Laplacian Segmentation Networks Improve Epistemic Uncertainty Quantification

Kilian Zepf[1], Selma Wanna[2], Marco Miani[1], Juston Moore[2], Jes Frellsen[1], Søren Hauberg[1], Frederik Warburg[3], and Aasa Feragen[1]

[1] Technical University of Denmark, Kongens Lyngby, Denmark
{kmze,mmia,jefr,sohau,afhar}@dtu.dk
[2] Los Alamos National Laboratory, Los Alamos, USA {slwanna,jmoore01}@lanl.gov
[3] Teton.ai, Copenhagen, Denmark frederik@teton.ai

Abstract. Image segmentation relies heavily on neural networks which are known to be overconfident, especially when making predictions on out-of-distribution (OOD) images. This is a common scenario in the medical domain due to variations in equipment, acquisition sites, or image corruptions. This work addresses the challenge of OOD detection by proposing Laplacian Segmentation Networks (LSN): methods which jointly model epistemic (model) and aleatoric (data) uncertainty for OOD detection. In doing so, we propose the first Laplace approximation of the weight posterior that scales to large neural networks with skip connections that have high-dimensional outputs. We demonstrate on three datasets that the LSN-modeled parameter distributions, in combination with suitable uncertainty measures, gives superior OOD detection.

Keywords: Uncertainty Quantification · Image Segmentation

1 Introduction

Segmentation is extensively used to quantify organs or anomalies and highlight important image features to clinicians. Its widespread application necessitates strict requirements for safe and interpretable operation. However, modern approaches rely on neural networks which are infamously overconfident on predictions outside of their training distributions [13]. As a result, downstream predictions may be confidently incorrect despite their high accuracy on in-distribution (ID) data, rendering every following analysis based on the prediction unreliable. Thus, detecting gradual distribution shifts is crucial for medical imaging tasks.

In this work, we study Laplace approximations (LA) for epistemic uncertainty quantification in binary image segmentation models. Current Laplace approximations [8] scale quadratically with the output dimension of the neural network, which prevent their usage in segmentation. We develop a fast Hessian approximation for deep architectures with skip connections, which are integral components of segmentation networks, e.g., U-net [31]. This enables us to combine the aleatoric logit distribution of Stochastic Segmentation Networks (SSN) [25] with Laplace approximations for epistemic uncertainty quantification.

We leverage aspects of the ValUES framework [15] to measure the effects of uncertainty estimation on OOD detection. We investigate how useful the inferred uncertainties are for OOD detection and how well LSNs can separate aleatoric and epistemic uncertainty. On three medical binary segmentation tasks, we show that the proposed method provides competitive outlier classification performance and assigns higher uncertainty to OOD datasets. The code is available.[4]

2 Background

While several different sources and taxonomies of uncertainties have been proposed [12,17], the Bayesian framework [4,16] distinguishes between two of them: aleatoric and epistemic. These can be derived directly from the Bayesian model average (BMA) predictive distribution

$$p(y|x, D) = \int \underbrace{p(y|x, \theta)}_{\text{likelihood}} \underbrace{p(\theta|D)}_{\text{posterior}} \, \mathrm{d}\theta, \tag{1}$$

where (x, y) is an input-output pair, D is the training data and θ the model parameters. The Shannon entropy $\mathrm{H}(\cdot)$ of the predictive distribution is a common measure of total predictive uncertainty [14]. This *predictive entropy* can be decomposed into the *expected entropy* as a measure of aleatoric uncertainty, and the *mutual information*, $\mathbf{I}$, representing epistemic uncertainty [16]:

$$\underbrace{\mathrm{H}(p(y|x, D))}_{\text{predictive entropy}} = \underbrace{\mathbb{E}_{p(\theta|D)}[\mathrm{H}(p(y|x, \theta))]}_{\text{expected entropy - aleatoric}} + \underbrace{\mathbf{I}[p(y, \theta|x, D)]}_{\text{mutual information - epistemic}} \quad [33]. \tag{2}$$

