Empirical Bayesian Methods and BNNs for Medical OOD Detection

Kevin Raina[1][0000−0002−6240−9675]

University of Ottawa, Ontario, Canada
krain033@uottawa.ca

Abstract. Bayesian Neural Networks (BNNs) are a principled way to incorporate epistemic uncertainty into deep learning, and they play a significant role in out-of-distribution (OOD) detection, especially in settings where estimating predictive uncertainty is crucial. Empirical Bayesian methods, which initialize priors and surrogate posteriors from the weights of pretrained deterministic neural networks, can help in OOD detection by providing well-informed models, thereby bridging the gap between data-driven learning and principled uncertainty estimation — especially when true Bayesian inference is intractable. In this work, the empirical Bayes method MOdel Priors with Empirical Bayes using Deterministic neural networks (MOPED) is adapted to include a Gaussian mixture prior. Experiments on the medical datasets D7P and BreastMNIST, with OOD images containing artefacts such as rulers and annotations, demonstrate marked improvements in OOD detection from the proposed prior with predictive entropy as the score. The proposed empirical Bayes methods also performs on par with state-of-the art OOD measures.

Keywords: Empirical bayes · Out-of-distribution detection · Bayesian neural networks.

1 Introduction

Deep neural networks (DNNs) have become a cornerstone of modern medical imaging analysis, demonstrating exceptional performance in a variety of classification tasks across domains such as radiology, pathology, and dermatology [21]. Their ability to automatically learn hierarchical feature representations from raw image data has significantly reduced the need for manual feature engineering, allowing for more accurate and scalable diagnostic tools. In particular, convolutional neural networks (CNNs), a class of DNNs, have achieved near-human or even superhuman performance in detecting diseases such as diabetic retinopathy, skin cancer, and pneumonia from medical images like fundus photographs, dermoscopic images, and chest X-rays, respectively [3,5,21]. These advances not only promise to enhance clinical decision-making and early disease detection but also offer potential solutions to medical resource shortages in underserved regions. As such, DNN-based methods are becoming increasingly central to the development of robust, efficient, and interpretable systems for medical image classification.

Despite the remarkable success of deep neural networks (DNNs) in medical image classification, their reliability in real-world clinical settings can be compromised by out-of-distribution (OOD) inputs—samples that differ significantly from the training data. In the medical domain, OOD inputs often arise from imaging artefacts, rare diseases, unusual anatomical variations, or scanner-specific distortions that were not adequately represented during training. These anomalous inputs can lead DNNs to produce confidently incorrect predictions, posing serious risks in high-stakes clinical environments [6, 20]. For instance, studies have shown that DNNs trained on chest X-rays may misclassify images with surgical implants, motion blur, or contrast variations, despite these features being irrelevant or misleading to the diagnostic task [26]. Consequently, the ability to detect and appropriately handle OOD inputs is critical for ensuring the safe and trustworthy deployment of AI models in medical imaging workflows.

Bayesian Neural Networks (BNNs) have emerged as a promising framework for addressing the limitations of standard DNNs, particularly in safety-critical applications like medical imaging [11, 12, 18]. Unlike conventional networks that produce point estimates, BNNs model uncertainty by treating network weights as probability distributions, allowing them to capture both epistemic uncertainty—which is model uncertainty due to limited training data—and aleatoric uncertainty, which arises from inherent noise in the input data [10]. This principled approach to uncertainty quantification is especially valuable when encountering OOD inputs, as high epistemic uncertainty often signals a lack of familiarity with the input, prompting caution in clinical interpretation. Predictive uncertainty derived from BNNs can thus serve as a powerful signal for flagging unreliable predictions, triaging ambiguous cases, or guiding human-in-the-loop review systems. In the context of medical imaging, leveraging such uncertainty-aware models can substantially improve decision confidence, model interpretability, and ultimately, patient safety.

Empirical Bayesian methods, such as MOdel Priors with Empirical Bayes using DNNs (MOPED) [11], offer a practical and effective approach to improving OOD detection in Bayesian Neural Networks by leveraging information from pretrained deterministic networks to define informative weight priors [11]. Unlike standard Bayesian initializations that often rely on vague or uninformative priors (e.g., isotropic Gaussians), MOPED uses the pretrained weights to construct data-informed priors, effectively anchoring the Bayesian model to a well-performing solution. This strategy enhances epistemic uncertainty estimation, especially in regions far from the training distribution, making the model more sensitive to OOD samples. In medical imaging, where subtle shifts in input distributions—due to demographic, device, or acquisition differences—can be clinically significant, such improved uncertainty modeling directly translates to more robust and calibrated decision-making. As a result, empirical Bayesian techniques like MOPED bridge the gap between high-performance deterministic models and uncertainty-aware Bayesian frameworks, offering a compelling tool for safer AI deployment in healthcare

