Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation

Citation for published version (APA):

Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., Van Der Lugt, A., & De Bruijne, M. (2020). Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation. In C. H. Sudre, H. Fehri, T. Arbel, C. F. Baumgartner, A. Dalca, R. Tanno, K. van Leemput, W. M. Wells, A. Sotiras, B. Papiez, E. Ferrante, & S. Parisot (Eds.), *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. UNSURE 2020, GRAIL 2020* (pp. 32-41). Springer, Cham. https://doi.org/10.1007/978-3-030-60365-6_4

Document status and date:
Published: 05/10/2020

DOI:
10.1007/978-3-030-60365-6_4

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Quantitative Comparison of Monte-Carlo Dropout Uncertainty Measures for Multi-class Segmentation

Robin Camarasa[1,2](✉), Daniel Bos[2,3](✉), Jeroen Hendrikse[4](✉),
Paul Nederkoorn[5](✉), Eline Kooi[6](✉), Aad van der Lugt[2](✉),
and Marleen de Bruijne[1,2,7](✉)

[1] Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands
{r.camarasa,marleen.debruijne}@erasmusmc.nl
[2] Department of Radiology and Nuclear Medicine, Erasmus MC,
Rotterdam, The Netherlands
{d.bos,a.vanderlugt}@erasmusmc.nl
[3] Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
[4] Department of Radiology, University Medical Center Utrecht,
Utrecht, The Netherlands
j.hendrikse@umcutrecht.nl
[5] Department of Neurology, Academic Medical Center University of Amsterdam,
Amsterdam, The Netherlands
p.j.nederkoorn@amsterdamumc.nl
[6] Department of Radiology, Cardiovascular Research Institute Maastricht (CARIM),
Maastricht University Medical Center, Maastricht, The Netherlands
eline.kooi@mumc.nl
[7] Department of Computer Science, University of Copenhagen,
Copenhagen, Denmark

Abstract. Over the past decade, deep learning has become the gold standard for automatic medical image segmentation. Every segmentation task has an underlying uncertainty due to image resolution, annotation protocol, etc. Therefore, a number of methods and metrics have been proposed to quantify the uncertainty of neural networks mostly based on Bayesian deep learning, ensemble learning methods or output probability calibration. The aim of our research is to assess how reliable the different uncertainty metrics found in the literature are. We propose a quantitative and statistical comparison of uncertainty measures based on the relevance of the uncertainty map to predict misclassification. Four uncertainty metrics were compared over a set of 144 models. The application studied is the segmentation of the lumen and vessel wall of carotid arteries based on multiple sequences of magnetic resonance (MR) images in multi-center data.

C. H. Sudre et al. (Eds.): UNSURE 2020/GRAIL 2020, LNCS 12443, pp. 32–41, 2020.
https://doi.org/10.1007/978-3-030-60365-6_4

1 Introduction

Bayesian methods for neural networks [2,6,14] offer a mathematically grounded framework to analyse uncertainties. Nonetheless, the early Bayesian networks were computationally expensive to train, hard to implement and required more storage than conventional ones. The work of Gal et al. [3] renewed the interest in the field demonstrating the Bayesian properties of networks using dropout. The fast uptake of this technique in the field can be mainly attributed to the light alteration of the original model required.

The uncertainty estimation provided by Bayesian deep learning methods can be considered in every downstream tasks such as biomarkers extraction, surgery planning etc. Therefore, Bayesian techniques have known a rising interest in medical imaging for classification [12], segmentation [15,22] and registration [18].

Little research focuses on comparing the quality of different uncertainties metrics. A straightforward approach is to investigate the relationship between different uncertainty metrics and inter-observer variability [1,4]. Alternatively, in a classification problem, Van Molle et al. [22] introduces an uncertainty metric based on distribution similarity of the two most probable classes. Authors recommend the use of this uncertainty metric compared to variance based ones since it is more interpretable. In another work, Mehrtash et al. [9] compares calibrated and uncalibrated segmentation with negative log likelihood and Brier score. Finally, Nair et al. [13] compared the gain in segmentation performance when filtering out the most uncertain voxels for different uncertainty metrics. However, none of these approaches compare uncertainty metrics for multi-class segmentation which provide a larger spectrum of uncertainty measures.

To the best of our knowledge, this is the first work, in medical imaging, that compares quantitatively and statistically the ability of different uncertainty measures to predict misclassification in a multi-class segmentation context over a large set of models with widely varying performance, including different variations of Monte-Carlo dropout (MC dropout) techniques.

