Uncertainty-Aware Classification: A Human-Guided Bayesian Deep Learning Framework

Peter J.T. Kampen$^{\star 1 [0009-0004-4656-946X]}$, Marcel Reimann$^{\star 1,2 [0009-0002-9240-7213]}$, Morten Rieger Hannemose$^{1 [0000-0002-9956-9226]}$, Anders Nymark Christensen$^{1 [0000-0002-3668-3128]}$, Miriam Kolko$^{2,3,4 [0000-0001-8697-0734]}$, Anders Bjorholm Dahl$^{1 [0000-0002-0068-8170]}$, and Josefine Vilsbøll Sundgaard$^{1,5 [0000-0003-2872-4660]}$

[1] Technical University of Denmark, Kongens Lyngby, Denmark
{pjtka,mohab,anym,abda,josh}@dtu.dk
[2] University of Copenhagen, Copenhagen, Denmark
{marcel.reimann,miriamk}@sund.ku.dk
[3] Copenhagen University Hospital, Copenhagen, Denmark
[4] Your Eye Doctors, Ringsted, Denmark
[5] Novo Nordisk A/S, Søborg, Denmark
⋆ These authors contributed equally.

Abstract. While neural networks achieve strong performance in medical image analysis, effectively combining their predictions with human expertise remains a critical challenge for clinical deployment. We examine how different choices of stochastic parameter subsets used in approximate Bayesian inference impact the posterior predictive distributions and, consequently, the performance of a combined human-AI decision model. Using two medical classification tasks, we analyze the relationship between the resulting model and human uncertainty. We demonstrate that uncertainty estimates correlate differently with human uncertainty depending on the stochastic subsets. Building on these findings, we propose a framework that optimizes the choice of stochastic subsets to improve a final decision model that considers human uncertainty, enabling more reliable and interpretable integration of human and AI predictions in clinical settings. Our implementation is publicly available at https://github.com/mkreimann/uncertainty-guided-classification.

Keywords: Uncertainty · Model calibration · Trustworthy AI

1 Introduction

Medical image classification systems enhanced by artificial intelligence (AI) have demonstrated remarkable diagnostic accuracy in recent years. Deep learning models, trained on large datasets, can now rival or even surpass human physicians in specific diagnostic tasks [2, 21]. However, AI models have shown a lack of reliability as stand-alone systems in clinical settings [20]. Recent research suggests that AI may be best utilized not as a replacement for physicians but as a support tool [2, 18]. Across multiple fields, augmenting human decision-making with AI has been shown to increase predictive performance [2]. For example, Reverberi *et al.* [18] show that AI-augmented clinical decision-making in the subject of colon lesion diagnostics results in higher performance than humans or AI alone. Humans can interact with AI systems through the combination of the models' decision and the uncertainty of that decision [1]. However, for optimal interaction, the uncertainty estimates must be calibrated towards optimized combined performance. Bayesian methodology has been widely used to improve model calibration and uncertainty estimation [24]. However, accurate Bayesian inference over all model weights is often computationally infeasible. Consequently, studies have investigated how the application of Bayesian inference over different subsets of model weights affects the posterior predictive distribution and the calibration of the model [6, 19]. Partially stochastic models have been shown to alleviate computational limitations while retaining the added properties of improved uncertainty estimation [11]. In this work, we investigate the relation between the posterior predictive distribution elicited by different choices of stochastic subsets and their relation to the uncertainty of human annotators, particularly towards optimizing the potential AI-human interaction.

Our contributions are as follows: First, we demonstrate that applying Bayesian inference over certain subsets of the model improves calibration. Simultaneously, it leads to improved alignment with human annotator uncertainty, as Bayesian inference primarily affects the probabilities of uncertain samples. Secondly, we are the first to propose a framework for combining (partially) stochastic models and human uncertainty into a *Final Decision Model* (FDM). We find this combination to consistently yield better performance than the already-known benefits of combining an uncalibrated model and humans. Finally, we demonstrate that our method requires very few samples with uncertainty annotations to significantly improve classification performance, making deployment feasible.