This work proposes an extension of the MOPED framework by incorporating a Gaussian mixture instead of a single Gaussian prior on each network weight. A Gaussian mixture prior introduces greater flexibility than a single Gaussian by allowing the posterior to adapt to multimodal structures in the weight space, which may better reflect the complex inductive biases inherent in pretrained networks. When applied in an empirical Bayes setting, this richer prior formulation can encode both local variations and global uncertainty more effectively, potentially leading to sharper epistemic uncertainty estimates in ambiguous or OOD regions. In high-stakes domains such as medical imaging, this added expressiveness may yield better detection of unfamiliar or artefactual inputs, without compromising the strong inductive performance of pretrained deterministic models. By grounding the prior in observed weight statistics while expanding its capacity to model uncertainty, this approach aims to strike a balance between Bayesian rigor and practical utility for robust, uncertainty-aware classification. The particular applications of interest are the D7P (dermatology) [9] and BreastMNIST (ultrasound) [25] datasets, where OOD data contain rulers and annotations respectively.

2 Background: Bayesian Neural Networks and Predictive Uncertainty

Bayesian Neural Networks (BNNs) extend standard neural networks by placing probability distributions over their weights, thereby enabling explicit modeling of uncertainty in predictions. Let ω denote the set of neural network weights. In a Bayesian formulation, a prior distribution $p(\omega)$ is defined over ω, and the goal is to compute the posterior distribution given training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ as:

$$p(\omega \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \omega) p(\omega)}{p(\mathcal{D})}, \tag{1}$$

where $p(\mathcal{D} \mid \omega)$ is the likelihood and $p(\mathcal{D})$ is the marginal likelihood or evidence. This posterior is generally intractable for deep networks due to the high-dimensional and nonlinear nature of ω.

To address this, *mean-field variational inference* is commonly used to approximate the posterior. A simpler, factorized variational distribution $q(\omega)$ is introduced, typically assumed to be a fully factorized Gaussian:

$$q(\omega) = \prod_j \mathcal{N}(\omega_j \mid \mu_j, \sigma_j^2), \tag{2}$$

and optimized by minimizing the Kullback–Leibler divergence between $q(\omega)$ and the true posterior. This is equivalent to maximizing the *evidence lower bound* (ELBO):

$$\mathcal{L}_{\text{ELBO}} = E_{q(\omega)}[\log p(\mathcal{D} \mid \omega)] - \text{KL}[q(\omega) \,\|\, p(\omega)]. \tag{3}$$

Once trained, BNNs yield a predictive distribution for a new input x^* by marginalizing over the approximate posterior:

$$p(y^* \mid x^*, \mathcal{D}) = \int p(y^* \mid x^*, \omega)\, q(\omega)\, d\omega. \tag{4}$$

In practice, this integral is approximated via Monte Carlo sampling, by drawing T samples $\{\omega^{(t)}\}_{t=1}^{T}$ from $q(\omega)$:

$$p(y^* \mid x^*, \mathcal{D}) \approx \frac{1}{T} \sum_{t=1}^{T} p(y^* \mid x^*, \omega^{(t)}). \tag{5}$$

A powerful measure of predictive uncertainty is the *predictive entropy*:

$$\mathcal{H}[y^* \mid x^*, \mathcal{D}] = -\sum_{c} p(y_c^* \mid x^*, \mathcal{D}) \log p(y_c^* \mid x^*, \mathcal{D}), \tag{6}$$

which captures both *aleatoric* uncertainty (from noise in the data) and *epistemic* uncertainty (from uncertainty in the model parameters), as described in [4, 10]. High entropy indicates uncertain predictions, making this score particularly useful for detecting out-of-distribution (OOD) samples or ambiguous cases in medical imaging. In all MOPED experiments, predictive entropy is the applied score.

3 Empirical Bayesian Priors: MOPED and Gaussian Mixture Extensions

3.1 MOPED: Model Priors with Empirical Bayes

MOPED [11] is an empirical Bayesian method that enables Bayesian neural networks (BNNs) to incorporate knowledge from deterministic pretrained models by defining *informative Gaussian priors* over the model weights. Rather than adopting standard zero-centered or uninformative priors, MOPED sets the prior mean to the value of a pretrained deterministic model:

$$p(\omega_i) = \mathcal{N}(\omega_i \mid \mu_i^{\text{pre}}, \sigma_i^2), \tag{7}$$

where μ_i^{pre} is the pretrained weight value and σ_i^2 is the prior variance, often defined via a heuristic or treated as a hyperparameter.

