Variational Visible Layers: A Practical Framework for Uncertainty Estimation

Zeinab Abboud[1,2] Herve Lombaert[1,2] and Samuel Kadoury[1,3]

[1] Polytechnique Montreal, Montreal, QC, Canada
[2] MILA, Montreal, QC, Canada
[3] CHUM Hospital Research Center, Montreal, QC, Canada
zeinab.abboud@polymtl.ca

Abstract. Uncertainty estimation is critical for reliable decision-making in medical imaging. State-of-the-art uncertainty methods require significant computational overhead and complex modeling. In this work, we present and explore a simple, effective approach to incorporating Bayesian uncertainty into deterministic networks by replacing the first and/or last layer (visible layers) with their variational Bayesian counterpart. This lightweight modification enables uncertainty quantification through mean-field variational estimation, making it practical for real-world medical applications. We evaluate the methods on ISIC and LIDC-IDRI for the segmentation task and DermaMNIST and ChestMNIST for the classification task using post-hoc and jointly-trained visible layers. We demonstrate that variational visible layers enable uncertainty-based failure detection for both in-distribution and near-out-of-distribution samples, preserving task performance while reducing the number of variational parameters required for Bayesian estimation. We provide an easy-to-implement solution for integrating uncertainty estimation into existing pipelines.

Keywords: Uncertainty · Mean-Field Variational Inference · Bayesian Neural Networks · Classification · Segmentation

1 Introduction

Applications of deep neural networks (DNNs) to medical image analysis have advanced significantly in the past few years; however, DNNs still suffer from overconfident predictions, which poses a risk in safety-critical applications. The performance of DNNs is also measured against highly catered benchmark datasets, which are assumed to be independent and identically distributed [22]. The predictive performance of DNNs does not hold beyond the given test sets; when deployed into a real-world environment, these models tend to perform poorly and without any indicators to users of their intrinsic uncertainties [20, 30, 2, 15]. Therefore, uncertainty estimation rises as an important quantitative indicator of the reliability of the model's predictive performance. More importantly, uncertainty quantification bridges the trust gap that hinders clinical adoption of DNNs, providing transparency and enabling informed decision-making.

Bayesian neural networks (BNNs) learn approximate distributions over the weights, in contrast to deterministic DNNs, which rely on point estimates. Since exact Bayesian inference in BNNs is intractable, various approximation methods are employed to estimate the posterior, including Markov Chain Monte Carlo (MCMC) [21, 24], the Laplace approximation [18], and variational inference [14, 16, 5]. MCMC provides a means to approximate the true posterior but incurs a high computational cost, requiring a large number of samples for accurate estimation. The Laplace approximation, while capable of capturing high predictive uncertainty, depends on an approximation of the Hessian, which becomes computationally prohibitive for large-class classification tasks or high-dimensional outputs such as segmentation [8]. Variational inference offers a more scalable alternative for Bayesian inference in large neural networks, by which the mean-field variational inference (MFVI) models network parameters as independent distributions with diagonal covariance. While MFVI scales better than MC sampling and the Laplace approximation, it suffers from a noisy loss landscape, increased training time, and a doubling of the number of learnable parameters [5]. Despite its robustness to distributional shifts, MFVI struggles to scale efficiently to larger datasets, limiting its widespread adoption [22, 9].

