitive results to the state-of-the-art DE method using only a single network. Compared to DE, our approach produces a more meaningful uncertainty heatmaps and allows estimating example difficulty in a single pass. Experimental results showed the effectiveness of the proposed AULA metric to measure an image-level uncertainty measure. The capabilities of both AULA and PD were demonstrated in segmentation and image quality control experiments. We believe that the efficient and reliable uncertainty estimation that LE demonstrates will pave the way for more trustworthy DL applications in healthcare.

6 Acknowledgements

This study has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 952103.

References

- [1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* **76**, 243–297 (2021)
- [2] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance* **20**(1), 1–12 (2018)
- [3] Baldock, R., Maennel, H., Neyshabur, B.: Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems* **34** (2021)

- [4] Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9368–9377 (2018)
- [5] Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. *IEEE Transactions on Medical Imaging* **40**(12), 3543–3554 (2021)
- [6] Cinelli, L.P., Marins, M.A., Barros da Silva, E.A., Netto, S.L.: Bayesian neural networks. In: Variational Methods for Machine Learning with Applications to Deep Networks, pp. 65–109. Springer (2021)
- [7] Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. pp. 1050–1059. PMLR (2016)
- [8] Gal, Y., et al.: Uncertainty in deep learning (2016)
- [9] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), <http://www.deeplearningbook.org>
- [10] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
- [11] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
- [12] Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q.: Snapshot ensembles: Train 1, get M for free. In: International Conference on Learning Representations (2017), <https://openreview.net/forum?id=BJYwwY911>
- [13] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* **30** (2017)
- [14] Liu, Y., Pagliardini, M., Chavdarova, T., Stich, S.U.: The peril of popular deep learning uncertainty estimation methods. In: Bayesian Deep Learning (BDL) Workshop at NeurIPS 2021 (2021)

- [15] Morcos, A., Raghu, M., Bengio, S.: Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems* **31** (2018)
- [16] Moura, D.C., López, M.A.G., Cunha, P., Posada, N.G.d., Pollan, R.R., Ramos, I., Loureiro, J.P., Moreira, I.C., Araújo, B.M., Fernandes, T.C.: Benchmarking datasets for breast cancer computer-aided diagnosis (CADx). In: *Iberoamerican Congress on Pattern Recognition*. pp. 326–333. Springer (2013)
- [17] Ng, M., Guo, F., Biswas, L., Wright, G.A.: Estimating uncertainty in neural networks for segmentation quality control. In: *32nd Conf. Neural Inf. Process. Syst.(NIPS 2018)*, Montréal, Canada, no. Nips. pp. 3–6 (2018)
- [18] Quinonero-Candela, J., Rasmussen, C.E., Sinz, F., Bousquet, O., Schölkopf, B.: Evaluating predictive uncertainty challenge. In: *Machine Learning Challenges Workshop*. pp. 1–27. Springer (2005)
- [19] Rezaei, Z.: A review on image-based approaches for breast cancer detection, segmentation, and classification. *Expert Systems with Applications* **182**, 115204 (2021)
- [20] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
- [21] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 240–248. Springer (2017)
- [22] Toneva, M., Sordoni, A., des Combes, R.T., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. In: *International Conference on Learning Representations* (2019), <https://openreview.net/forum?id=BJlxm30cKm>
- [23] Valdenegro-Toro, M.: Deep sub-ensembles for fast uncertainty estimation in image classification. In: *Bayesian Deep Learning (BDL) Workshop at NeurIPS* (2019)
- [24] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging* **23**(7), 903–921 (2004)

- [25] Xu, Z., Liu, D., Yang, J., Raffel, C., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=BVSM0x3EDK6>
- [26] Young, A.T., Amara, D., Bhattacharya, A., Wei, M.L.: Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *The Lancet Digital Health* **3**(9), e599–e611 (2021)