1.1 Related work

The approach of treating only a subset of model parameters as stochastic has gained traction in recent years [6, 19]. In Daxberger *et al.* [5], the linearized Laplace approximation (LA) [9] is extended to subnetworks and made more practical for neural networks by leveraging the generalized Gauss-Newton Hessian approximation. Sharma *et al.* [19] argue that partially stochastic models are theoretically as sound as fully Bayesian models and can even, at times, lead to better performance. Kampen *et al.* [11] show that various choices of stochastic subnetworks lead to significant performance increases, particularly in terms of calibration. Bansal *et al.* [1] optimized the model predictive distribution to maximize utility as a team member with simulated human interactions and uncertainty. This methodology has since been extended to include actual human annotations in a post-hoc final training procedure [12, 16]. Ju *et al.* [10] utilize

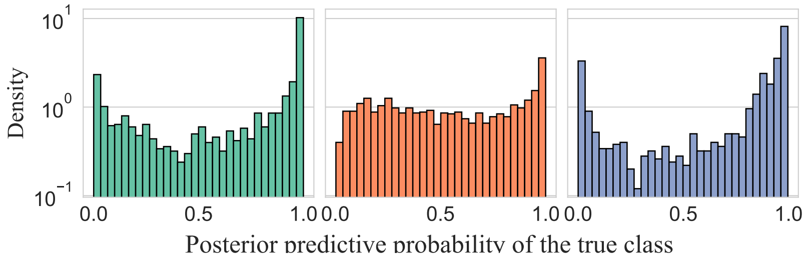

Fig. 1: Histogram of the posterior predictive probability of the true class (x-axis) on the skin lesion test set, showcasing the predictive distribution change elicited by different stochastic subsets. The posterior distribution is estimated using the Laplace approximation of a `ViT` for 3 different choices of stochastic subsets (blocks: 0-1, 13-14, and 19-20 from left).

model calibration in combination with human uncertainty labels to improve the selection of samples for training on noisy datasets.

2 Methods

Our *Final Decision Model* (FDM) improves human-AI collaboration by optimizing model uncertainty and including human uncertainty estimates. Let $\pi(\boldsymbol{x})$ be a measure of the human uncertainty of image $\boldsymbol{x}$ and $g(\boldsymbol{x}) = p\,(t|f(\boldsymbol{x}))$ be the posterior predictive distribution of a neural network f on the same input with target t. Based on this, we define $h : (g(\boldsymbol{x}), \pi(\boldsymbol{x})) \to \mathbb{R}^C$ as the FDM. On the surface, two factors affect the performance of the FDM, namely the choice of parameterization of h and the posterior predictive distribution g. We choose to focus on the latter, and select h to be a logistic regression because it requires few samples and is easy to interpret.

Partial Bayesian inference is widely used to alter the posterior predictive distribution of a neural network to improve calibration while maintaining computational feasibility [11, 19]. In the partially stochastic setting, posterior inference is performed over a subset of the model weights $\boldsymbol{w}_S$, while keeping the remaining parameters deterministic. As seen in Fig. 1, different choices of stochastic subsets can significantly alter the posterior predictive distribution. Therefore, we explore the problem of maximizing the predictive performance of the FDM over the computationally feasible choices of stochastic subsets and the resulting posterior predictive distributions g_S. For each choice of S, the parameters in the logistic regression h_S are estimated using the maximum a posteriori estimate with a Gaussian prior.

$$\text{FDM} = \underset{\text{S}}{\arg\max} \int_{p(\mathcal{D})} \text{Utility}(t, h_S(g_S(\boldsymbol{x}), \pi(\boldsymbol{x})))\ d\boldsymbol{x}dt. \tag{1}$$

Table 1: Calibration metrics on both datasets for the deterministic model (Base model), LA, and SWAG. The best value over model depth is reported for LA and SWAG ± 1 standard deviation. Best performance in **bold**, lower is better.

