Last Layer Laplacian Pseudocoresets for Robust Medical Image Analysis

Franciskus Xaverius Erick[1], Johanna Paula Müller[1], Zhe Li[1], and Bernhard Kainz[1,2]

[1] Image Data Exploration and Analysis Lab,Friedrich-Alexander University Erlangen-Nürnberg
[2] Department of Computing, Imperial College London

Abstract. Developing robust machine learning algorithms is of utmost importance for their applications to biomedical imaging applications. This issue is non-trivial, as networks are generally trained with datasets taken from relatively homogeneous samples dominated by statistically more probable disease classes, leading to unbalanced class distributions. One possible solution is to resolve the intrinsic biases towards certain dominating classes in the training datasets through more data collection with a more diverse sample, which is often prohibitively expensive. Another solution is to directly implement established uncertainty estimation measures for more robust predictions, which are nevertheless computationally demanding and insensitive to class imbalance. To address this issue, we propose a novel class-aware and uncertainty-aware pseudocoreset framework consisting of the following components: 1) An efficient framework with last layer Laplacian approximation 2) Class-aware calibration with error-based regularization, and 3) a Wasserstein distance-based regularization which explicitly imposes uncertainty-awareness. We evaluate our method for In-Distribution calibration, Out-of-Distribution inference, and class balance evaluations in two public skin cancer datasets taken from samples from different geographical location with differing skin colors. Our method outperforms various baseline uncertainty quantification and Bayesian pseudocoreset methods.

Keywords: Coresets · Uncertainty Estimation · OOD.

1 Introduction

The recent success of large deep learning models led to a substantial increase in training and deployments of models in various applications in clinical practice [30]. However, the expected benefits of deploying such models are severely hindered by two major limiting factors: the required number of labeled training data and the unavoidable class imbalance of the samples. Due to such limitations, deep learning models are prone to overfit specifically to the limited attributes found within the dominating subclasses, subsequently leading to overconfident, erroneous predictions. Failure to generalize to data distribution shifts can potentially be fatal for safety-critical applications such as medical imaging.

One medical imaging application in which the aforementioned variances occur is the automated detection of skin cancers, a globally relevant disease where approximately 14 million cases of new cancer cases are detected annually, 9.6 million of which lead to death [23]. In response to the pressing global mortality figures, communities worldwide strive to streamline the screening and detection of skin cancers. One possible solution is to adopt automated detection with handheld devices [22]. Skin cancers develop primarily on the outermost epidermis layer, which renders them visible to the naked eye. However, their morphological features may vary due to differing skin colors, skin textures, and preexisting confounding skin issues. These confounding features, while seemingly trivial for humans to distinguish, might easily be ignored by trained neural networks [3,13,28,1]. Moreover, samples from healthy patients or more common variants of skin lesions are more readily obtainable from the population than other variants, leading to class imbalances within available training data. The machine learning community has proposed various uncertainty estimation solutions to alleviate the lingering robustness problems of deep learning models. Methods such as Deep ensembles [18], MC-Dropout [10], and Bayesian neural networks [24,2,29] are prominent approaches that are intended to improve the in-distribution calibration and out-of-distribution performance across various datasets. However, often the computational demands for training and inference of these methods render lightweight, real-life applications infeasible as they require multiple trainings and inferences of modified networks with larger parameter counts. Recent works on Bayesian pseudocoresets highlight the possibility of condensing datasets into synthetic datasets (pseudocoresets) for lightweight uncertainty-estimation [16,17,27]. Nevertheless, these methods focus solely on networks without considering the potential class imbalance of the training data [15], with performances recorded on class-balanced natural image datasets such as CIFAR100 or ImageNet. Given the class imbalance in medical image datasets, we propose to perform a class-balanced pseudocoreset framework that minimizes the training and inference cost of a neural network's uncertainty estimation. This leads us to introduce a novel class- and uncertainty-aware Last Layer Laplacian Pseudocoreset (LLLP) framework with the following **contributions**: (1) We propose a pseudocoreset construction framework with neural networks equipped with a stochastic Laplacian last layer. By restricting the number of images per class and introducing explicit class-aware regularization terms, we ensure a compact yet balanced dataset that streamlines the training and inference of last-layer Laplacian neural networks. (2) We introduce an uncertainty-aware calibration regularization term by explicitly penalizing calibration errors incurred from the training process. This term guarantees that the pseudocoresets are not solely optimized on their in-distribution accuracy but also on their calibration quality. (3) We incorporate a Wasserstein-2 distance-aware divergence regularization term to minimize the divergence between the Laplace approximated posteriors obtained from the pseudocoresets and the original dataset. This term ensures that the improved uncertainty-estimation performance from the original dataset is transferred similarly to the pseudocoresets. (4) We conduct in-distribution (ID)

