# Evidence Reconciled Neural Network for Out-of-Distribution Detection in Medical Images

Wei Fu[1], Yufei Chen[1](✉), Wei Liu[1], Xiaodong Yue[2,3], and Chao Ma[1,4]

[1] College of Electronics and Information Engineering,
Tongji University, Shanghai, China
yufeichen@tongji.edu.cn

[2] Artificial Intelligence Institute of Shanghai University, Shanghai, China

[3] VLN Lab, NAVI MedTech Co., Ltd., Shanghai, China

[4] Department of Radiology, Changhai Hospital of Shanghai, Shanghai, China

**Abstract.** Near Out-of-Distribution (OOD) detection is a crucial issue in medical applications, as misdiagnosis caused by the presence of rare diseases inevitablely poses a significant risk. Recently, several deep learning-based methods for OOD detection with uncertainty estimation, such as the Evidential Deep Learning (EDL) and its variants, have shown remarkable performance in identifying outliers that significantly differ from training samples. Nevertheless, few studies focus on the great challenge of near OOD detection problem, which involves detecting outliers that are close to the training distribution, as commonly encountered in medical image application. To address this limitation and reduce the risk of misdiagnosis, we propose an Evidence Reconciled Neural Network (ERNN). Concretely, we reform the evidence representation obtained from the evidential head with the proposed Evidential Reconcile Block (ERB), which restricts the decision boundary of the model and further improves the performance in near OOD detection. Compared with the state-of-the-art uncertainty-based methods for OOD detection, our method reduces the evidential error and enhances the capability of near OOD detection in medical applications. The experiments on both the ISIC2019 dataset and an in-house pancreas tumor dataset validate the robustness and effectiveness of our approach. Code for ERNN has been released at https://github.com/KellaDoe/ERNN.

**Keywords:** Near out-of-distribution detection · Uncertainty estimation · Reconciled evidence representation · Evidential neural network

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_30.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 305–315, 2023.
https://doi.org/10.1007/978-3-031-43898-1_30

## 1 Introduction

Detecting out-of-distribution (OOD) samples is crucial in real-world applications of machine learning, especially in medical imaging analysis where misdiagnosis can pose significant risks [7]. Recently, deep neural networks, particularly ResNets [9] and U-Nets [15], have been widely used in various medical imaging applications such as classification and segmentation tasks, achieving state-of-the-art performance. However, due to the typical overconfidence seen in neural networks [8,18], deep learning with uncertainty estimation is becoming increasingly important in OOD detection.

Deep learning-based OOD detection methods with uncertainty estimation, such as Evidential Deep Learning (EDL) [10,17] and its variants [2,11–13,24], have shown their superiority in terms of computational performance, efficiency, and extensibility. However, most of these methods consider identifying outliers that significantly differ from training samples(e.g. natural images collected from ImageNet [5]) as OOD samples [1]. These approaches overlook the inherent near OOD problem in medical images, in which instances belong to categories or classes that are not present in the training set [21] due to the differences in morbidities. Failing to detect such near OOD samples poses a high risk in medical application, as it can lead to inaccurate diagnoses and treatments. Some recent works have been proposed for near OOD detection based on density models [20], preprocessing [14], and outlier exposure [16]. Nevertheless, all of these approaches are susceptible to the quality of the training set, which cannot always be guaranteed in clinical applications.

To address this limitation, we propose an Evidence Reconciled Neural Network (ERNN), which aims to reliably detect those samples that are similar to the training data but still with different distributions (near OOD), while maintain accuracy for In-Distribution (ID) classification. Concretely, we introduce a module named Evidence Reconcile Block (ERB) based on evidence offset. This module cancels out the conflict evidences obtained from the evidential head, maximizes the uncertainty of derived opinions, thus minimizes the error of uncertainty calibration in OOD detection. With the proposed method, the decision boundary of the model is restricted, the capability of medical outlier detection is improved and the risk of misdiagnosis in medical images is mitigated. Extensive experiments on both ISIC2019 dataset and in-house pancreas tumor dataset demonstrate that the proposed ERNN significantly improves the reliability and accuracy of OOD detection for clinical applications. Code for ERNN can be found at https://github.com/KellaDoe/ERNN.