Crucially, MOPED also controls the *initialization of the surrogate posteriors* $q(\omega_i) = \mathcal{N}(\omega_i \mid \mu_i^q, {\sigma_i^q}^2)$, used in variational inference. Rather than initializing the posterior variance with a fixed value, MOPED follows a *delta-scaled initialization* scheme. Specifically, the surrogate posterior is initialized as:

$$\mu_i^q = \mu_i^{\text{pre}}, \quad \log \sigma_i^q = \log \delta + \log |\mu_i^{\text{pre}}|, \tag{8}$$

where $\delta > 0$ is a small scalar hyperparameter controlling the spread of the posterior around the pretrained mean. This strategy ties the uncertainty in each weight to its magnitude in the pretrained model, under the intuition that larger weights may reflect more confident parameters, and thus should have proportionally calibrated uncertainty.

By leveraging pretrained weights in both the prior and the initialization of the variational posterior, MOPED improves the efficiency and quality of posterior inference in Bayesian neural networks, particularly under limited data and in uncertainty-sensitive tasks such as OOD detection.

3.2 A Gaussian Mixture Prior for Bayesian Neural Networks

Consider an extension to the empirical Bayes approach used in MOPED by replacing the single Gaussian prior with a *Gaussian Mixture Model* (GMM) prior over the weights. The prior for each independent weight ω_i is modeled as a mixture of two Gaussians:

$$p(\omega_i) = \pi \cdot \mathcal{N}(\omega_i \mid \mu_i^{(1)}, \sigma_i^{(1)\,2}) + (1 - \pi) \cdot \mathcal{N}(\omega_i \mid \mu_i^{(2)}, \sigma_i^{(2)\,2}), \tag{9}$$

where $\pi \in [0, 1]$ is the mixture weight (hyperparameter), and each component has its own mean and variance. Two configurations are explored for this Gaussian mixture prior:

1. **Dual-Pretrained Initialization**: Both components are initialized using the pretrained weight μ_i^{pre}, but are assigned different variances, allowing the model to represent both high-confidence and exploratory uncertainty centered around the same mean.
2. **Hybrid Initialization**: One component is centered at the pretrained mean μ_i^{pre} with a small variance (as in MOPED), while the second component is initialized with a zero mean and unit variance:

   $$\mu_i^{(1)} = \mu_i^{\text{pre}}, \quad \sigma_i^{(1)} = \delta \cdot |\mu_i^{\text{pre}}|, \qquad \mu_i^{(2)} = 0, \quad \sigma_i^{(2)} = 1.$$

   This allows the model to hedge between learned inductive biases and a non-informative component, enabling better generalization under distribution shift or ambiguity.

This mixture formulation provides a more expressive prior that can capture multi-modal beliefs over weights and offers greater flexibility in uncertainty modeling. As with MOPED, the surrogate posterior is initialized using delta-scaled rules based on the pretrained weights.

4 Experimental Setting

4.1 Datasets and Out-of-Distribution Detection Task

This study adopts the experimental setup proposed by Anthony and Kamnitsas [2] for evaluating out-of-distribution (OOD) detection in medical image classification, with the key difference that synthetic artefact augmentations are excluded. The focus is on naturally occurring artefacts, such as rulers and annotation markers, which commonly appear in clinical image data and may indicate a distributional shift not seen during training.

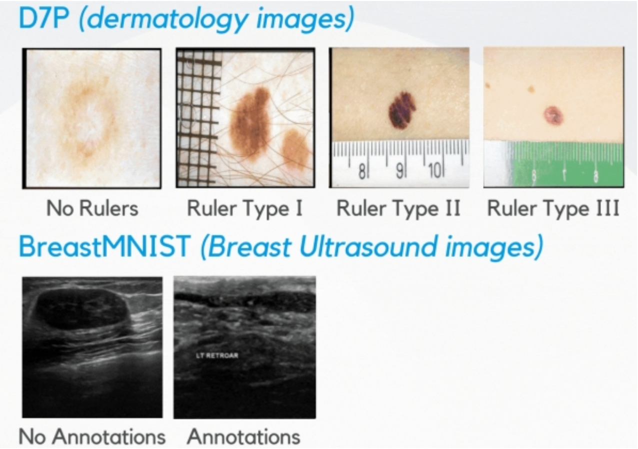

Fig. 1: Example medical images (in-distribution and OOD) from D7P and BreastMNIST. Figure included from Anthony and Kamnitsas [2] with permission from the authors.