Aleatoric uncertainty represents noise or variations that arise from ambiguities in the data, e.g., vague tumor boundaries resulting from gradual tissue infiltration. Epistemic uncertainty, in this framework measured as mutual information $\mathbf{I}$, quantifies the degree to which the model itself should be trusted. However, recent works caution against using this decomposition by demonstrating behavioral incoherences of mutual information [33,34,37], calling for new formulations to calculate epistemic uncertainty. Table 1 lists the measures included in our study, such as expected pairwise KL-divergence (EPKL) [33]: a recently proposed method which uses pairwise comparisons between the predictive distributions of possible models and weights to calculate epistemic uncertainty. We also consider the pixel-wise variance of mean predictions averaged over samples from the posterior distribution, a measure we call Pixel Variance. Prior work [15,26] suggests there is no universal method for uncertainty estimation. Thus, the measure of magnitude of a targeted uncertainty type becomes a design choice that needs to be selected for the dataset and task at hand.

Kendall et al. [16] recommend jointly modelling aleatoric and epistemic uncertainty for regression and classification tasks in computer vision. Prior methods that model both types of uncertainty typically combine Mean-Variance networks

[4] https://github.com/kilianzepf/laplacian_segmentation

Table 1: Overview of uncertainty measures with targeted uncertainty types.

Targeted Type	Uncertainty Measure	Definition
Predictive	Predictive Entropy	$H(\mathbb{E}_{q(\theta\|D)}[p(y\|x,\theta)])$
Aleatoric	Expected Entropy	$\mathbb{E}_{q(\theta\|D)}[H(p(y\|x,\theta))]$
Epistemic	Mutual Information	$I(p(y,\theta\|x,D))$
Epistemic	Expected Pairwise KL	$\mathbb{E}_{q(\theta\|D)}[\mathbb{E}_{q(\tilde{\theta}\|D)}[D_{\mathrm{KL}}(p(y\|x,\theta)\|\|p(y\|x,\tilde{\theta}))]]$
Epistemic	Pixel Variance	$\mathrm{Var}_{q(\theta\|D)}[\mathrm{sigmoid}(\mu_\theta)]$

with diagonal covariance matrices and Dropout [16] or utilize Gaussian-Process based convolutional layers [30]. Recent works, however, focus on modeling either one or the other component [18,25]. Generally, aleatoric uncertainty techniques rely on mixing deterministic segmentation architectures with generative components [3,19,35]. Common epistemic modeling techniques incorporate dropout, ensembles and multi-head models [16,20,21,32].

Fairly evaluating the performance of uncertainty estimation techniques on real-world tasks is challenging due to combinatorial design factors. Kahl et al. [15] address this issue by providing a framework which standardizes evaluations for uncertainty estimation methods. Our work coheres to this framework in the sense that we determine how Prediction Models, Uncertainty Measures and Aggregation strategies impact uncertainty estimation on OOD detection tasks for a fixed U-net architecture. Our results add insight to the recent discussion which calls into question the suitability of common uncertainty measures [33,34,37].

3 The Laplacian Segmentation Networks

To model the posterior distribution in Eq. (1) we apply Laplace's method which approximates the weight posterior with a Gaussian distribution $q(\theta)$ around a local mode θ_{MAP} using the Hessian matrix $\mathbf{H}$ [22]

$$q(\theta) = \mathcal{N}(\theta|\theta_{\mathrm{MAP}}, \mathbf{H}^{-1}). \tag{3}$$

Evaluating $\mathbf{H}$ is computationally infeasible because of the quadratic complexity in network parameters and the large output dimensions for segmentation. We improve upon Hessian approximation techniques [8,5] by extending recent progress in scaling LA for images [24] to segmentation networks with skip connections.