Bayesian uncertainty methods are rarely adopted in medical image analysis due to their high training complexity and computational cost, which pose challenges for clinical integration and real-time decision-making. In practice, approximate Bayesian methods such as MC Dropout [10], and deep ensembles [17] are more commonly used for uncertainty estimation due to their ease of implementation. However, deep ensembles require training and evaluating multiple networks, making them computationally impractical, while MC Dropout produces uncalibrated predictions under distributional shifts, reducing its reliability in safety-critical applications like medical image analysis. Therefore, the medical imaging community requires simpler implementations of Bayesian neural networks that provide predictive uncertainty without the additional implementation and training costs. Recent research [8, 26, 23, 11, 1] indicates that reparameterizing a subset of network's parameters probabilistically is sufficient for uncertainty estimation, with strong evidence suggesting that introducing stochasticity near the network input and output yields reliable uncertainty estimates without jeopardizing model performance. Recent work on variational Bayesian last layers (VBLL) [11] has shown strong performance in regression, medium-scale classification on natural images, and classification using LLM features. By restricting sampling to the last layer, VBLL provides efficient uncertainty estimation with minimal computational overhead. However, the method introduces hyperparameters that are challenging to tune, potentially limiting its practicality in real-world applications [11]. Additionally, a sparse sub-network variational inference, guided by sensitivity analysis, has shown that selecting variational parameters near the first and last layers can achieve uncertainty estimates comparable to deep ensembles while using significantly fewer parameters [1]. Building on these insights, we propose and investigate a practical training

strategy for uncertainty estimation that can be applied both post-hoc and joint-training. Our key contributions include:

1. We examine variational visible layers for failure detection, demonstrating their effectiveness in in-distribution and near-out-of-distribution for medical image segmentation and classification.
2. We benchmark uncertainty on ISIC, LIDC-IDRI, and Derma/ChestMNIST under covariate shifts to evaluate uncertainty failure detection.
3. We automate KL divergence weighting in ELBO loss, eliminating the need for hyperparameter tuning.
4. We release a plug-and-play package in PyTorch, enabling easy integration into existing neural network models.[4]

2 Methods

2.1 Problem setup

Given a training dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, where $x_n \in \mathbb{R}^d$ represents a d-dimensional input and $y_n \in \mathbb{C}^k$ denotes the corresponding label among k classes for classification and $y_n \in \mathcal{Y}^d$ for segmentation. Our goal is to train a neural network to model the predictive distribution $p_\theta(y \mid x, \mathcal{D})$, parameterized by θ. In a Bayesian neural network (BNN) formulation, we place a prior $p(\theta)$ over the parameters, and the posterior distribution is given by Bayes' rule: $p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$. The predictive distribution is then obtained by marginalizing over the posterior $p(y \mid x, \mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})}[p(y \mid x, \theta)]$ which we approximate by drawing N Monte Carlo samples from the posterior.

To improve the scalability of Bayesian inference, we employ Gaussian Mean-Field Variational Inference [5] via the reparameterization trick [16], which approximates the posterior with a factorized distribution (independent parameter assumption). To further reduce computational complexity, we limit Bayesian inference to a subset of parameters $\theta_v \in \theta$, treating them as variational parameters $\mathcal{N}(\mu, \sigma^2)$ while keeping the rest of the network deterministic.

2.2 Variational Visible Layers

We introduce two approaches to incorporate Bayesian inference into the first and/or last layers of a neural network, guided by strong evidence [11, 26, 1, 23, 29] that these layers play a crucial role in capturing uncertainty.

1. Post-hoc Reparameterization: Given a pre-trained deterministic model, we reparameterize the selected layer(s) by initializing their posterior means μ to the corresponding point estimates from the pre-trained model, and fine-tune the model. This enables uncertainty estimation without requiring full retraining from random initialization, preserving the original model's learned representations while introducing Bayesian uncertainty.

[4] https://github.com/zabboud/Variational-Visible-Layers

2. Joint Training: The variational Bayesian layer(s) are jointly trained with the deterministic backbone, with all parameters randomly initialized.

In both cases the loss is defined as the weighted sum of the negative log-likelihood and the KL divergence term (-ELBO):

$$\mathcal{L}_{ELBO} = \mathbb{E}[-\log p(\mathcal{D} \mid \theta)] + \beta \cdot D_{KL}\left(q_\phi(\theta) \| p(\theta)\right). \tag{1}$$

We introduce an automated method for scaling the hyperparameter β as equal to the ratio of Bayesian parameters to the total model parameters, as opposed to current methods that require manual hyperparameter tuning. This adaptive weighting prevents the KL divergence term from dominating the likelihood, ensuring that the model effectively captures both uncertainty and task-specific performance. This method also eliminates the need for hyperparameter tuning.