2 Methods

2.1 MC Dropout

In the following, $\theta \in \Omega$ represents the parameters of the model, f_θ the network with parameters θ, $(n_x, n_y, n_z) \in \mathbb{N}^3$ the dimensions of the input images, M the number of input modalities, C the number of output classes, $(x, t) \in \mathbb{R}^{n_x \times n_y \times n_z \times M} \times \mathbb{R}^{n_x \times n_y \times n_z \times C}$ a pair input image x with ground truth label image t, and $j \in J = \{0, ..., n_x - 1\} \times \{0, ..., n_y - 1\} \times \{0, ..., n_z - 1\}$ a 3D coordinate.

The different models used for carotid artery segmentation are based on the MC dropout method [3]. To obtain several estimates of the multi-class segmentation at test time, we sample T sets of parameters $(\theta_1, ..., \theta_T)$. From those parameters, we can evaluate T outputs $(f^{\theta_1}(x), ..., f^{\theta_T}(x))$ which represent a sample of the output distribution $q(y|x)$. From this sample, one can derive the mean and the covariance of output probabilities at a voxel level in Eq. 1.

$$\begin{cases} \mathbb{E}(q(y_j|x)) & \approx \frac{1}{T}\sum_{t=1}^{T} f_j^{\theta_t}(x) \\ \mathrm{Var}(q(y_j|x)) & \approx \frac{1}{T}\sum_{t=1}^{T} f_j^{\theta_t}(x)^T f_j^{\theta_t}(x) - \mathbb{E}(q(y_j|x))^T \mathbb{E}(q(y_j|x)) \end{cases} \tag{1}$$

An alternative to the original (Bernoulli) dropout that applies binary multiplicative noise is to use Gaussian multiplicative noise [20]. To make the two dropout methods comparable, one has to match the expected mean and the variance of the dropout distributions as shown in the following Eq. 2.

$$\begin{cases} B = \lambda A \\ \lambda_{Bernoulli} \sim \frac{1}{1-p}\mathcal{B}(1-p) \\ \lambda_{Gaussian} \sim \mathcal{N}(1, \frac{p}{1-p}) \end{cases} \tag{2}$$

where A is part of the feature maps of a dropout layer input, B is the corresponding feature map of that dropout layer output, $\lambda \in \mathbb{R}$ is randomly sampled from the dropout distribution, p is the dropout rate, $\mathcal{B}$ is a Bernoulli distribution and $\mathcal{N}$ is a Gaussian distribution.

2.2 Uncertainty Metrics

Distribution Description. A conventional approach to estimate uncertainty in a multi-class segmentation is to average the variance over classes [5,19]. In practice, this is obtained, at a voxel level, averaging the diagonal elements of the covariance matrix, Eq. 3.

$$u^v(q(y_j|x)) = \frac{1}{C}\mathrm{Tr}[\mathrm{Var}(q(y_j|x))] \tag{3}$$

where u^v is the averaged variance uncertainty metric and Tr is the trace of the matrix.

Another widely used uncertainty metric for segmentation is the entropy [23]. In contrast with the variance metric which can be directly computed from data sampled with MC dropout from the distribution $q(y_j|x)$, it requires the estimation of an integral defined in Eq. 4.

$$u^h(q(y_j|x)) = \frac{1}{C}\sum_{i=0}^{C-1}\int_0^1 -q_c(y_j = t|x)\log[q_c(y_j = t|x)]dt \tag{4}$$

where u^h is the averaged entropy uncertainty metric, $q_c(y_j|x)$ is the output distribution of the class c of the voxel j.

Distribution Similarity. Another option to define (voxelwise) classification uncertainty, is to consider the overlap of the distributions of the two most probable classes for a given voxel. Van der Molle et al. [22] considered the Bhattacharya coefficient, since it is interpretable (0: certain, 1: uncertain), Eq. 5.