and out-of-distribution (OOD) inference experiments, as well as a class balance evaluation to demonstrate the benefit of our method in comparison to the other baseline methods.

2 Method

The primary aim of Bayesian pseudocoresets (BPC) is to extract synthetic, condensed datasets for more efficient training and inference of Bayesian neural networks (BNN) [21,16], with the main training objective of aligning the uncertainty estimation performance of the networks on the pseudocoresets to that on the original dataset. We denote the original, full dataset F with the image samples $x = \{x_i, x_2, ..., x_{|F|}\}$ and their corresponding labels $y = \{y_i, y_2, ..., y_{|F|}\}$, and the pseudocoreset C with the synthetic image samples $u = \{u_i, u_2, ..., u_{|C|}\}$ and labels $v = \{v_i, v_2, ..., v_{|C|}\}$ such that $|F| >> |C|$. Consider the parameters θ of a probabilistic encoder π. The optimal pseudocoreset C^* is obtained by minimizing the divergence between π_F, the posterior of the parameters conditioned to F, and π_C, the posterior of the parameters conditioned to C:

$$C^* = \arg\min_C D\left(\pi_F | \pi_C\right), \tag{1}$$

with the posterior terms computed as follows:

$$\begin{aligned} \pi_F &= \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{|F|} \log \pi(y_i|x_i, \theta)\right) \pi_0(\theta), \\ \pi_C &= \frac{1}{Z(u)} \exp\left(\sum_{i=1}^{|C|} \log \pi(v_i|u_i, \theta)\right) \pi_0(\theta), \end{aligned} \tag{2}$$

where $\pi_0(\theta)$ is the parameters' prior and $Z(x)$ is the margin likelihood term. As $Z(x) = \int_0^x \pi_0(\theta) \exp\left(\sum_{i=1}^{|F|} \log \pi(y_i|x_i, \theta)\right) d\theta$ is intractable, it poses a significant computational overload, which is why implementations had been limited to simpler, lower-dimensional data domains so far.

2.1 Efficient, Tractable Pseudocoreset Optimization

The existing purely Bayesian pseudocoreset (BPC) frameworks necessitate performing computationally expensive Markov Chain Monte Carlo(MCMC) sampling on the BNNs output to approximate the posteriors required for the synthetic dataset generation optimization [16]. To mitigate this, we utilize neural networks with last-layer Laplacian Approximation to compute the posteriors required for the optimization [7]. The last layer Laplacian approximation allows a scalable and efficient approximation of the posteriors as Gaussian distributions N of mean θ^*, which are the maximum-a-posteriori (MAP) weights, and the variance term Σ [7]. Following the notations used previously

in Eq. 2, the Laplacian approximation of the posterior is $\pi_F \approx N(\theta_F^*, \Sigma_F)$, where $\Sigma_F = \left(\nabla_\theta^2 \mathcal{L}(D;\theta)|_{\theta_F^*}\right)^{-1}$ is the inverse of the Hessian matrix of the loss computed at MAP. Therefore, Laplace approximations can be performed after training a normal deterministic neural network without compute-heavy probabilistic training or inferences that are otherwise required in other uncertainty estimation methods. With Laplace approximation, we can consider only the last-layer weights of a pre-trained network as probabilistic and perform the appropriate posterior update. Recent works have demonstrated that restricting stochasticity into the last layer of a neural network sufficiently delivers competitive uncertainty-estimation capability [8,7,26].