The two training schemes are further divided into subcases, as illustrated in Fig. 1. The Post-hoc Reparameterization method includes six configurations, introducing stochasticity in (1) the first layer, (2) the last layer, or (3) both layers (VVL). For each setup, we compare the performance of freezing the pre-trained deterministic backbone versus allowing it to remain trainable. If freezing produces similar results, it offers a computational advantage by reducing GPU usage during the fine-tuning stage. In the Joint Training scheme, we train a randomly initialized model with variational Bayesian first, last, or both layers (VVL), establishing a baseline to compare against the post-hoc approach.

2.3 Implementation and Experimental Methodology

The approach described above is both model- and task-agnostic. We illustrate the comparative performance of different training strategies in Fig. 1 on medical

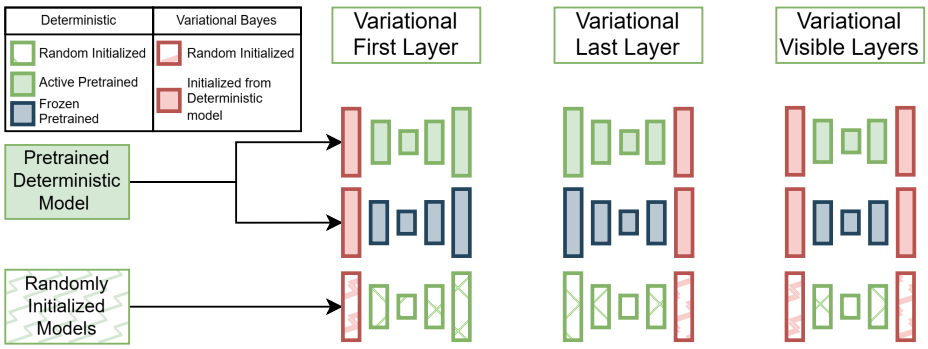


Fig. 1: Training schemes: (1) Post-hoc Reparameterization, introducing stochasticity into a pre-trained deterministic network by reparameterizing one or both visible layers, with the backbone either frozen or trainable; and (2) Joint Training, where a randomly initialized model is trained with one or both variational visible layers (VVL).

image classification and segmentation tasks. To evaluate the performance of our method, we compare it against the following baselines: an ensemble of networks [17], last layer variational inference (VI) [22], which corresponds to the Joint-Last layer training scheme in Fig. 1, a variational inference model, where all parameters are stochastic [5], and Sparse Bayesian Networks, which reparameterize a subnetwork of variational parameters (denoted as Sparse VI) [1]. Additionally, a standard deterministic model is included for comparison. Although MC-Dropout [10] is a widely used approach, we do not include it as a comparative method, as it has been shown to produce overconfident predictions, making it an unreliable estimator of predictive uncertainty [17, 22].

We demonstrate the performance across all the 9 model variants in Fig. 1 compared to the baselines on segmentation of ISIC (2,594/100/1,000) [7, 27] and LIDC-IDRI [3, 4, 6], and classification of DermaMNIST (7,007/1,003/2,005 train/val/test) and ChestMNIST (78,468/11,219/22,433) datasets [28]. LIDC-IDRI volumes were cropped into 15,096 patches of size 128×128, each with segmentation masks from four experts resulting in a 10,567/4,529 split. An input size of 224×224×3 was used for ISIC and MedMNIST datasets.