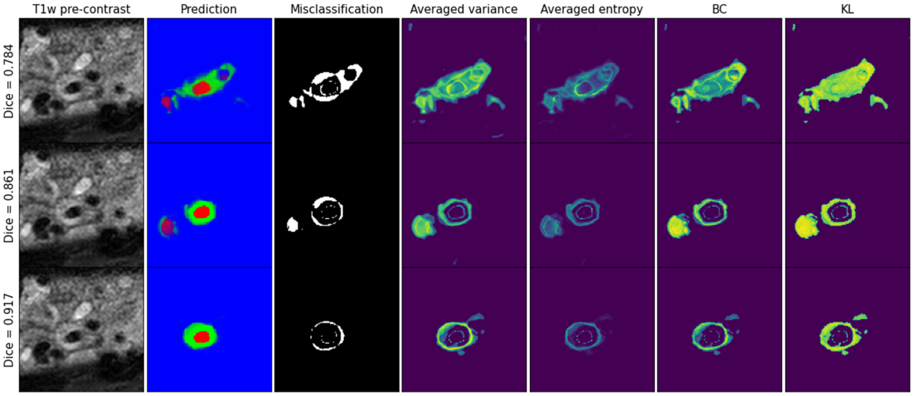

Fig. 1. Example of the different uncertainty metrics. From left to right columns represent, the T1w pre-contrast MR image, the multi-class prediction (blue = background, green = vessel wall, red = lumen, the level of brightness corresponds to the probability of the predicted class), the misclassification map and the different uncertainty maps (averaged variance, averaged entropy, BC, and KL). The rows correspond to predictions with different networks and level of performances. The indicated Dice is the averaged Dice over classes (Color figure online)

$$u^b(q(y_j|x)) = \int_0^1 \sqrt{q_{c_1}(y_j = t|x)q_{c_2}(y_j = t|x)}dt \tag{5}$$

where u^b is the Bhattacharya coefficient based uncertainty metric (BC), c_1 and c_2 are the two top classes for voxel j.

Alternatively, Kullback-Leibler divergence provides another measure of distribution similarity. However, unlike the previous presented uncertainty measures, a high value represents a small overlap among distributions. Therefore, the negative of the metric is considered. In addition, the Kullback-Leibler is made symmetric with respect to the classes c_1 and c_2, resulting in Eq. 6.

$$u^{kl}(q(y_j|x)) = -D_{KL}[q_{c_1}(y_j|x)||q_{c_2}(y_j|x)] - D_{KL}[q_{c_2}(y_j|x)||q_{c_1}(y_j|x)] \tag{6}$$

where u^{kl} is the Kullback-Leibler based uncertainty metric (KL) and D_{KL} is the Kullback-Leibler divergence.

The distribution $q(y_j|x)$ and $q_c(y_j|x)$ of Eqs. 3, 4, 5 and 6 are approximated by the distribution of the T outputs $(f^{\theta_1}(x), ..., f^{\theta_T}(x))$. The integrals of Eqs. 4, 5 and 6 are estimated with a left Riemann sum with a discretisation of the interval in n_{bins}.

2.3 Evaluation

Uncertainty Map Quality. To assess the quality of the uncertainty metrics, we applied the framework developed by Mobiny et al. [11] to the different type of uncertainty maps. The main idea is to consider uncertainty as a score that

Table 1. Uncertainty as a predictor of misclassification

	Uncertain ($u(q(y_j\|x)) \geq th$)	Certain ($u(q(y_j\|x)) < th$)
Misclassified	$TP(th)$	$FN(th)$
Correctly classified	$FP(th)$	$TN(th)$

predicts misclassification. From the MC estimates and the ground-truth, one can obtain a misclassification map m and an uncertainty map u (either u^v, u^h, u^b, u^{KL}) as described in Fig. 2. Once an uncertainty map is thresholded at a value th, one can define four types of voxels as summarized in Table 1: misclassified and uncertain (True Positive (TP) in a sense that the uncertainty of the voxel accurately predicts its misclassification), misclassified and certain (False Negative (FN)), correctly classified and uncertain (False Positive (FP)) and correctly classified and certain (True Negative (TN)).

For a given value of the uncertainty threshold th, it is possible to compute the precision and the recall of uncertainty as a misclassification predictor following: $\text{Pr}(th) = \frac{TP(th)}{TP(th)+FP(th)}$ and $\text{Rc}(th) = \frac{TP(th)}{TP(th)+FN(th)}$. By varying the threshold over the range of the values of u, one can derive the area under the precision recall curve (AUC-PR), Eq. 7.

$$\text{AUC-PR} = \sum_{i=1}^{|J|} Pr(u_i).[Rc(u_i) - Rc(u_{i+1})] \tag{7}$$

where $u_1 = u(q(y_{\phi^{-1}\circ\sigma(1)}|x)) \leq u_2 = u(q(y_{\phi^{-1}\circ\sigma(2)}|x)) \leq \ldots \leq u_{|J|} = u(q(y_{\phi^{-1}\circ\sigma(|J|)}|x))$ with $\phi : i, j, k \rightarrow 1 + k + i.n_x + j.n_x.n_y$ transforms 3D coordinates into indices and $\sigma \in \mathfrak{S}_{|J|}$ is a permutation.