Two medical imaging datasets are used for in-distribution (ID) training and evaluation:

- **D7P**: A digital pathology dataset comprising histopathological image patches across seven diagnostic categories. It captures diverse tissue morphologies and is used for multi-class classification tasks.
- **BreastMNIST**: A dataset of breast ultrasound images annotated for benign and malignant cases. It represents a different imaging modality (ultrasound) compared to D7P (histopathology), but is still treated as in-distribution for the purposes of evaluating model confidence under known conditions.

Models are trained and validated on clean, artefact-free subsets of these datasets. The OOD inputs are composed of naturally occurring artefact-containing images extracted from the same datasets. These include visual anomalies such as:

- **Rulers**: Measurement guides overlaid during the imaging process.
- **Annotations**: Text labels, arrows, or segmentation markings embedded in the image.

Such artefacts are not present in the training data and can significantly alter the model's feature representation, potentially leading to unreliable predictions. Figure 1 depicts examples of images with and without artefacts for both datasets. These inputs are withheld during training and used only at test time to evaluate OOD detection.

4.2 BNN Training Details

All Bayesian Neural Networks (BNNs) were trained using the Adam optimizer with a learning rate of 5×10^{-5} and a batch size of 32. Training was run for a fixed 500 epochs, and the model was saved whenever validation accuracy improved. Models were trained on 90 percent of the artefact-free in-distribution dataset, with 10 percent used as held-out ID test examples [2]. During training, a single Monte Carlo sample was used to estimate the evidence lower bound (ELBO), while 100 samples were used at test time for evaluation. The variational posterior means were initialized from pretrained weights, and the posterior variances were initialized using a fixed MOPED delta value of 0.01. For standard MOPED, the prior was a diagonal Gaussian with mean equal to the pretrained weights and variance set to 1.0.

In the Dual Gaussian Mixture prior variant, both mixture components were Gaussians centered at the pretrained weight, with variances of 0.1 and 1.0, respectively. The mixture parameter $\pi = 0.25$ is applied to the narrow (0.1 variance) component. In the Hybrid variant, one component was centered at zero and the other at the pretrained weights; both used variance 1.0. The mixture parameter $\pi = 0.25$ is applied to the pretrained component.

Using a mixture weight of $\pi = 0.25$ assigns moderate prior belief to the pretrained component without overwhelming the model's ability to adapt. This value reflects a reasonable balance: it gives the model flexibility to diverge from pretrained weights when the data demands it (via the 75 percent weight on the broader or zero-centered component), while still retaining some inductive bias toward known useful features. Empirically, this balance has been found to support generalization, as measured by improved validation accuracy on in-distribution data.

5 Results

OOD performance data for the proposed empirical Bayesian methods and existing methods reported in Anthony and Kamnitsas [2] are shown in Table 1. Each AUC score represents the mean over 5 independently trained models, each initialized with a different random seed. On the D7P (ruler OOD) benchmark, Dual achieves AUC_{OOD} scores of 78.6 (ResNet18) and 76.3 (VGG16), while Hybrid reaches 80.0 and 74.4. On BreastMNIST (annotator OOD), Dual scores 79.2 (ResNet18) and 76.3 (VGG16), with Hybrid close behind at 75.8 and 74.4. Both methods outperform the baseline MOPED model substantially. Compared to existing *confidence-based methods*, which generally yield lower AUCs, and *feature-based methods* such as RMS and GRAM that show variable performance, empirical Bayesian approaches provide strong and consistent improvements. The proposed empirical Bayes methods demonstrate competitive performance when compared to the leading Multi-branch Mahalanobis (MBM) method.

The improved results can be attributed to the structured uncertainty modeled by the mixture priors. This enables the methods to better represent multiple plausible weight configurations and capture epistemic uncertainty, which is particularly important for small-scale medical imaging datasets prone to overfitting.