3.1 Laplace Approximation of the Mean Network

We can reformulate the integral for the predictive distribution over the binary predictions y in Eq. (1) by integrating over logits η to obtain

$$p(y|x,D) = \iint p(y|\eta)p(\eta|x,\theta)p(\theta|D)\,\mathrm{d}\eta\,\mathrm{d}\theta. \tag{4}$$

Following [16] and [25], we model the conditional distribution over logits $p(\eta|x,\theta)$ as a normal distribution parametrized by neural networks μ and Σ :

$$\eta|x \sim \mathcal{N}(\mu(x,\theta_1), \Sigma(x,\theta_2)), \tag{5}$$

and assume pixel-wise independence for the predicted labels given the logits. Thus, we can model $p(y|\eta)$ for each pixel s as a Bernoulli distribution parametrized by the sigmoid of the respective logit. Since the size of the covariance matrix Σ scales quadratically with the number of pixels S in the image, we use the low-rank parameterisation of [25]:

$$\Sigma(x) = D(x) + P(x)^T P(x), \tag{6}$$

i.e. the variance network $\Sigma(x)$ is implemented with two networks $D(x)$ and $P(x)$.

The vectors $\theta_1 \in \Theta_1 = \mathbb{R}^T$ and $\theta_2 \in \Theta_2 = \mathbb{R}^T$ parameterize the mean and variance networks (c.f. Eq. 5) and share the first t entries, i.e.we define the shared weight vector θ_t of the network by

$$\theta_t \coloneqq (\theta_{1_1}, \ldots, \theta_{1_t}) = (\theta_{2_1}, \ldots, \theta_{2_t}) \in \Theta_t = \mathbb{R}^t. \tag{7}$$

Then $\theta \in \Theta = \mathbb{R}^{(t+2\cdot(T-t))}$ contains all model parameters

$$\theta \coloneqq (\theta_t, \theta_{1_{t+1}}, \ldots, \theta_{1_T}, \theta_{2_{t+1}}, \ldots, \theta_{2_T}). \tag{8}$$

The post-hoc Laplace approximation first finds a mode θ_{MAP} by minimizing

$$\begin{aligned}\mathcal{L}(\theta) = &- \log \mathbb{E}_{p(\eta|x,\theta)}[p(y|\eta)] - \log p(\theta) \approx \\ &- \text{logsumexp}_{m=1}^{M} \left(\sum_{s=1}^{S} \log p(y_s|\eta_s^{(m)}) \right) + \log(M),\end{aligned} \tag{9}$$

where M logits η are sampled from the distribution in Eq. (5) and where the term $\log p(\theta)$ vanishes assuming a flat prior $\nabla_\theta p(\theta) = 0$. Since current algorithms for fast Hessian computations have no implementation for this loss function, we instead make use of the shared weights in the parameter vectors to estimate the mean and variance of the logit distribution based on the feature maps of a deep deterministic segmentation model. Using only one convolutional layer each for mean and variance estimation, we omit the entries of the variance heads on the parameter vector θ_{MAP}, i.e. we set

$$\theta^*_{\text{MAP}} \coloneqq \theta_{\text{MAP}}\big|_{(\theta_t, \theta_{1_{t+1}}, \ldots, \theta_{1_T})} \in \Theta_{\text{mean}} = \mathbb{R}^T. \tag{10}$$

We can make use of the fact that the SSN loss function reduces to the binary cross entropy loss under zero variance, which allows us to fall back on the fast Hessian computation frameworks available. The posterior is then found by Laplace's method resulting in a Gaussian approximation in the parameter space Θ_{mean}

$$q(\theta^*) = \mathcal{N}\left(\theta^* | \theta^*_{\text{MAP}}, \mathbf{H}^{*-1}\right), \tag{11}$$

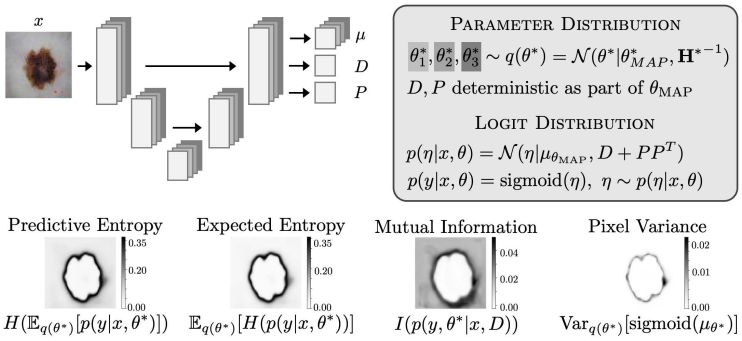

Fig. 1: Model overview - uncertainty measures are calculated by approximating expectations by Monte Carlo-sampling mean networks from the Laplace approximation $q(\theta^*)$ and predicting the respective logit distributions $p(\eta|x,\theta)$ for x.