The main advantage of this metric is its independence from uncertainty map scaling and distribution, as only the order of the voxels in the uncertainty map matters. For this reason, AUC-PR provides a quantitative evaluation of the uncertainty map quality that can reliably compare different uncertainty metrics.

Statistical Significance. To assess the statistical significance of our findings a Bayesian point of view is adopted. One can estimate the posterior distribution $p_{A,B}$ of the proportion of experiments where the uncertainty metric A has a higher average of AUC-PR (Eq. 7) over the test set than uncertainty metric B (with metric A and metric B different). In a Bayesian fashion, we choose a non-informative prior distribution of $p_{A,B} \sim \text{Beta}(1, 1)$ which corresponds to a uniform distribution. Over the N experiments, we observe $k_{A,B}$ experiments where metric A gives a better estimate of misclassification than metric B. Then, using Bayes rules, the posterior distribution is the following beta distribution, $p_{A,B} \sim \text{Beta}(1+k_{A,B}, 1+N-k_{A,B})$. From this Bayesian analysis, one can derive $I_{95\%}$, the 95% equally tailed credible interval of the parameter $p_{A,B}$, [7,8].

3 Experiments

Dataset. We used carotid artery MR images acquired within the multi-center, multi-scanner PARISK study [21], a large prospective study to improve risk stratification in patient with mild to moderate carotid artery stenosis (<70%). The standardized MR images acquisition protocol is described in Table 2. We used the images of all enrolled subjects (n = 145) at three of the four study centers as these centers have used the same protocol resulting in a homogeneous set of data: Amsterdam Medical Center (AMC), the Maastricht University Medical Center (MUMC) and the University Medical Center of Utrecht (UMCU). The dataset was split as followed 69 patients in the training set (all from MUMC), 24 patients in the validation set (all from MUMC) and 52 patients in the test set (15 from MUMC, 24 from UMCU and 13 from AMC).

Table 2. MR images scan parameters (QIR = quadruple inversion recovery, TSE = turbo spin echo, IR = inversion recovery, FFE = fast field echo and, TFE = turbo field echo, FA = flip angle, AVS = acquired voxel size, RVS = reconstructed voxel size)

Pulse	T1wQIR TSE		TOF FFE	IR-TFE	T2w TSE
sequence	Pre-contrast	Post-contrast			
Repetion time (ms)	800	800	20	3.3	4800
Echo time (ms)	10	10	5	2.1	49
Inversion time (ms)	282,61	282,61		304	
FA (degrees)	90	90	20	15	90
AVS (mm^2)	0.62 × 0.67	0.62 × 0.67	0.62 × 0.62	0.62 × 0.63	0.62 × 0.63
RVS (mm^2)	0.30 × 0.30	0.30 × 0.30	0.30 × 0.30	0.30 × 0.24	0.30 × 0.30
Slice thickness (mm)	2	2	2	2	2

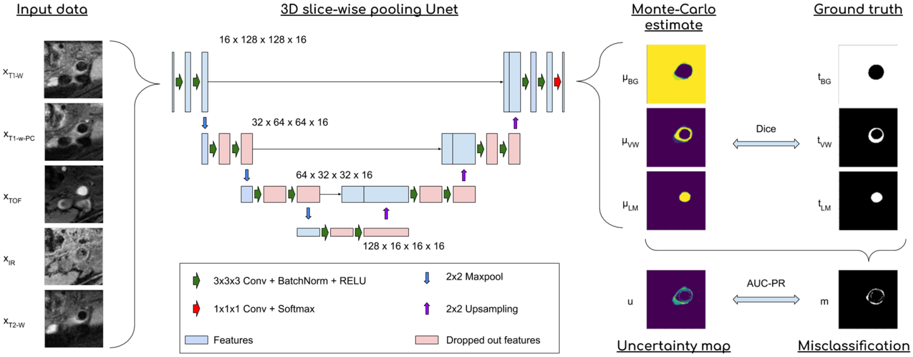

Fig. 2. Description of the network architecture and of the uncertainty map testing framework. The dimensions corresponds to the number of feature maps and the size of the feature maps

MR sequences were semi-automatically affinely and elastically registered to the T1w precontrast sequence. The vessel lumen and outer wall were annotated manually slice-wise, by trained observers with 3 years of experience, in the T1w precontrast sequence. Registration and annotation were achieved with VesselMass software[1]. The intensity histogram was linearly scaled per image such that the 5^{th} % was set to 0 and the 95^{th} % was set to 1. The networks were trained and tested on a region of interest of 128 × 128 × 16 voxels covering one of the common and internal carotid arteries per scan (either left or right).