	Fundus Dataset			Skin Lesion Dataset		
	Base model	LA	SWAG	Base model	LA	SWAG
NLL	0.15 ± 0.01	0.15 ± 0.01	$\mathbf{0.13 \pm 0.01}$	1.24 ± 0.21	$\mathbf{0.89 \pm 0.03}$	1.09 ± 0.19
WNLL	0.22 ± 0.01	$\mathbf{0.20 \pm 0.01}$	0.22 ± 0.01	1.82 ± 0.29	$\mathbf{1.12 \pm 0.12}$	1.52 ± 0.42
MCE	0.53 ± 0.02	0.49 ± 0.02	$\mathbf{0.42 \pm 0.05}$	0.20 ± 0.03	$\mathbf{0.15 \pm 0.03}$	0.16 ± 0.05

As exact posterior inference even over a subset of weights remains computationally infeasible, we rely on approximate inference, and thus g_S also depends on the choice of inference method. We explore the optimization in Eq. (1) for the two widely used approximate inference methods, the LA [5] and Stochastic Weight Averaging Gaussian (SWAG) [14]. Both methods are applied post-hoc and rely on a maximum likelihood estimate (MLE) of the model weights and approximate the posterior distribution using a Gaussian distribution $p(\boldsymbol{w}_\text{S}|\mathcal{D}) \approx q(\boldsymbol{w}_\text{S}) = \mathcal{N}(\boldsymbol{w}_\text{S}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We use the linearized Laplace approximation with Kronecker factorization as described by Immer *et al.* [9] and implemented in [5]. This version does not alter the decision given by the argmax of the MLE model's predictions. Instead, it computes an input-dependent temperature $T_{LA}(\boldsymbol{x})$ such that the predictive distribution becomes $p(t|\boldsymbol{x}) \approx \text{Softmax}\left(\left[\frac{f_{\boldsymbol{w}}(\boldsymbol{x})_1}{T_{LA}(\boldsymbol{x})}, \dots, \frac{f_{\boldsymbol{w}}(\boldsymbol{x})_C}{T_{LA}(\boldsymbol{x})}\right]\right)$. In contrast, the posterior mean in SWAG is not equal to the MLE estimate. Consequently, predictions can change significantly from the MLE model, leading to some variance in model performance depending on initialization, hyperparameter configuration, and the number of samples used in the posterior predictive estimate. We refer to [6, 9, 14] for a more comprehensive understanding of the two methods. We also explore the performance of the FDM using a deep ensemble as a fully Bayesian model, and the widely used baseline Monte-Carlo Dropout [7]. Finally, we include Platt-scaling [8] to compare with an input "independent" method that only seeks to improve the calibration of the models.

We explore the relation between human uncertainty and posterior predictive distribution of the partially stochastic models, and how these correlate with the calibration measures on the backbone neural network. Finally, we correlate all three measures to the performance of the resulting FDM. We choose the negative log-likelihood (NLL) and the Maximum Calibration Error (MCE) with adaptive bin placement [17] as MCE is robust toward imbalanced datasets commonly found in the medical domain [15]. For the same reason, we choose balanced accuracy as a measure of the performance of the final decision function.

3 Experimental setup

To demonstrate the general applicability of our methods, we present results on two datasets from different medical domains: dermatology and ophthalmology.