with $\mathbf{H}^*$ defined as $\mathbf{H}^* = -\nabla_{\theta^*}\nabla_{\theta^*} \log p(\theta^*|D)|_{\theta^*=\theta^*_{\text{MAP}}}$. During inference we can now sample segmentation networks from the posterior distribution in form of the Laplace approximation. Each sampled segmentation network predicts one logit distribution. Figure 1 gives an schematic overview of the proposed Laplacian Segmentation Network (LSN) and derived uncertainty measures.

3.2 Fast Hessian Approximations for Segmentation Networks with Skip Connections

Computation of second order derivatives for Segmentation Networks is expensive due to the vast amount of parameters and pixels in the output. Standard methods approximate the Hessian with the diagonal of the Generalized Gauss Newton (GGN) matrix [10,5]. This approximation, besides enforcing positive definiteness, also allows for an efficient backpropagation-like algorithm. The required compute scales linearly in the number of parameters and quadratic in the number pixels. The quadratic dependency is prohibitive already with images of size 64×64. We therefore make use of the diagonal backpropagation (DB) proposed by [24], which returns a trace-preserving approximation of the diagonal of the GGN. The complexity of this approximation scales linearly with the number of pixels, allowing the computation of the Hessian also for larger images. The idea is to add a diagonal operator $\mathcal{D}$ in-between each backpropagation step. For each layer l

$$
\begin{aligned}
[\nabla_\theta \nabla_\theta \log p(\theta|D)]_l &\overset{\text{GGN}}{\approx} [J_\theta f_\theta(x)^\top \mathbf{H}^{(L)} J_\theta f_\theta(x)]_l = \\
&= {J_\theta f^{(l)}}^\top \left(\prod_{i=l+1}^{L} {J_x f^{(i)}}^\top \mathbf{H}^{(L)} \prod_{i=L}^{l+1} J_x f^{(i)} \right) J_\theta f^{(l)} \\
&\overset{\text{DB}}{\approx} {J_\theta f^{(l)}}^\top \mathcal{D}\left({J_x f^{(l+1)}}^\top \mathcal{D}\left(\dots \right) J_x f^{(l+1)} \right) J_\theta f^{(l)}
\end{aligned}
\tag{12}
$$

where J_θ denotes the Jacobian and $\mathbf{H}^{(L)}$ the Hessian of the binary cross entropy loss with respect to the logits. The Hessian matrix can be expressed in closed form as a diagonal matrix plus an outer product matrix.

Moreover, we extend the `StochMan` library [9] with support for skip-connection layers. For a given submodule f_θ, a skip-connection layer SC_f concatenates the function with the identity, such that $\text{SC}_f(x) = (f_\theta(x), x)$. The Jacobian is then defined as $J_x\text{SC}_f(x) := (J_x f_\theta(x), \mathbb{I}_x)$. We utilize the block structure of the Jacobian matrix and efficiently backpropagate its diagonal only. With a recursive call on the submodule f, the backpropagation supports nested skip-connections, i.e. when some submodules of f are skip-connections as well. This unlocks the use of various curvature-based methods for segmentation architectures with skip connection in future research. For a technical description of the used Hessian approximation we refer to the supplementary material.

4 Experiments

Our method validation is based on the ValUES framework by [15] for evaluating segmentation uncertainty. All benchmark models are constructed by combining an aleatoric and an epistemic component, using the same U-net backbone architecture for comparability. As aleatoric components we consider mean predictions of a U-net and mean-variance predictions of SSN, with diagonal and low-rank covariance matrices. The epistemic components are implemented as Ensembles, MC-Dropout and our post-hoc Laplace approximation. For the nine *method combinations* we calculate the following uncertainty measures: Predictive Entropy, Expected Entropy, Mutual Information, Expected Pairwise KL (EPKL) [33] and Pixel Variance. With exception of the EPKL all measures yield pixel-wise uncertainty heatmaps. We consider sum aggregation as well as a patch based strategy to take the uncertainty from pixel to image level. Patch aggregation sums uncertainties within a 10^2 sliding window across the image, selecting the patch with the highest uncertainty as the image-level score. Table 1 lists the calculated uncertainty measures along their targeted uncertainty type and their definition.