With the Laplace-approximated posteriors, we can perform the divergence minimization operation from Eq. 1 given appropriate divergence measures. We choose the 2-Wasserstein distance[9,25], which possesses a tractable form for Gaussian distributions.

$D\left(\pi_F|\pi_C\right) = W_2^2(\pi_F, \pi_C) = \left\|\theta_F^* - \theta_C^*\right\|^2 + Tr(\Sigma_F + \Sigma_C - 2(\Sigma_F^{1/2}\Sigma_F\Sigma_C^{1/2})^{1/2})$.

We further streamline the optimization procedure by considering a combined stochastic and deterministic approach. Prior BPC methods [16,17,27] generally adapt the trajectory matching (TM) framework [5] as a basis for the optimization objective in Eq. 1, particularly to accommodate the computationally expensive purely BNN base encoders. Trajectory matching is a dataset condensation framework optimized by aligning the training trajectories of networks on the original dataset and the condensed dataset. However, trajectory matching necessitates pre-training several expert teacher networks as checkpoints for training trajectories, which in combination with the purely Bayesian formulation, results in an overall expensive optimization and sampling inferences. Here, we adopt a combination of the stochastic Laplacian optimization component with a more efficient, distribution-matching optimization [31] with a Maximum Mean Discrepancy (MMD)-based objective[11],

$$\mathcal{L}_D(F, C) = \mathbb{E}_{\theta \sim P_\theta} \left| \frac{1}{|F|} \sum_{i=1}^{|F|} \psi_\theta(x_i) - \frac{1}{|C|} \sum_{j=1}^{|C|} \psi_\theta(v_j) \right|^2, \tag{3}$$

where ψ is the feature embedding function of the input samples. The deterministic term $\mathcal{L}_D$ encourages the minimization of the discrepancy between features of the original dataset and the pseudocoresets without pre-training multiple expert networks.

2.2 Dual Class-Aware Regularizations

Furthermore, we also introduce a dual class-aware distribution regularization term, taking into account both the predictive performance and calibration capability of the network for the different classes. Medical image datasets generally exhibit class imbalance, with healthy or common classes outweighing less commonly occurring disease classes. Consequently, classical data condensation methods might lead to a more unbalanced condensed dataset, as the diversity

of features of each class in the original dataset is further reduced to a mere number of samples per class in the condensed dataset. We use the cross-entropy loss-based class-aware features balancing regularization term $\mathcal{L}_{ACE}$ [32] with an additional classification calibration error term $\mathcal{L}_{ECE}$, thereby encouraging pseudocoresets with balanced features representation and calibration capability of the different classes, denoted as follows,

$$\mathcal{L}_{ACE} = \text{Acc}_{\phi}\mathcal{L}_{CE}(C) + \mathcal{L}_{ECE}(C), \tag{4}$$

where ϕ denotes a sampled model where accuracy (Acc), cross-entropy loss $\mathcal{L}_{CE}$, and expected calibration error are evaluated with real data.

2.3 Overall loss function and framework

The optimal pseudocoreset is determined by minimizing the combined loss functions,

$$C^* = \arg\min_{C}(D\left(\pi_F|\pi_C\right) + \mathcal{L}_D(F, C) + \lambda_{ACE}\mathcal{L}_{ACE}(C)). \tag{5}$$

The combined objective encourages extracting pseudocoresets with a more efficient predictive and uncertainty-estimation capability employing Laplace approximation while considering the class imbalance in the data.

3 Experimental Settings

Training Parameters. We conduct all our experiments with the ConvNet architecture, which is consistently used in prior BPC and dataset condensation works. We utilize the recommended training hyperparameters for DM [32] and Hessian optimization parameters for Laplace Approximation [7]. The training images undergo the default Differentiable Siamese Augmentation (DSA) procedure consisting of color jitters, flips, crops, cutouts, scaling, and rotation. We set the regularization term λ_{ACE} to 0.1.
Data. We train and construct pseudocoresets (PC) with the ISIC2019 skin lesion dataset [6]. ISIC2019 consists of 8 classes, with some class categories containing significantly more samples than others, thus showcasing a clear class imbalance. For OOD and class balance evaluation, we perform inference on the ASAN skin lesion dataset with 12 classes [12]. The ASAN skin dataset is sampled from a population of darker skin color in contrast to the ISIC2019 dataset, where samples are taken from a population with the majority having lighter skin color.
Evaluation Metrics. We consider predictive accuracy to assess the predictive performance of the PC. We also report on the following metrics for calibration, OOD performance, and class balance evaluations. **NLL** $\downarrow$: Negative log-likelihood between the prediction logits and the ground truth labels, with lower values indicating a better calibration performance. **ECE** $\downarrow$: Expected calibration error measures how well aligned the model's confidence to true positive