Each dataset is partitioned into certain and uncertain samples based on human annotations. The dermatology, or skin lesion, dataset is a concatenation of the ISIC 2019 challenge dataset [3, 4, 22] and additional proprietary data consisting of 11,530 dermoscopic images with the same eight class labels as the ISIC dataset. 3,030 of the additional images have been classified by 3,836 users of a medical diagnosis training app. The ground truth labels are based on expert diagnoses, usually supported by histopathological findings. The full dataset attributes $[15\%, 54\%, 12\%, 3\%, 12\%, 1\%, 1\%, 2\%]$ of the total number of samples to the respective classes, which are described in the ISIC dataset. The class label distribution for the 3,030 samples with human uncertainty information is $[15\%, 29\%, 17\%, 3\%, 21\%, 9\%, 4\%, 3\%]$. We define certain samples as those with an entropy of votes less than the 75th percentile of the entropy of votes over the whole dataset, and the uncertain as the complement. Additionally, $\pi(\boldsymbol{x})$ is a vector that holds the proportion of overall votes for the majority class and zeros everywhere else. Importantly, the ground-truth labels are not given by the votes, which enables this one-hot-like encoding of the uncertainty. We motivate this definition by how clinicians might, in practice, be able to represent their uncertainty, i.e, by providing a diagnosis and an associated uncertainty. We choose the publicly available JustRAIGS dataset [13] as our ophthalmology dataset. It contains approximately 101k fundus images of the AIROGS study [23] with expert consensus labels for referable (RG) or non-referable glaucoma (NRG) and the individual label from each of the two to three experts per image. The class distribution exhibits 3% RG samples. The set of certain samples on the fundus dataset is $\{\boldsymbol{x} \in \mathcal{D} | \operatorname{Entropy}(\pi(\boldsymbol{x})) = 0\}$, thus $\pi(\boldsymbol{x})$ is defined as zero if the annotators agree and 1 if they do not. On average, human annotators are uncertain in 4% of NRG samples and 18% of RG samples. Contrary to the skin lesion dataset, the ground truth labels are defined by a few annotations per sample, therefore, we only allow the model to know if there is disagreement or not. If the same procedure as for the skin lesion dataset were applied, the model input would include a smoothed version of the ground truth. Training on the fundus dataset utilizes the ophthalmology-specific `ViT` called RETFound [25] with the recommended fine-tuning procedure. For the skin lesion dataset, we use the same architecture with an ImageNet pretrained checkpoint. Additionally, we compare to a convolutional neural network of similar architecture, namely the ConvNeXt V2 Base model. Training details can be found in Section 6.2 and the code repository. All models are trained with a weighted cross-entropy loss and early stopping based on the macro-averaged accuracy on the validation set. The selection of the subset of stochastic weights for SWAG and LA is based on the block structure of `ConvNeXt` and `ViT`. We include two consecutive blocks in each subset, resulting in 18 and 11 subsets for the `ConvNeXt` and `ViT`. This is motivated by the results presented in [11], where the inclusion of 8 weight matrices in the stochastic subset is shown to be near optimal for calibration, corresponding to the number of weight matrices in two `ViT` blocks. The consecutive block structure was chosen to see if any tendencies emerged as a function of model depth. The performance of LA and SWAG is influenced by the choice of prior precision

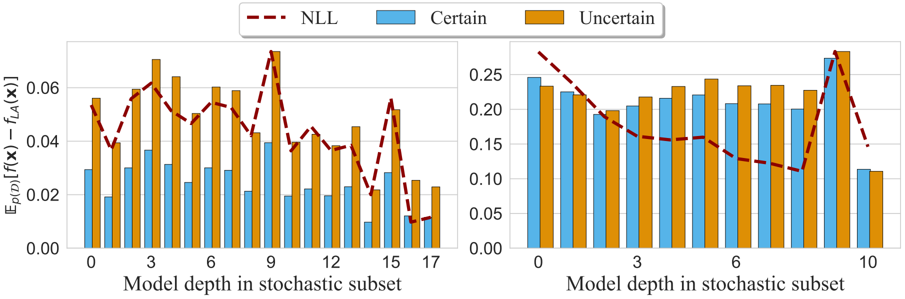

Fig. 2: Average change in predicted probabilities by applying LA over the different stochastic subsets of the `ConvNeXt` on the fundus dataset (left) and the `ViT` on the skin lesion dataset (right). Distances are stratified by sample certainty according to the human annotation. NLL is scaled linearly to the minimum and maximum height of the corresponding bars for easier visualization.

for LA and, particularly, the learning rate used during the SGD iterations for SWAG. We optimize these to minimize the NLL on the validation set. Search parameters can be found in Table 5. Both datasets are split into multiple folds, to allow benchmarking against a deep ensemble. For each fold, we train three models with different seeds. An additional hold-out test set was created with at least 10% of the total number of available images.

4 Results

We report our findings on the `ViT` for the skin lesions dataset and on the `ConvNeXt` for the fundus dataset, as these obtained the best classification performance. On the fundus dataset, we obtain balanced accuracy scores of 89.23% ± 0.56% for the `ViT` and 90.92% ± 0.67% for the `ConvNeXt`. On the skin lesions dataset, the balanced accuracy was 63.8%±1.05% and 61.8%±1.54% for the `ViT` and `ConvNeXt`, respectively. The balanced accuracy of the human annotators on the skin lesion dataset is 61.3% i.e, lower than the performance of the models.