OOD-D method	**D7P (ruler OOD)**		**BreastMNIST (anno. OOD)**	
	ResNet18	**VGG16**	**ResNet18**	**VGG16**
	AUC_{OOD}	AUC_{OOD}	AUC_{OOD}	AUC_{OOD}
Confidence-based Methods				
MCP [6]	49.3	51.9	55.8	52.4
SE [6]	49.5	52.8	55.8	51.4
MLS [7]	48.6	51.5	57.9	52.4
Energy Score [17]	48.5	51.5	57.6	51.9
MCDP-MCP [19]	49.3	52.0	55.8	51.9
MCDP-PE [19]	49.5	51.7	56.7	50.3
MCDP-MI [19]	49.5	51.7	56.7	50.3
DE-MCP [13]	49.9	52.7	56.0	53.3
GradNorm [8]	49.4	51.9	53.8	53.2
ODIN* [16]	64.6	52.0	58.7	53.6
ReAct* [24]	67.2	61.5	60.2	58.0
DICE* [15]	68.5	57.7	58.0	59.1
Feature-based Methods				
Mahal. Score [14]	76.9	72.5	77.1	72.5
MBM [1]	**80.7**	73.8	77.4	**76.8**
RMS [22]	70.2	60.5	70.5	52.7
GRAM [23]	53.6	72.3	63.6	71.3
Empirical Bayesian Methods				
MOPED [11]	64.8	49.8	72.1	67.9
MOPED GMM Dual	78.6	**76.3**	**79.2**	76.3
MOPED GMM Hybrid	80.0	74.4	75.8	74.4

Table 1: AUC_{OOD} scores for different OOD detection methods. *Feature-based* and *confidence-based* statistics are retrieved from Anthony and Kamnitsas [2]. The best performing method (specified by architecture and OOD task is bolded column-wise. * methods incorporate OOD data for hyperparameter tuning.

6 Future Work

Bayesian Neural Networks are strong contenders in uncertainty quantification applications such as OOD detection. Empirical Bayesian methods are practically significant for BNN training as they incorporate prior model information.

This work proposes novel empirical Bayesian methods for improved and state-of-the-art OOD detection in the context of benchmark medical imaging datasets. Specifically, the MOPED method is adapted to include a Gaussian mixture prior, from which two experimental settings are derived; one that incorporates a slab and spike prior on the model weights for flexible learning, and another that balances the effects of model information and regularization. Further investigations could include experimentation with more components in the prior, diversifying dataset applications, and testing novel Bayesian scores.

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

References

1. Anthony, H., Kamnitsas, K.: On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, pp. 136–146. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-44336-7_14,
2. Anthony, H., Kamnitsas, K.: Evaluating reliability in medical dnns: A critical analysis of feature and confidence-based ood detection. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 160–170. Springer (2024)
3. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)
4. 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)
5. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama **316**(22), 2402–2410 (2016)
6. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (ICLR) (2017), https://arxiv.org/abs/1610.02136
7. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Scaling out-of-distribution detection for realistic settings. In: International Conference on Learning Representations (ICLR) (2020), https://arxiv.org/abs/1911.11132
8. Jeong, S.W., Kim, S.J., Kwak, N.: Consistency regularization for out-of-distribution detection. In: NeurIPS (2021), https://arxiv.org/abs/2102.12231
9. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
10. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
11. Krishnan, R., Subedar, M., Tickoo, O.: Specifying weight priors in bayesian deep neural networks with empirical bayes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4477–4484 (2020)

12. Kwon, Y., Won, J.H., Kim, B.J., Paik, M.C.: Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In: Medical Imaging with Deep Learning (2022)
13. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017), https://arxiv.org/abs/1612.01474
14. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018), https://arxiv.org/abs/1807.03888
15. Li, S., Jiang, J., Zhang, R., Li, L., Li, S.: Dice: Detecting out-of-distribution samples via classifier ensembles. In: CVPR (2021), https://arxiv.org/abs/2103.03444
16. Liang, S., Li, Y., Srikumar, V.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (ICLR) (2018), https://arxiv.org/abs/1706.02690
17. Liu, K., Li, L., Li, S., Tran, D., Rudin, C., Srikumar, V., Feris, R., Li, S.: Energy-based out-of-distribution detection. In: NeurIPS (2020), https://arxiv.org/abs/1910.10755
18. Mackay, D.J.C.: Bayesian methods for adaptive models. California Institute of Technology (1992)
19. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H., Gal, Y.: Deep deterministic uncertainty: A new simple baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24384–24394 (2023)
20. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM conference on health, inference, and learning. pp. 151–159 (2020)
21. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: Ai in health and medicine. Nature medicine **28**(1), 31–38 (2022)
22. Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A simple fix to mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022 (2021)
23. Schulam, P., Saria, S.: Distributional robustness and regularization for structured prediction. In: ICML (2019), https://arxiv.org/abs/1810.03720
24. Sun, Y., Tang, Y., Liu, Y., Huang, L., Zhang, X., Zhang, L., Lin, D.: React: Out-of-distribution detection with rectified activations. In: NeurIPS (2021), https://arxiv.org/abs/2106.08147
25. 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), 41 (2023)
26. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine **15**(11), e1002683 (2018)