The models were trained for 100 epochs using cross-entropy loss and an SGD optimizer with an initial learning rate of 0.1. The learning rate was reduced by a factor of 0.1 at epochs 30, 60, and 90. For Post-hoc methods, the initial learning rate was set to 0.01. Weight decay was set to 1×10^{-4} for L2 regularization. Batch normalization was applied to all models, with a batch size of 64, 128 for DermaMNIST and ChestMNIST classification, respectively. A batch size of 10 was used for segmentation of both datasets. Five posterior samples were used for VI sampling, and an ensemble of five members was trained to ensure comparable results. Data augmentation includes random horizontal and vertical flips. To assess reproducibility, each model was trained using five different random seeds (selected sequentially from 0 to 25). For ensembles, each ensemble consists of five models trained with distinct seeds. We report the median performance for each metric along with its 95% confidence interval.

Assessing robustness to covariate shift. To evaluate the calibration and reliability of uncertainty estimates under distributional shifts, we created a variant of the benchmark test sets using the perturbations outlined in [13]. This assessment is critical, as natural variations in imaging protocols across institutions, countries, and equipment can introduce distribution shifts that can significantly impact model performance. Ensuring robustness to these variations is essential for the safe deployment of medical image analysis models. We inject seven types of corruptions, selected from the methods outlined in [13], that are most relevant to variations in medical imaging. These include Gaussian noise, defocus blur, zoom blur, brightness, contrast, saturation, and elastic deformations, each applied at five levels of severity. This creates a robustness benchmark for the ISIC and LIDC-IDRI datasets, specifically targeting segmentation failure detection under covariate distributional shifts.

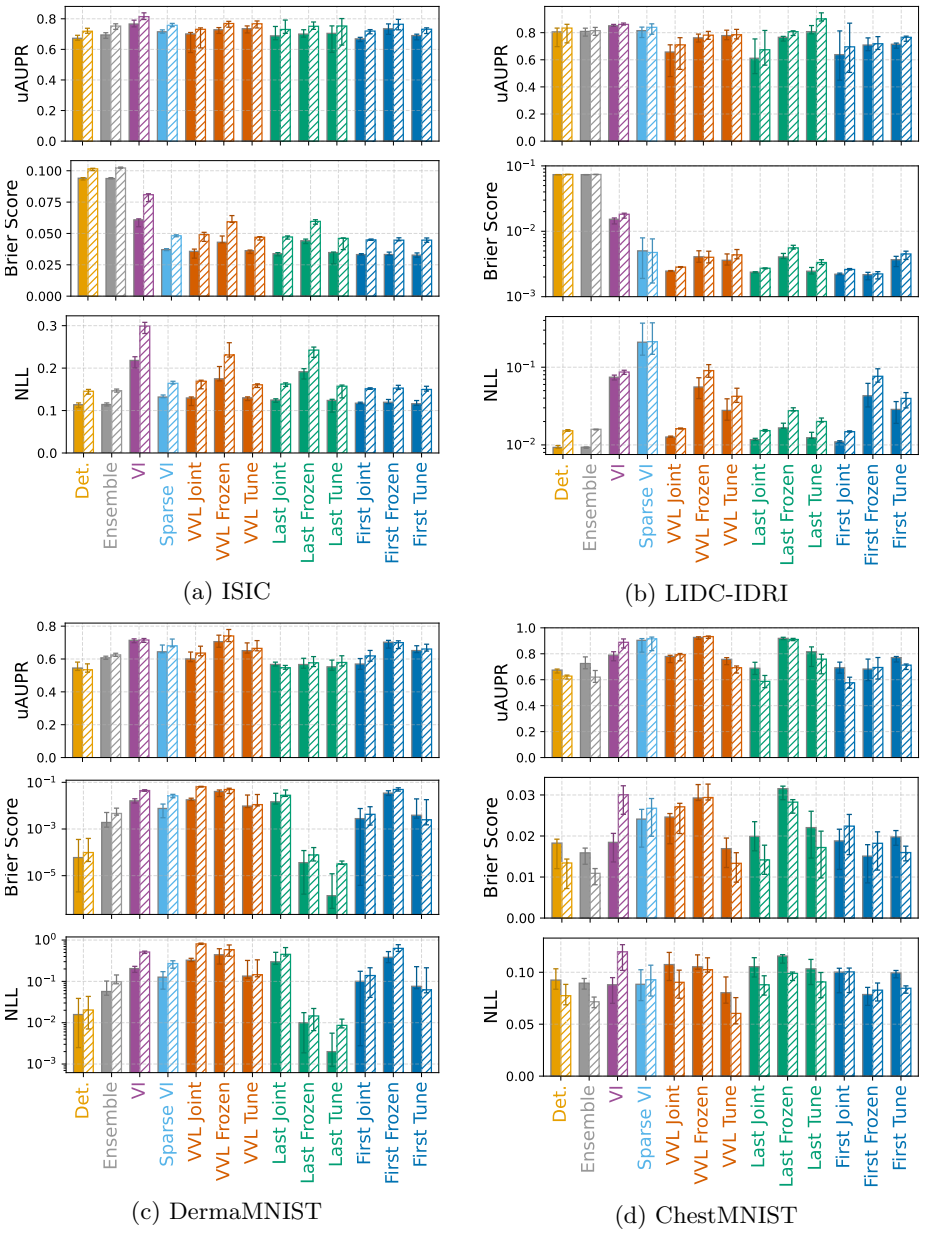
Performance metrics. For ChestMNIST and DermaMNIST, classification accuracy is used as the performance metric. Segmentation performance is as-