predictions across various accuracy bins. Lower values signify a better alignment. **OOD-AUROC** ↑ : OOD Area Under Operating Curve gauges the model's ability to distinguish between positive (ID) and negative classes (OOD), with higher values signifying better OOD performance. **BACC** ↑ : Balanced accuracy measure, which takes into account of unbalanced class samples [4]. **F1** ↑ : F1 measure across various classes. **GM** ↑ : Geometric mean of recalls across various classes.

Compared Methods. For in-distribution and OOD evaluations, we compare our method with 3 previously developed BPC methods (BPC-F, BPC-W, and BPC-R) and the deterministic framework DM without and with additional uncertainty estimation baselines, namely Deep Ensembles, MC-Dropout, and Spectral-normalized Neural Gaussian Process(SNGP) [19].

4 Results and Discussions

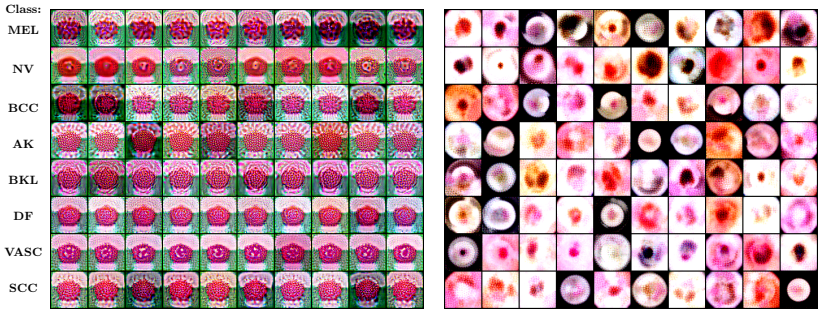
In-Distribution (ID) and Out-of-Distribution Experiments. For the ID and OOD analysis, we first extracted PC with 1,10, and 20 instances per class (ipcs). We trained ConvNets with the extracted PC for 200 epochs and analyzed their corresponding inference performance on the test datasets. For uncertainty estimation baselines, we trained an ensemble of $k = 10$ and MC-Dropout with a dropout rate of 0.1. The results are summarized in Table 1. The explicit Bayesian optimization of the BPCs results in improved calibrations at the cost of diminished predictive and OOD inference. Applying dropouts and SNGP training to deterministic coresets results in strongly diminished performances, highlighting the complexity of performing classic uncertainty estimation methods in small data domains. Our method with combined stochastic and deterministic terms, showcases a balanced predictive, calibration, and OOD inference.

Class Balance Experiments. We trained ConvNets with PC extracted with 10 ipcs and evaluated the class imbalance performance with the best performing methods from the previous experiments. The findings summarized in 2 reveal in conjunction with the OOD experiment results, that the BPCs and uncertainty estimation methods potentially suffer from overfitting to the dominant classes, especially for smaller medical datasets with less diversity of features. Adopting the class-aware regularization terms facilitates balanced class sampling for an effective deterministic (predictive) and stochastic (calibration, OOD) optimization of the PC.

Qualitative Analysis. We visually compare the generated PC for both BPC and LLLP shown in Figure 1. BPC displays less diverse, artificial image features, while our method results in compact datasets of more diverse and more realistic features. While BPC performs well with natural image datasets with balanced class samples and diverse features, its potential is limited when used on smaller, more imbalanced medical image datasets. We further emphasize the importance of non-Bayesian optimization terms and class-balancing terms for PC generations in such domains.

Table 1: In-distribution calibration and OOD experiment results for the differing BPCs and uncertainty estimation methods. The networks are trained with generated PC with 3 different instances per classes(ipcs), with best performing metrics indicated in **bold**.