All experiments are conducted on three datasets: the ISIC19 skin lesion dataset [7,6,36], the BRATS dataset [23,1,2] and the first Prostate segmentation task from the QUBIQ 2021 challenge[5]. For the ISIC19 dataset we assign three OOD datasets, representing distribution shifts: Derm-Skin (DERM), Clin-Skin (CLINIC) [27] and the PAD-UFES-20 dataset [28]. The Derm-Skin dataset contains 1,565 images of healthy skin, cropped out of the ISIC dataset. The Clin-Skin dataset contains 723 images showing healthy skin gathered from social networks. The PAD-UFES-20 dataset contains 1570 photos of skin lesions collected from smartphone cameras. For the BRATS and Prostate datasets, we follow the experimental setup of [11] and augment the images with Motion, Spike, Ghosting and Noise artifacts, which are regularly observed in MR images. For training and implementation details we refer to the supplementary material.

[5] https://qubiq21.grand-challenge.org

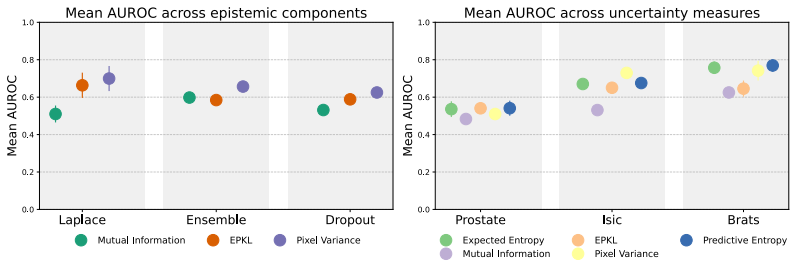

Fig. 2: **Left** - OOD performance measured by AUROC across epistemic components for Mutual Information (MI), Expected Pairwise KL (EPKL) and Pixel Variance (PV). Models using Laplace Approximations with EPKL and PV reach highest AUROC values on average. **Right** - Uncertainty Measures for predictive and aleatoric uncertainty perform on par indicating weak disentanglement.

Image Level OOD detection can be viewed as a binary classification task across the ID test set and all OOD test sets. We therefore calculate the Area Under the Receiver Operating Characteristic Curve (AUROC) for all method combinations, uncertainty measures and aggregation strategies to assess how well different combinations can separate OOD from ID images. The AUROCs were calculated with `sklearn` [29] using per image ground truth binary labels (0-ID, 1-OOD) versus uncertainty scores as target predictions. Figure 3 shows the mean AUROC values for different method combinations and uncetainty measures across datasets. Note that for the method combinations, which use a U-net as an aleatoric component, the EPKL is not defined since it uses a KL divergence term, which is again not defined for two Dirac measures. For the Prostate and ISIC dataset, combinations of the LSN with EPKL and PV yield the highest AUROC values. On the BRATS dataset method combinations that use the Laplace approximation range after models using Dropout. We find that over all datasets method combinations that use the Laplace approximation for their parameter distribution yield the best OOD capabilities when combined with Pixel Variance and EPKL as shown in the left plot of Figure 2.

Additionally, we evaluate how well certain distribution shifts are detected by calculating the mean epistemic uncertainty assigned to a given OOD test set and normalizing it by the uncertainty assigned to the respective ID test set obtained from our ID dataset split. This compares how different distribution shifts influence the absolute values of epistemic uncertainty assigned, providing intuition on each method's sensitivity to OOD data. In Table 2 we display which method combinations on the EPKL measure are able to assign 5% and 10% more uncertainty to the OOD set. Since the EPKL measure does not require aggregation over pixels, we can compare the method combinations directly on all datasets without bias introduced by a aggregation strategy. We find that the LSN model can detect most distribution shifts in terms of assigning a higher