Appendix

Corrupted image examples to increase prediction depth

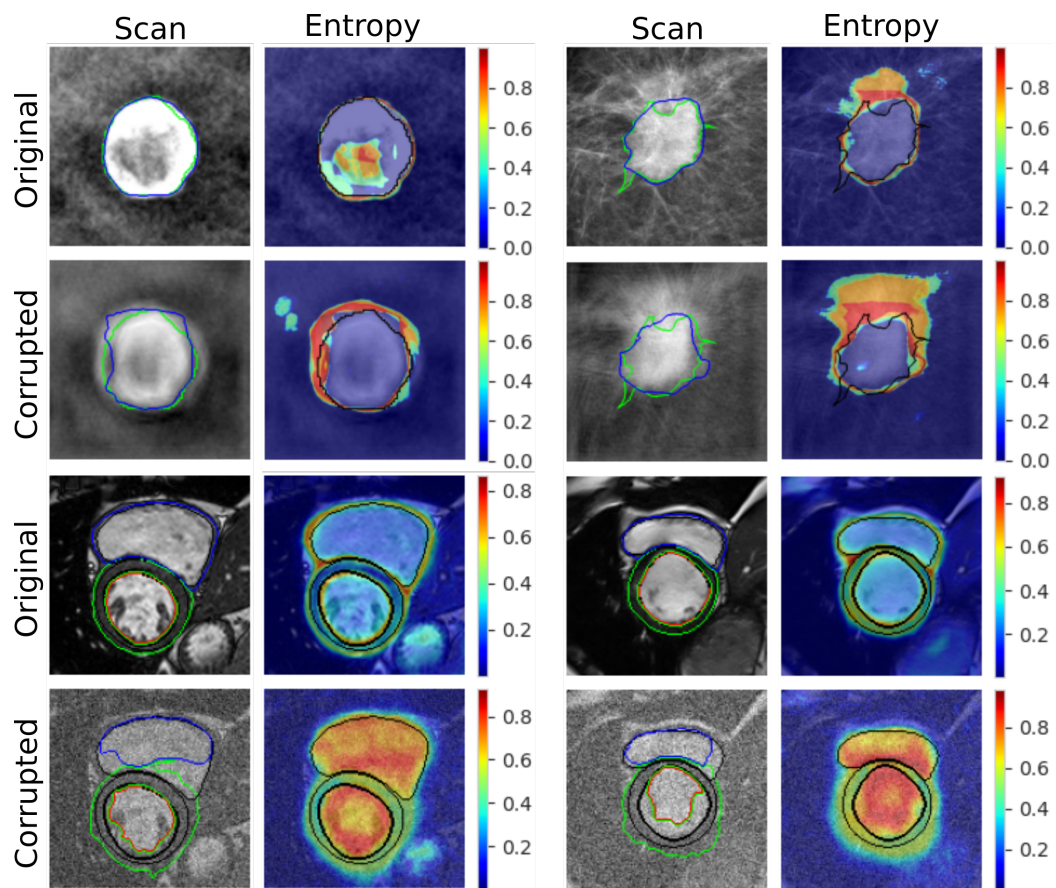


Figure 6: Examples of corrupted images and their effect on the entropy map. Black and green contours correspond to ground truth.