sessed using Dice and intersection over union (IoU) scores for the LIDC-IDRI and ISIC datasets, respectively. Brier Score (↓) is used to assess calibration of probabilities (mean square error between one-hot inputs and predicted probabilities). Entropy of expectation of probabilities is used to quantify the total uncertainty. Negative log likelihood (NLL↓) is also used to assess model calibration. To assess the reliability of a given model's uncertainty, we evaluate the area under the precision-recall curve (AUPR↑) for misclassified or poorly segmented samples. For classification, we treat incorrectly classified examples as the positive class [19]. For segmentation, we define a pass/fail criterion based on clinically relevant thresholds for segmentation quality. In this study, we evaluate AUPR using IoU or Dice thresholds, based on the average performance of the model on in-distribution (ID) data, considering any score below the threshold as a failure. The AUPR is then computed using the entropy of expectation (total uncertainty) to measure the quality of the model uncertainty estimates, for both ID test set shifted out-of-distribution (OOD) test set. The uncertainty AUPR (uAUPR), therefore, serves as a failure detection metric.

3 Experimental Results

Model complexity and compute cost vary across approaches. If we take the deterministic model with θ parameters and m compute cost per iteration, then N-Ensembles scale linearly $N\theta, Nm$. Variational first layer and Variational Visible Layers (VVL) add d and $b + d$ parameters, respectively, where d and b correspond to the number of parameters in the first and last layers, while maintaining Nm compute. Last-model reduces compute to $m+b(N-1)$ per iteration. VI doubles parameters 2θ with Nm cost, while Sparse VI limits overhead to 1% extra parameters with Nm cost/iteration.

MedMNIST Classification results on the DermaMNIST and ChestMNIST datasets using a ResNet18 [12] base model are presented in Figure 2. Accuracy and Brier scores were found to be fairly stable across all the baselines and the 9 variants, with an accuracy of 0.735 ± 0.007, 0.947 ± 0.001 for DermaMNIST and ChestMNIST, respectively. For ID performance on DermaMNIST dataset, the deterministic, ensemble, and Post-hoc Last Layer models achieve the best performance based on NLL. However, for OOD failure detection using uAUPR, the VVL-Frozen, VI, and First Layer-Frozen approaches demonstrate the best performance in both ID and OOD settings. In contrast, for the larger ChestMNIST dataset, over an order of magnitude larger than DermaMNIST, the VVL-Frozen, Last Layer-Frozen, and Sparse VI models achieve the highest overall performance for both ID and OOD evaluation. A visualization of the uAUPR performance on both ID and OOD misclassified samples is shown in Figures 2c and 2d. The VI model achieves strong OOD performance, aligning with previous findings that VI outperforms other uncertainty baselines in OOD but does not scale well to large datasets [22]. Last-layer models have among the lowest NLL performances, consistent with previous findings [26], and perform reasonably well on ID/OOD failure detection as a post-training method (Fig. 2c, 2d).