Method	ipc	**Acc.**	**NLL**(↓)	**ECE**(↓)	**AUROC**(↑)
BPC-fKL	1	$0.305_{\pm 0.013}$	$2.100_{\pm 0.030}$	$\mathbf{0.355_{\pm 0.008}}$	$0.617_{\pm 0.018}$
	10	$0.448_{\pm 0.010}$	$1.937_{\pm 0.009}$	$0.241_{\pm 0.002}$	$0.625_{\pm 0.006}$
	20	$0.479_{\pm 0.007}$	$1.825_{\pm 0.008}$	$0.220_{\pm 0.003}$	$0.656_{\pm 0.004}$
BPC-rKL	1	$0.298_{\pm 0.015}$	$2.210_{\pm 0.026}$	$0.371_{\pm 0.010}$	$0.609_{\pm 0.023}$
	10	$0.443_{\pm 0.013}$	$2.054_{\pm 0.010}$	$0.272_{\pm 0.001}$	$0.621_{\pm 0.004}$
	20	$0.475_{\pm 0.005}$	$1.941_{\pm 0.005}$	$0.237_{\pm 0.002}$	$0.648_{\pm 0.001}$
DM	1	$0.327_{\pm 0.024}$	$2.328_{\pm 0.023}$	$0.420_{\pm 0.035}$	$0.613_{\pm 0.015}$
	10	$0.477_{\pm 0.006}$	$2.144_{\pm 0.005}$	$0.352_{\pm 0.002}$	$0.630_{\pm 0.003}$
	20	$\mathbf{0.510_{\pm 0.003}}$	$1.986_{\pm 0.002}$	$0.259_{\pm 0.001}$	$0.665_{\pm 0.001}$
DM-Ensembles	1	$0.320_{\pm 0.015}$	$2.193_{\pm 0.017}$	$0.360_{\pm 0.012}$	$0.648_{\pm 0.003}$
	10	$0.472_{\pm 0.005}$	$2.006_{\pm 0.005}$	$0.239_{\pm 0.001}$	$0.667_{\pm 0.002}$
	20	$0.501_{\pm 0.002}$	$1.877_{\pm 0.003}$	$0.218_{\pm 0.001}$	$0.702_{\pm 0.001}$
DM-MC Dropout	1	$0.157_{\pm 0.016}$	$2.257_{\pm 0.031}$	$0.408_{\pm 0.022}$	$0.622_{\pm 0.008}$
	10	$0.435_{\pm 0.008}$	$2.167_{\pm 0.012}$	$0.339_{\pm 0.007}$	$0.638_{\pm 0.007}$
	20	$0.487_{\pm 0.007}$	$1.979_{\pm 0.005}$	$0.256_{\pm 0.005}$	$0.672_{\pm 0.004}$
DM-SNGP	1	$0.113_{\pm 0.028}$	$2.401_{\pm 0.033}$	$0.405_{\pm 0.040}$	$0.631_{\pm 0.010}$
	10	$0.254_{\pm 0.014}$	$2.208_{\pm 0.027}$	$0.389_{\pm 0.011}$	$0.651_{\pm 0.013}$
	20	$0.386_{\pm 0.008}$	$2.050_{\pm 0.012}$	$0.283_{\pm 0.004}$	$0.678_{\pm 0.006}$
LLLP(ours)	1	$\mathbf{0.332_{\pm 0.016}}$	$\mathbf{2.017_{\pm 0.026}}$	$0.359_{\pm 0.006}$	$\mathbf{0.662_{\pm 0.007}}$
	10	$\mathbf{0.487_{\pm 0.008}}$	$\mathbf{1.862_{\pm 0.011}}$	$\mathbf{0.225_{\pm 0.001}}$	$\mathbf{0.680_{\pm 0.005}}$
	20	$0.507_{\pm 0.004}$	$\mathbf{1.760_{\pm 0.005}}$	$\mathbf{0.210_{\pm 0.002}}$	$\mathbf{0.731_{\pm 0.002}}$



(a) BPC (b) LLLP (ours)

Fig. 1: Generated PC for for ipc=10. Rows correspond to the classes **SCC**: Squamous Cell Carcinoma; **VASC**: Vascular Lesion; **DF**: Dermatofibroma; **BKL**: Benign Keratosis-Like Lesion; **AK**: Actinic Keratosis; **BCC**: Basal Cell Carcinoma; **NV**: Nevus; **MEL**: Melanoma.