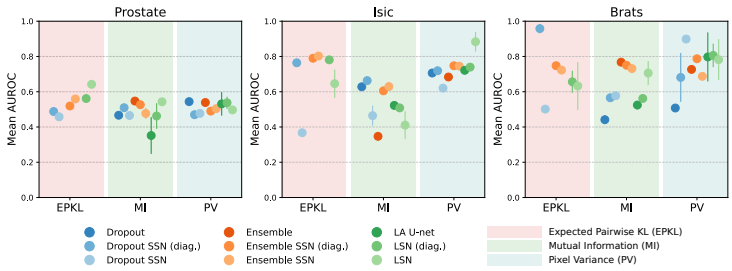

Fig. 3: **OOD performance** measured by AUROC across models, marginalized over aggregation strategies, for Mutual Information (MI), Expected Pairwise KL (EPKL) and Pixel Variance (PV). LSN models with EPKL and PV reach highest AUROC values for Prostate and ISIC respectively.

EPKL. The method, however, fails to identify Random Motion augmentations as ID, which have been used to augment images during training.

Table 2: OOD datasets with uncertainty measured by EPKL compared to ID datasets. A orange check (✓) signifies an average uncertainty that is at least 5% higher than the ID dataset, while a green check (✓) indicates a 10% or greater increase in uncertainty. For the ID Motion dataset an orange check (✓*) signifies an average uncertainty that is at most 5% lower than the ID dataset.

	Prostate				Brats				ISIC				
Model	Motion	Ghost	Spike	Blur	Motion	Ghost	Spike	Blur	Clin	Derm	Padufes	✓	✓
Ensemble SSN (diag.)	✗	✓✓	✓✓	✗	✗	✓✓	✓✓	✗	✓✓	✓✓	✓✓	7	7
Ensemble SSN	✗	✓✓	✓	✗	✗	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	8	7
Dropout SSN (diag.)	✗	✓✓	✓✓	✗	✗	✓✓	✓✓	✗	✓✓	✓✓	✓✓	7	7
Dropout SSN	✓*	✓✓	✓✓	✗	✗	✓✓	✓✓	✗	✗	✗	✓✓	5	6
LSN (diag.)	✓*	✓	✗	✗	✗	✓✓	✓✓	✗	✓✓	✓✓	✓✓	7	5
LSN	✗	✓✓	✓✓	✓	✗	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	**9**	**8**

5 Discussion and Conclusion

In this paper, we have demonstrated how Laplace approximations can scale to image segmentation tasks, through a trace-preserving diagonal Hessian approximation. Importantly, this scales linearly with the number of image pixels, unlike past work which exhibited a quadratic complexity. We have demonstrated across different datasets that the parameter distributions obtained by Laplace's

method, in combination with suitable uncertainty measures, can lead to superior OOD detection performance on image level.

Our experimental findings support the recent initiative in research for finding better measures for epistemic uncertainty than Mutual Information [33]. Marginalizing over all datasets and aggregations strategies, our findings show that EPKL and Pixel Variance, not Mutual Information, provide the strongest discriminative power for classifying images as either ID or OOD (cf. Fig. 2, left). Further we find that there is still a strong correlation between aleatoric and epistemic measures across all method combinations, visible by comparable AUROC performance over all datasets(cf. Fig. 2, right).

Further research might investigate in more depth how different logit distributions interplay with Laplace approximations in theoretically suggested uncertainty measures that target aleatoric and epistemic uncertainty. The presented method provides an extendable framework for other researchers to build upon.

Acknowledgments. This work was supported by VILLUM FONDEN (grants 15334, 42062), the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant 757360), Novo Nordisk Foundation (NNF20OC0062606), LANL (LA-UR-24-23937) LDRD grant 20210043DR (U.S. DOE NNSA Contract 89233218CNA000001), and the Pioneer Centre for AI (DNRF grant P1).