(a) ISIC (b) LIDC-IDRI

(c) DermaMNIST (d) ChestMNIST

Fig. 2: uAUPR ↑, NLL ↓, Brier Score ↓, for ID (filled bars) and OOD (hashed bars) performance evaluation of both performance and mis-classification and mis-segmentation detection. (Det.=Deterministic, VVL= Variational Visible Layers, VI=Variational Inference). Error is based on 95% confidence interval of runs trained with five different seeds.

Segmentation on ISIC and LIDC-IDRI, using a UNet architecture [25], shows a similar pattern to classification on smaller datasets. For ISIC segmentation performance, all models achieve an IoU within the range of 0.844 ± 0.006, except for VI, which has the lowest performans with an IoU of 0.76 ± 0.01. VI model, however, outperforms the rest in uncertainty failure detection at the cost of a lower *absolute* segmentation performance as measured by IoU and NLL. This is also evident in the qualitative examples in Fig. 3, where VI has a poor segmentation quality. For the ISIC dataset, ensemble models demonstrate strong OOD performance, though they come with a significant drawback: an approximately five-fold increase in the number of parameters compared to the VVL approaches. In contrast, the Post-VVL and Post-Last approaches display strong ID/OOD performance on both datasets. Qualitative visualization of predictions and uncertainty are shown in Fig. 3 for the best and worst models.

4 Conclusions and Recommendations

We conducted an evaluation of post- and joint-training approaches for integrating variational visible layers into existing model architectures. Our analysis, covering segmentation and classification of medical images in different imaging modalities, assessed predictive uncertainty under covariate shifts. Key findings include: (1) For small-scale datasets, post-hoc reparameterization of visible layers matches or exceeds the performance of variational inference (VI) and ensembles while using significantly fewer parameters; (2) on larger datasets, post-hoc reparameterization of visible or final layers achieve the top failure detection performance; and (3) across all experiments, VI and Sparse VI consistently demonstrate strong failure detection, though often at the cost of higher NLL. Based

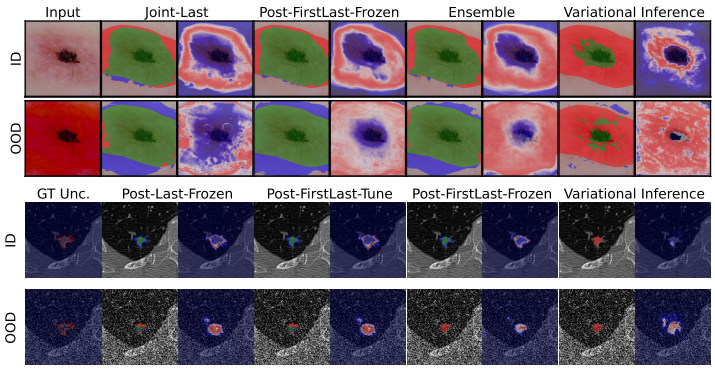


Fig. 3: ISIC, LIDC segmentation ID and OOD samples, with inputs, binary predictions are displayed with colors (first column) representing TP (green), FP (blue), and FN (red) predictions, and associated uncertainty (second column), showing regions of high (red), white (mid), and low (blue) uncertainty. Joint last shows very low uncertainty in OOD, VI displays high FN.

on these findings, we recommend post-hoc reparameterization of VVL layers for failure detection. For applications requiring fast inference, post-hoc reparameterization of the last layers is recommended. Future work should explore the scalability of these methods on datasets $> 10^5$ samples to further validate their effectiveness in large-scale applications, as well as explore threshold-independent metrics for failure detection.