Table 2: Pearson correlations for both LA and SWAG between the balanced accuracy of the different stochastic subsets combined with human uncertainty against RHD from Fig. 2, MCE, and NLL; and RHD against NLL and MCE (fundus/skin lesion dataset). **Bold** denotes statistical significance at $p \leq 0.05$.

	Balanced accuracy			RHD	
	RHD	NLL	MCE	NLL	MCE
LA	**0.40** / 0.13	**-0.33** / **-0.66**	**-0.38** / -0.02	**-0.73** / **-0.32**	-0.23 / 0.03
SWAG	0.11 / -0.09	0.12 / **-0.65**	**0.28** / **-0.59**	**-0.72** / -0.27	**0.29** / -0.27

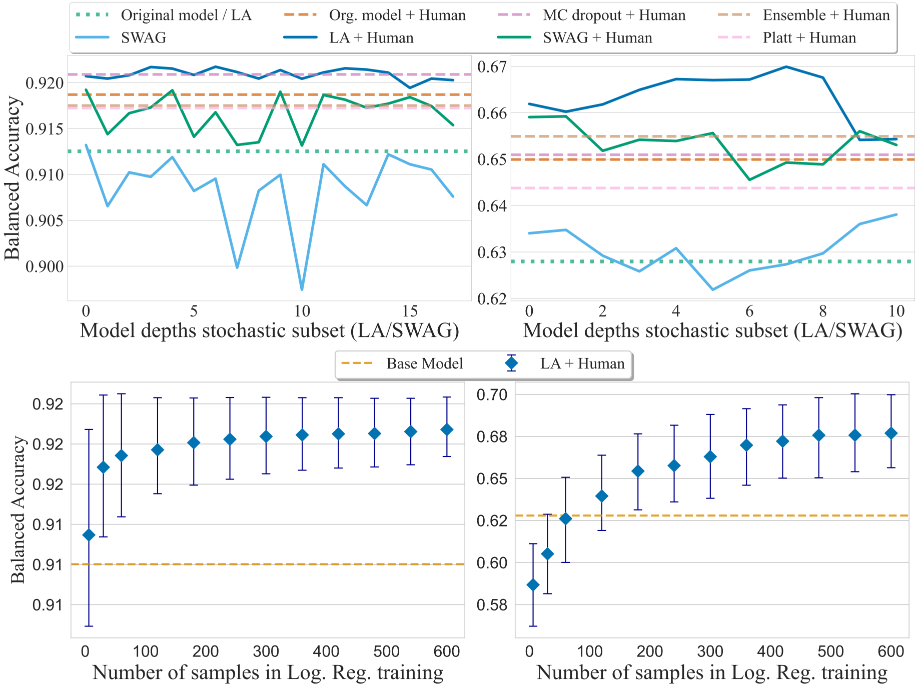

Fig. 3: Left: Fundus dataset, Right: Skin lesions dataset. Top: Mean FDM performance for all methods for stochastic subsets at different depths. Bottom: Balanced accuracy over the number of samples with uncertainty information used to train the logistic regression. Error bars correspond to the interquartile range across 400 random train set splits.

We confirm that applying LA and SWAG improves the calibration of our base models by reporting the NLL, class-weighted NLL (WNLL) and MCE in Table 1. Figure 2 shows the effects of applying LA on different choices of stochastic subsets. On both datasets, we observe larger changes in the model outputs for uncertain samples than for certain samples when applying LA. On the skin lesions dataset, the LA's ability to capture samples with higher human uncertainty is more dependent on the choice of the stochastic subset. For both datasets, the calibration as measured by the NLL depends on the choice of the stochastic subset. In Table 2 we report the partially stochastic models' ability to capture human uncertainty as measured by the correlation between the relative height difference (RHD) between the bars in Fig. 2 and the resulting calibration metrics. There is a significant negative correlation between the RHD and NLL of the models for both datasets. Hence, we observe that the best calibration is obtained when the stochastic model corrects samples with higher human uncertainty the most.