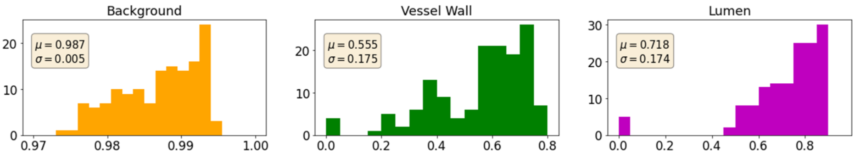

Fig. 3. Distribution over the experiments of the average Dice coefficient, for each of the three classes.

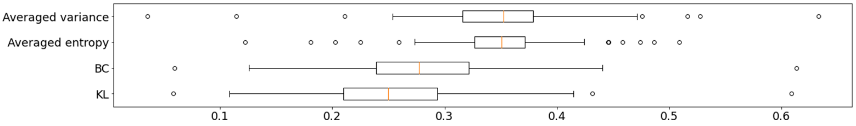

Fig. 4. Distribution over the experiments of the average AUC-PR computed on the test set. The whiskers represent the 5% and 95% interval.

Network Implementation. The networks used for our experiments are based on a 3D U-net architecture [17] as shown in Fig. 2. Because of the low resolution of our problem in the z-axis compared to the resolution on the x-and y-axis, we apply 2D max-pooling and 2D up-sampling slice-wise instead of their usual 3D alternatives. We trained the model using Adadelta optimizer [24] for 600 epochs with training batches of size 1. The network was optimized with the Dice loss [10]. As data augmentation, on the fly random flips along x axis were used. The networks were implemented in Python using Pytorch [16], on a NVIDIA GeForce 2080 RTX GPU.

Parameters Under Study. We varied three parameters in our experiments: the number of images in the training sample to analyse the robustness of the metrics to networks with different level of segmentation performances, the dropout rate, and the dropout type to test different variations of MC dropout. Eight values of number of images in the training set were used: 3, 5, 9, 15, 25, 30, 40 and 69 images. Also, nine dropout rates were used: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. Finally the two types of dropout (Bernoulli and Gaussian

[1] https://medisimaging.com/apps/vesselmass-re/.

dropout) described in Sect. 2 were considered. For every combination of those three parameters, we trained a network following the procedure detailed in previous paragraph. At evaluation time, we discretized the integrals of Eqs. 5, 4 and 6 in $n_{bins} = 100$ bins and we sampled $T = 50$ times using MC dropout method.

Results. A visualization of the different uncertainty measures for different level of performances can be found in Fig. 1. One can find the distribution of the average Dice per class in Fig. 3. The experiment with the highest averaged Dice over classes was observed with a model trained with Gaussian dropout and a dropout rate of 0.3 on the whole training set (69 samples). This method achieved Dice scores of 0.994 on the background, 0.764 on the vessel wall and 0.885 on the lumen. Figure 4 shows the distribution over experiments of the AUC-PR averaged over the test set for the four uncertainty metrics presented in this article.

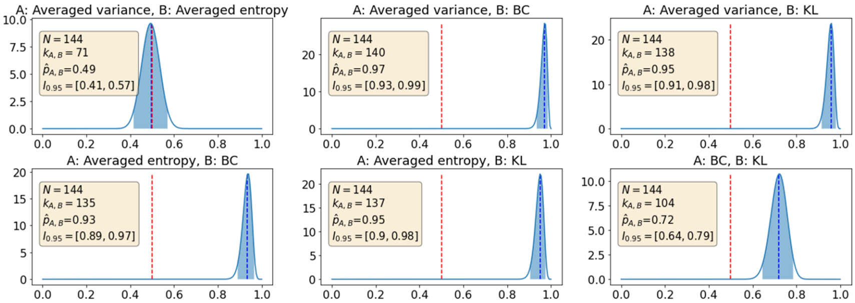

Fig. 5. Posterior distribution of $p_{A,B}$ for different metric pairs A and B, the red dashed line represents the expected value if compared metrics perform equally and the blue area under the curve represents the 95% credible interval

In our pairwise comparison of the four metrics, ten of the twelve combinations of metrics under study showed statistically significant differences over the 144 experiments ($I_{0.95}$ does not contain 0.5). Due to the nature of the beta distributions, the distribution of $p_{a,b}$ and $p_{b,a}$ are symmetric with respect to $y = 0.5$ axis. Therefore, to avoid redundancy only half of the combinations of metric analysis are reported in Fig. 5.