Acknowledgments. We acknowledge the support of Le Fonds de recherche du Québec – Nature et technologies (FRQNT) for Doctoral Research Fellowship (DOI: 10.69777/352230), Natural Sciences and Engineering Research Council of Canada (NSERC) for NSERC Discovery grant, and Canada Research Chair in Shape Analysis in Medical Imaging.

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

References

1. Abboud, Z., Lombaert, H., Kadoury, S.: Sparse bayesian networks: Efficient uncertainty quantification in medical image analysis. In: MICCAI (2024)
2. Abdar, M., Pourpanah, F., Hussain, S., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76** (2021)
3. Armato III, S.G., McLennan, G., Bidaut, et al.: The lung image database consortium and image database resource initiative: a completed reference database of lung nodules on CT scans. Medical Physics **38**(2) (2011)
4. Armato III, S.G., McLennan, G., Bidaut, L., et al.: Data From LIDC-IDRI. The Cancer Imaging Archive (2015), Data set
5. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: ICML (2015)
6. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The Cancer Imaging Archive: Maintaining and Operating a Public Information Repository. Journal of Digital Imaging **26**(6) (2013)
7. Codella, N., Rotemberg, V., Tschandl, P., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration. arXiv preprint arXiv:1902.03368 (2019)
8. Daxberger, E., Nalisnick, E., Allingham, J.U., Antorán, J., Hernández-Lobato, J.M.: Bayesian deep learning via subnetwork inference. In: ICML (2021)
9. Farquhar, S., Smith, L., Gal, Y.: Try depth instead of weight correlations: Mean-field is a less restrictive assumption for variational inference in deep networks. In: Bayesian Deep Learning Workshop At NeurIPS (2020)
10. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: ICML (2016)
11. Harrison, J., Willes, J., Snoek, J.: Variational bayesian last layers. In: ICLR (2024)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on CVPR (2016)

13. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
14. Hinton, G.E., Van Camp, D.: Keeping the neural networks simple by minimizing the description length of the weights. In: Proceedings of the sixth annual conference on Computational learning theory (1993)
15. Huang, L., Ruan, S., Xing, Y., Feng, M.: A review of uncertainty quantification in medical image analysis: probabilistic and non-probabilistic methods. Medical Image Analysis (2024)
16. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR. vol. 28 (2014)
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS. vol. 30 (2017)
18. MacKay, D.J.: A practical bayesian framework for backpropagation networks. Neural computation **4** (1992)
19. Malinin, A.: Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment. Ph.D. thesis, University of Cambridge (2019)
20. Malinin, A., Band, N., Chesnokov, G., Gal, Y., Gales, M.J., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., et al.: Shifts: A dataset of real distributional shift across multiple large-scale tasks. arXiv preprint arXiv:2107.07455 (2021)
21. Neal, R.: Bayesian learning via stochastic dynamics. NeurIPS **5** (1992)
22. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. NeurIPS **32** (2019)
23. Prabhudesai, S., Hauth, J., Guo, D., et al.: Lowering the computational barrier: Partially Bayesian neural networks for transparency in medical imaging AI. Frontiers in Computer Science **5** (2023)
24. Radford, M.N.: Bayesian Learning for Neural Networks. Ph.D. thesis, University of Toronto (1995)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
26. Sharma, M., Farquhar, S., Nalisnick, E., Rainforth, T.: Do Bayesian neural networks need to be fully stochastic? In: AISTATS (2023)
27. 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) (2018)
28. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST V2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1) (2023)
29. Zeng, J., Lesnikowski, A., Alvarez, J.M.: The relevance of Bayesian layer positioning to model uncertainty in deep Bayesian active learning. In: Third workshop on Bayesian Deep Learning in NeurIPS (2018)
30. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. Meta-Radiology (2023)