Qualitative examples

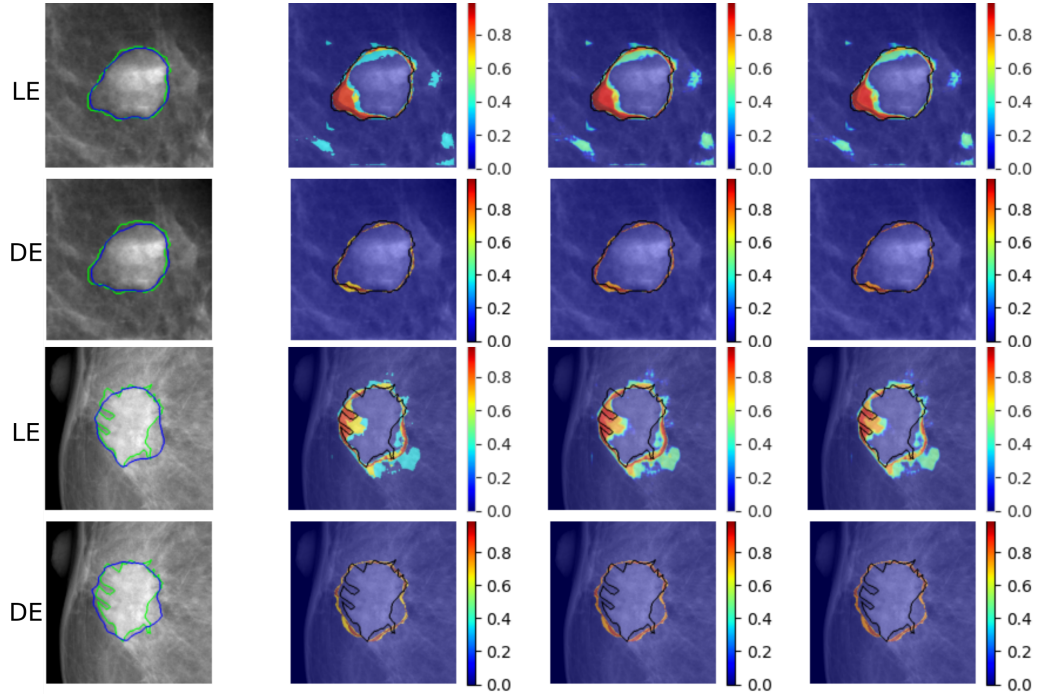


Figure 7: **BCDR**. Examples of visual uncertainty heatmaps based on variance, entropy, and mutual information. Black and green contours correspond to ground truth.

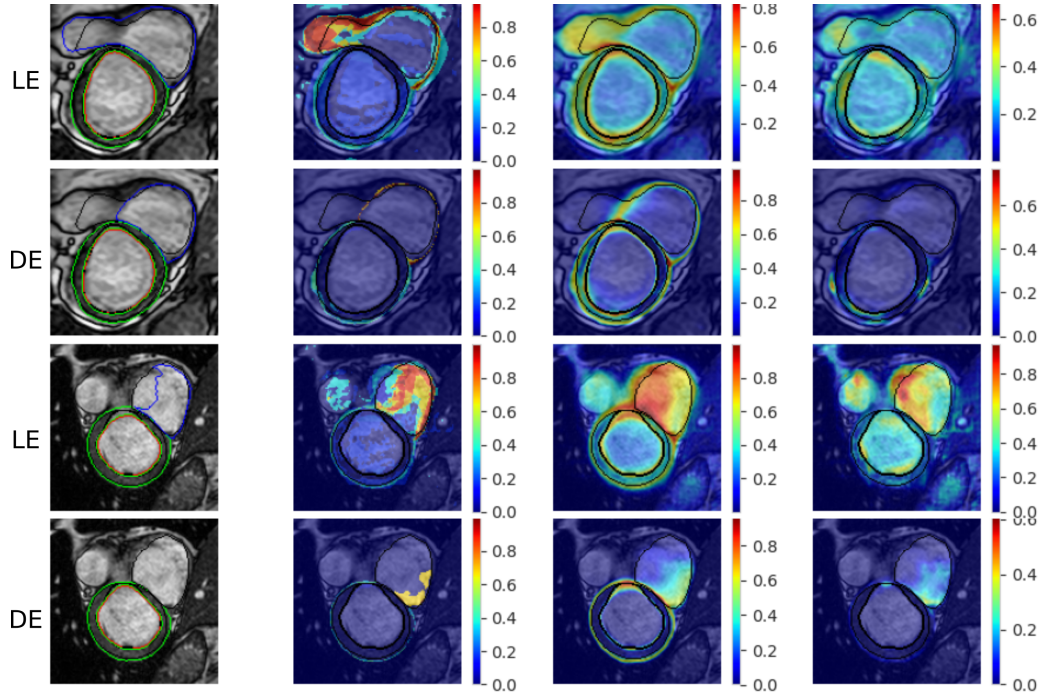


Figure 8: **MnM**. Examples of visual uncertainty heatmaps based on variance, entropy, and mutual information. Black contours correspond to ground truth.