Table 2: Class imbalance experiments results for the various baselines and our method. Experiments were run with ipc of 10. Best performing metrics are indicated in **bold**.

	BPC-fKL	BPC-rKL	DM	DM-Ensembles	**LLLP(ours)**
Bal.Acc. (↑)	$0.253_{\pm 0.002}$	$0.230_{\pm 0.003}$	$0.386_{\pm 0.002}$	$0.383_{\pm 0.001}$	$\mathbf{0.401_{\pm 0.001}}$
Macro F1(↑)	$0.207_{\pm 0.001}$	$0.193_{\pm 0.002}$	$0.285_{\pm 0.002}$	$0.282_{\pm 0.001}$	$\mathbf{0.292_{\pm 0.002}}$
GM(↑)	$0.469_{\pm 0.002}$	$0.432_{\pm 0.001}$	$0.529_{\pm 0.003}$	$0.520_{\pm 0.002}$	$\mathbf{0.573_{\pm 0.001}}$

Ablation study. We investigated the effect of each component in our method, namely Last-Layer Laplace (LL), its corresponding regularization component (LL-R), and the class-aware calibration regularization component(ACE-R) to the balanced accuracy, calibration error, and OOD inference. Our finding, summarized in Table 3, highlight the contribution of each component to class-balanced, robust inference. For cross architecture generalization experiments summarized in Table 4, we extracted the PC with ConvNet and performed in-distribution inference with ConvNet, AlexNet, and ResNet-18. Table 4 shows that LLLP delivers generalize well through architectures.

Table 3: LLLP component ablations.

LL	LL-R	ACE-R	**BAcc.**(↑)	**ECE** (↓)	**OOD**(↑)
-	-	-	0.285	0.352	0.630
✓	-	-	0.285	0.280	0.655
✓	✓	-	0.243	0.291	0.653
✓	-	✓	0.246	0.257	0.661
✓	✓	✓	**0.292**	**0.225**	**0.680**

Table 4: Accuracy Cross architecture generalization.

	BPC-f	BPC-r	LLLP
ConvNet	0.448	0.436	**0.487**
AlexNet	0.385	0.359	**0.402**
ResNet-18	0.363	0.344	**0.380**

5 Conclusion

We introduced an efficient pseudocoreset framework consisting of stochastic last-layer Laplace approximations to induce distance-awareness and a deterministic MMD-based pseudocoreset optimization term to facilitate faster pseudocoreset generation with satisfactory predictive performances. We also imbued a class-aware calibration regularization term to promote class-balanced learning. Our findings from the downstream performance in safety-critical medical domains with limited training samples and class imbalance reveal that our framework promotes robust calibration, OOD inference, and a more class-balanced performance. For our future works, we will explore further strategies to implement fairness measures into our pseudocoreset generation and investigate the correlation between uncertainty estimation and fairness. We would also

like to perform more robust uncertainty quantification and out-of-distribution evaluations as highlighted in [14], with other metrics such as KDE-ECE [20].. Our method implementation can be found in the following repository https://github.com/fx-erick/LLLP.

Acknowledgments. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR projects b143dc and b180dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Additional support was also received by the ERC - project MIA-NORMAL 101083647, DFG 513220538, 512819079, and by the state of Bavaria (HTA).

Disclosure of Interests. The authors have no competing interests for this work.