4 Discussion and Conclusion

We presented a quantitative analysis of four uncertainty metrics as predictors of misclassification over a large set of MC dropout variations applied to multi-class segmentation of carotid artery on MR images. This analysis which ranks voxels based on their uncertainty does not take into account the calibration of the different metrics. However, calibration can be performed easily for all metrics based on the validation set, without altering the rank of uncertainty values for individual voxels [12].

Our results showed that metrics considering the statistical description of a distribution averaged over classes performed significantly better than metrics based on distribution similarity of the top two classes when it comes to predict misclassification. Furthermore, BC performed better than KL. Those observations could be attributed to the over-confidence of the softmax output that tends to polarize the distributions to their extreme values (0 or 1) and how sensitive are the different metrics to this polarization. Therefore, in vessel segmentation, taking computation time and metrics performances into account, we advise the use of the averaged variance which does not require the discretisation of an integral voxel-wise. Finally, the good performances of the averaged variance and averaged entropy are consistent with their extensive use in the literature [5,19,23].

Acknowledgments. This work was funded by Netherlands Organisation for Scientific Research (NWO) VICI project VI.C.182.042. The PARISK study was funded within the framework of CTMM, the Center for Translational Molecular Medicine, project PARISK (grant 01C-202), and supported by the Dutch Heart Foundation.

References

1. Chotzoglou, E., Kainz, B.: Exploring the relationship between segmentation uncertainty, segmentation performance and inter-observer variability with probabilistic networks. In: Zhou, L., et al. (eds.) LABELS/HAL-MICCAI/CuRIOUS -2019. LNCS, vol. 11851, pp. 51–60. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33642-4_6
2. Denker, J.S., LeCun, Y.: Transforming neural-net output levels to probability distributions. In: Advances in Neural Information Processing Systems, pp. 853–859 (1991)
3. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
4. Jungo, A., et al.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 682–690. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_77
5. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680 (2015)
6. MacKay, D.J.: A practical Bayesian framework for backpropagation networks. Neural Comput. **4**(3), 448–472 (1992)
7. Makowski, D., Ben-Shachar, M., Lüdecke, D.: bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. J. Open Source Softw. **4**(40), 1541 (2019)
8. McElreath, R.: Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press, Boca Raton (2020)

9. Mehrtash, A., Wells III, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. arXiv preprint arXiv:1911.13273 (2019)
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
11. Mobiny, A., Nguyen, H.V., Moulik, S., Garg, N., Wu, C.C.: DropConnect is effective in modeling uncertainty of Bayesian deep networks. arXiv preprint arXiv:1906.04569 (2019)
12. Mobiny, A., Singh, A., Van Nguyen, H.: Risk-aware machine learning classifier for skin lesion diagnosis. J. Clin. Med. **8**(8), 1241 (2019)
13. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. **59**, 101557 (2020)
14. Neal, R.M.: Bayesian learning via stochastic dynamics. In: Advances in Neural Information Processing Systems, pp. 475–482 (1993)
15. Orlando, J.I., et al.: U2-Net: a Bayesian U-Net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1441–1445. IEEE (2019)
16. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035 (2019)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M.: Probabilistic image registration via deep multi-class classification: characterizing uncertainty. In: Greenspan, H., et al. (eds.) CLIP/UNSURE -2019. LNCS, vol. 11840, pp. 12–22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32689-0_2
19. Seeböck, P., et al.: Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. IEEE Trans. Med. Imaging **39**(1), 87–98 (2019)
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
21. Truijman, M., et al.: Plaque At RISK (PARISK): prospective multicenter study to improve diagnosis of high-risk carotid plaques. Int. J. Stroke **9**(6), 747–754 (2014)
22. Van Molle, P., et al.: Quantifying uncertainty of deep neural networks in skin lesion classification. In: Greenspan, H., et al. (eds.) CLIP/UNSURE -2019. LNCS, vol. 11840, pp. 52–61. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32689-0_6
23. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
24. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)