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

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
2. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
3. Baumgartner, C.F., et al.: PHiSeg: Capturing uncertainty in medical image segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI) pp. 119–127 (2019)
4. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)
5. Botev, A.: The Gauss-Newton matrix for deep learning models and its applications. Ph.D. thesis, UCL (University College London) (2020)
6. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: IEEE 15th International Symposium on Biomedical Imaging. pp. 168–172. IEEE (2018)
7. Combalia, M., et al.: BCN20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
8. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux–effortless Bayesian deep learning. In: NeurIPS (2021)
9. Detlefsen, N.S., Pouplin, A., Feldager, C.W., Geng, C., Kalatzis, D., Hauschultz, H., González-Duque, M., Warburg, F., Miani, M., Hauberg, S.: Stochman. GitHub. Note: https://github.com/MachineLearningLifeScience/stochman/ (2021)

10. Foresee, F.D., Hagan, M.T.: Gauss-Newton approximation to Bayesian learning. In: Proceedings of International Conference on Neural Networks (ICNN'97). vol. 3, pp. 1930–1935. IEEE (1997)
11. Fuchs, M., Gonzalez, C., Mukhopadhyay, A.: Practical uncertainty quantification for brain tumor segmentation. In: Medical Imaging with Deep Learning (2021)
12. Gawlikowski, J., et al.: A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342 (2021)
13. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
14. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning (2011)
15. Kahl, K.C., Lüth, C.T., Zenk, M., Maier-Hein, K., Jaeger, P.F.: Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. arXiv preprint arXiv:2401.08501 (2024)
16. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems **30** (2017)
17. Kiureghian, A.D., Ditlevsen, O.: Aleatory or epistemic? Does it matter? Structural Safety **31**(2), 105–112 (2009)
18. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. Advances in Neural Information Processing Systems **31** (2018)
19. Kohl, S.A.A., et al.: A hierarchical probabilistic U-Net for modeling multi-scale ambiguities. arXiv preprint arXiv:1905.13077 (2019)
20. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems **30** (2017)
21. Lee, S., Purushwalkam Shiva Prakash, S., Cogswell, M., Ranjan, V., Crandall, D., Batra, D.: Stochastic multiple choice learning for training diverse deep ensembles. Advances in Neural Information Processing Systems **29** (2016)
22. MacKay, D.J.: Bayesian interpolation. Neural computation **4**(3), 415–447 (1992)
23. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
24. Miani, M., Warburg, F., Moreno-Muñoz, P., Detlefsen, N.S., Hauberg, S.: Laplacian autoencoders for learning stochastic representations. In: Advances in Neural Information Processing Systems (2022)
25. Monteiro, M., et al.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems **33**, 12756–12767 (2020)
26. Mucsányi, B., Kirchhof, M., Oh, S.J.: Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks (2024)
27. Pacheco, A.G.C., Sastry, C.S., Trappenberg, T., Oore, S., Krohling, R.A.: On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3152–3161 (2020). https://doi.org/10.1109/CVPRW50498.2020.00374
28. Pacheco, A.G.h.o.: Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in brief **32**, 106221 (2020)
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

30. Popescu, S.G., Sharp, D.J., Cole, J.H., Kamnitsas, K., Glocker, B.: Distributional Gaussian process layers for outlier detection in image segmentation. In: International Conference on Information Processing in Medical Imaging. pp. 415–427. Springer (2021)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
32. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3591–3600 (2017)
33. Schweighofer, K., Aichberger, L., Ielanskyi, M., Hochreiter, S.: Introducing an improved information-theoretic measure of predictive uncertainty (2023), https://openreview.net/forum?id=c71B6zW70d
34. Schweighofer, K., Aichberger, L., Ielanskyi, M., Klambauer, G., Hochreiter, S.: Quantification of uncertainty with adversarial models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=5eu00pcLWa
35. Selvan, R., Faye, F., Middleton, J., Pai, A.: Uncertainty quantification in medical image segmentation with normalizing flows. In: Machine Learning in Medical Imaging. Lecture Notes in Computer Science, Springer, Switzerland (2020)
36. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 1–9 (2018)
37. Wimmer, L., Sale, Y., Hofman, P., Bischl, B., Hüllermeier, E.: Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In: Evans, R.J., Shpitser, I. (eds.) Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence. Proceedings of Machine Learning Research, vol. 216, pp. 2282–2292. PMLR (31 Jul–04 Aug 2023), https://proceedings.mlr.press/v216/wimmer23a.html