In our second experiment, we evaluate the performance of the FDM across the various methods for posterior inference and over different choices of stochastic subsets. Figure 3 shows a performance increase when combining any model with human uncertainty in an FDM. However, the performance is highest when using Bayesian methodology, either over the full model (MC dropout and ensembling) or over different subsets (SWAG and LA). Specifically, LA yields the best performance across both datasets. In contrast, using SWAG as the FDM backbone on the fundus dataset fails to improve performance, following the behavior of SWAG without human uncertainty information. The sample agnostic calibration of Platt scaling shows to be inferior to the other methods, as well as to the base model. Figure 3 demonstrates our framework's low requirement for human uncertainty labels. We vary the number of random samples with uncertainty labels, drawn from the validation set, that are included in the logistic regression training and present the balanced accuracy of the FDM. We choose g_S to be the LA with the best performing choice of S. The sampling process was class-balanced; however, a minimum of one uncertain sample per class was enforced. The results are averaged over three seeds. The most apparent increase is observed for the fundus dataset, where only 5%, or 30 samples, are required to achieve a mean performance that significantly exceeds that of the base model. A decrease in performance until the inclusion of 100 samples can be observed for the skin lesion dataset, which is most likely caused by eight times as many classes the model has to differentiate. However, for both datasets, we observe a significant improvement over the base model with less than 200 samples for training.

5 Discussion & Conclusion

Our results indicate that improving the uncertainty estimation of the models may simultaneously align the uncertainty distribution with that of humans. Furthermore, we show that using the predictive posterior distribution of Bayesian models in a Final Decision Model (FDM) rather than the base model consistently increases performance, and using LA for posterior inference yielded the best performance on both datasets. Crucially, we also observed a significant correlation between the performance of the FDM and the NLL of the underlying LA model. Hence, we find that NLL is indicative of the alignment between humans and AI models and the resulting diagnostic performance when combined. This suggests that determining a near-optimal choice of stochastic subset can be done independently of human annotations and, therefore, before estimating the parameters of the FDM. The positive correlation coefficient between the human/AI uncertainty alignment and the performance of the FDM is perhaps counterintuitive, as it would seem to violate the notion that an optimal ensemble consists of uncorrelated models. We argue that this is not the case because samples with high human uncertainty are likely to lie close to the decision boundary in the FDM. Hence, when the average change in model predictive probabilities is larger on those cases specifically, it is likely that a better separating hyperplane can be estimated. Although we could not attribute fine-grained uncertainty values to the fundus images, we were able to demonstrate that a simple indication of an uncertain sample is beneficial for the FDM. The low sample requirement necessary to outperform the base model demonstrates the usefulness of selecting a 'simple' model as the decision function h_s, see Fig. 3. The framework is sufficiently general to support arbitrarily complex models; however, training such models would require more data, reducing the framework's overall applicability. We consider the primary use case of this methodology to be integration into the diagnostic pathway. If physicians can reliably assess their uncertainty, then the model can balance the predictions of AI and physicians, potentially improving diagnostic accuracy. The white-box nature of logistic regression enables physicians to understand the importance of their predictions and associated uncertainty, fostering a transparent and strengthened human/AI collaboration.

In summary, we present a pipeline for combining partially stochastic Bayesian models with human uncertainty through a simple classifier that achieves significantly better classification performance. Finally, integrating human expertise directly into the decision takes another step towards improving reliability and enhancing trust in deep learning models in medical applications.

Acknowledgments. This work was partially supported by VELUX Fonden under project number 00048932. Furthermore, the authors would like to extend their gratitude to Tareen Dawood for her continued availability to discuss the paper's content and her invaluable suggestions.