References

1. Skin analytics. https://skin-analytics.com/, accessed: 2025-02-14
2. Abdar, M., Samami, M., Mahmoodabad, S.D., Doan, T., Mazoure, B., Hashemifesharaki, R., Liu, L., Khosravi, A., Acharya, U.R., Makarenkov, V., et al.: Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. Computers in biology and medicine p. 104418 (2021)
3. Bevan, P.J., Atapour-Abarghouei, A.: Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. In: Kamnitsas, K., Koch, L., Islam, M., Xu, Z., Cardoso, J., Dou, Q., Rieke, N., Tsaftaris, S. (eds.) Domain Adaptation and Representation Transfer. pp. 1–11. Springer Nature Switzerland, Cham (2022)
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. pp. 3121–3124 (2010). https://doi.org/10.1109/ICPR.2010.764
5. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories (2022)
6. Codella, N.C.F., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S.W., Gutman, D.A., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M.A., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). CoRR **abs/1902.03368** (2019), http://arxiv.org/abs/1902.03368
7. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux – effortless bayesian deep learning (2022)
8. Daxberger, E., Nalisnick, E., Allingham, J.U., Antoran, J., Hernandez-Lobato, J.M.: Bayesian deep learning via subnetwork inference. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 2510–2521. PMLR (18–24 Jul 2021)
9. Dowson, D.C., Landau, B.V.: The fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis **12**, 450–455 (1982)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning (2016)

11. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**(25), 723–773 (2012), http://jmlr.org/papers/v13/gretton12a.html
12. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. Journal of Investigative Dermatology **138**(7), 1529–1538 (2018). https://doi.org/https://doi.org/10.1016/j.jid.2018.01.028
13. Hauser, K., et al.: Explainable artificial intelligence in skin cancer recognition: A systematic review. European Journal of Cancer **167**, 54–69 (2022). https://doi.org/https://doi.org/10.1016/j.ejca.2022.02.025, https://www.sciencedirect.com/science/article/pii/S095980492200123X
14. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification (2023), https://arxiv.org/abs/2211.15259
15. Khan, S., Hayat, M., Zamir, W., Shen, J., Shao, L.: Striking the right balance with uncertainty (2019), https://arxiv.org/abs/1901.07590
16. Kim, B., Choi, J., Lee, S., Lee, Y., Ha, J.W., Lee, J.: On divergence measures for bayesian pseudocoresets (2022), https://arxiv.org/abs/2210.06205
17. Kim, B., Lee, H., Lee, J.: Function space bayesian pseudocoreset for bayesian neural networks (2023), https://arxiv.org/abs/2310.17852
18. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles (2017)
19. Liu, J.Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., Lakshminarayanan, B.: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness (2020)
20. Maier-Hein, L., et al.: 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
21. Manousakas, D., Xu, Z., Mascolo, C., Campbell, T.: Bayesian pseudocoresets. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 14950–14960. Curran Associates, Inc. (2020)
22. Modaragama, S.: Decision intelligence: Transformational developments in skin cancer screening using mobile applications. Available at SSRN 3687522 (2023)
23. National Cancer Institute: Cancer Statistics, accessed: 2024-03-03. https://www.cancer.gov/ (2024)
24. Neal, R.M.: Bayesian learning for neural networks, vol. 118. Springer Science & Business Media (2012)
25. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra and its Applications **48**, 257–263 (1982)
26. Sharma, M., Farquhar, S., Nalisnick, E., Rainforth, T.: Do bayesian neural networks need to be fully stochastic? In: Ruiz, F., Dy, J., van de Meent, J.W. (eds.) Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 206, pp. 7694–7722. PMLR (25–27 Apr 2023)
27. Tiwary, P., Shubham, K., Kashyap, V.V., P, P.A.: Bayesian pseudo-coresets via contrastive divergence (2024), https://arxiv.org/abs/2303.11278
28. Wen, D., Khan, S.M., Xu, A.J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A.K., Liu, X., et al.: Characteristics of publicly available skin cancer image datasets: a systematic review. The Lancet Digital Health **4**(1), e64–e74 (2022)

29. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. In: Advances in Neural Information Processing Systems. vol. 33 (2020)
30. Wu, K., Wu, E., Theodorou, B., Liang, W., Mack, C., Glass, L., Sun, J., Zou, J.: Characterizing the clinical adoption of medical ai devices through us insurance claims. NEJM AI **1**(1), AIoa2300030 (2024)
31. Zhao, B., Bilen, H.: Dataset condensation with distribution matching (2022)
32. Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation (2023), https://arxiv.org/abs/2307.09742