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

References

1. Bansal, G., Nushi, B., Kamar, E., Horvitz, E., Weld, D.S.: Is the most accurate ai the best teammate? optimizing ai for teamwork. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11405–11414 (2021)
2. Becker, J., Decker, J.A., Römmele, C., Kahn, M., Messmann, H., Wehler, M., Schwarz, F., Kroencke, T., Scheurig-Muenkler, C.: Artificial intelligence-based detection of pneumonia in chest radiographs. Diagnostics **12**(6), 1465 (2022)
3. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., 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: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
4. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
5. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux-effortless bayesian deep learning. Advances in neural information processing systems **34**, 20089–20103 (2021)

6. Daxberger, E., Nalisnick, E., Allingham, J.U., Antorán, J., Hernández-Lobato, J.M.: Bayesian deep learning via subnetwork inference. In: International Conference on Machine Learning. pp. 2510–2521. PMLR (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. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks (2017), https://arxiv.org/abs/1706.04599
9. Immer, A., Korzepa, M., Bauer, M.: Improving predictions of bayesian neural nets via local linearization (2021)
10. Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Zhou, Q., Liu, T., Ge, Z.: Improving medical images classification with label noise using dual-uncertainty estimation. IEEE transactions on medical imaging **41**(6), 1533–1546 (2022)
11. Kampen, P.J., Als, G.R., Andersen, M.R.: Towards scalable bayesian transformers: investigating stochastic subset selection for NLP. In: The 40th Conference on Uncertainty in Artificial Intelligence (2024)
12. Kerrigan, G., Smyth, P., Steyvers, M.: Combining human predictions with model probabilities via confusion matrices and calibration. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 4421–4434. Curran Associates, Inc. (2021)
13. Lemij, H.G., de Vente, C., Sánchez, C.I., Vermeer, K.A.: Characteristics of a large, labeled data set for the training of artificial intelligence for glaucoma screening with fundus photographs. Ophthalmology Science **3**(3), 100300 (2023)
14. Maddox, W., Garipov, T., Izmailov, P., Vetrov, D., Wilson, A.G.: A simple baseline for bayesian uncertainty in deep learning (2019), https://arxiv.org/abs/1902.02476
15. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B.e.a.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (Feb 2024). https://doi.org/10.1038/s41592-023-02151-z, http://dx.doi.org/10.1038/s41592-023-02151-z
16. Narasimhan, H., Jitkrittum, W., Menon, A.K., Rawat, A., Kumar, S.: Post-hoc estimators for learning to defer to an expert. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 29292–29304. Curran Associates, Inc. (2022)
17. Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., Tran, D.: Measuring calibration in deep learning (2020), https://arxiv.org/abs/1904.01685
18. Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Cherubini, A.: Experimental evidence of effective human–ai collaboration in medical decision-making. Scientific reports **12**(1), 14952 (2022)
19. Sharma, M., Farquhar, S., Nalisnick, E., Rainforth, T.: Do bayesian neural networks need to be fully stochastic? (2023)
20. Ternov, N.K., Christensen, A.N., Kampen, P.J., Als, G., Vestergaard, T., Konge, L., Tolsgaard, M., Hölmich, L.R., Guitera, P., Chakera, A.H., et al.: Generalizability and usefulness of artificial intelligence for skin cancer diagnostics: an algorithm validation study. JEADV Clinical Practice **1**(4), 344–354 (2022)
21. Tschandl, P., Codella, N., Akay, B.N., Argenziano, G., Braun, R.P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al.: Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. The lancet oncology **20**(7), 938–947 (2019)

22. 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)
23. de Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.D., Galdran, A., Ballester, M.A.G., Carneiro, G., G, D.R., S, H.P., Puthussery, D., Liu, H., Yang, Z., Kondo, S., Kasai, S., Wang, E., Durvasula, A., Heras, J., Zapata, M.A., Araújo, T., Aresta, G., Bogunović, H., Arikan, M., Lee, Y.C., Cho, H.B., Choi, Y.H., Qayyum, A., Razzak, I., van Ginneken, B., Lemij, H.G., Sánchez, C.I.: AIROGS: Artificial Intelligence for RObust Glaucoma Screening Challenge (Feb 2023), http://arxiv.org/abs/2302.01738
24. Wang, C.: Calibration in deep learning: A survey of the state-of-the-art. arXiv preprint arXiv:2308.01222 (2023)
25. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)