## 2 Method

In this section, we introduce our proposed Evidence Reconciled Neural Network (ERNN) and analyze its theoretical effectiveness in near OOD detection. In our approach, the evidential head firstly generates the original evidence to support the classification of each sample into the corresponding class. And then, the proposed Evidence Reconcile Block (ERB) is introduced, which reforms the derived evidence representation to maximize the uncertainty in its relevant opinion and better restrict the model decision boundary. More details and theorical analysis of the model are described below.

### 2.1 Deep Evidence Generation

Traditional classifiers typically employ a softmax layer on top of feature extractor to calculate a point estimation of the classification result. However, the point estimates of softmax only ensure the accuracy of the prediction but ignore the confidence of results. To address this problem, EDL utilizes the Dirichlet distribution as the conjugate prior of the categorical distribution and replaces the softmax layer with an evidential head which produces a non-negative output as evidence and formalizes an opinion based on evidence theory to explicitly express the uncertainty of generated evidence.

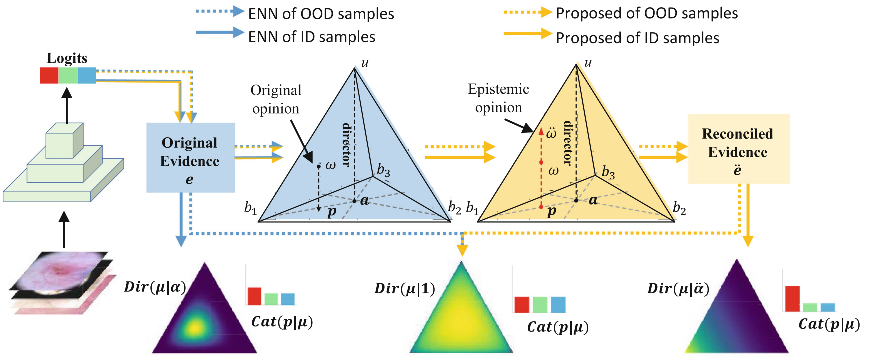


**Fig. 1.** Workflow of the proposed Evidence Reconciled Neural Network. In the Dirichlet distribution derived from evidence, the greater the predictive entropy (i.e., uncertainty), the closer the distribution expectation is to the center.

Formally, the evidence for $K$-classification task uniquely associates with a multinomial opinion $\boldsymbol{\omega} = (\boldsymbol{b}, u, \boldsymbol{a})$ which can be visualized shown in Fig 1 (in case of $K = 3$). In this opinion, $\boldsymbol{b} = (b_1, \ldots, b_K)^{\mathrm{T}}$ represents the belief degree, $\boldsymbol{a} = (a_1, \ldots, a_K)^{\mathrm{T}}$ indicates the prior preference of the model over classes, and $u$ denotes the overall uncertainty of generated evidence which is also called vacuity in EDL. The triplet should satisfy $\sum_{k=1}^{K} b_k + u = 1$, and the prediction probability can be defined as $p_k = b_k + a_k u$. Typically, in OOD detection tasks, all values of the prior $\boldsymbol{a}$ are set to $1/K$ with a non-informative uniform distribution.

Referring to subjective logic [10], the opinion $\boldsymbol{\omega}$ can be mapped into a Dirichlet distribution $Dir\,(\boldsymbol{p}\,|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^T$ represents the Dirichlet parameters and we have $\boldsymbol{\alpha} = \boldsymbol{e} + \boldsymbol{a}K$, $\boldsymbol{e} = (e_1, \ldots, e_K)^{\mathrm{T}}$ represents the evidence that is a measure of the amount of support collected from data in favor of a sample to be classified into known categories. When there is no preference over class, the Dirichlet parameters $\boldsymbol{\alpha} = \boldsymbol{e} + 1$, then the belief mass and uncertainty are calculated as:

$$\boldsymbol{b} = \frac{\boldsymbol{e}}{S} = \frac{\boldsymbol{\alpha} - 1}{S}, u = \frac{K}{S}, \tag{1}$$

where $S = \sum_{k=1}^{K}(e_k+1) = \sum_{k=1}^{K}\alpha_k$ is the Dirichlet strength. Based on the fact that the parameters of the categorical distribution should obey Dirichlet distribution, the model prediction $\hat{\boldsymbol{y}}$ and the expected cross entropy loss $L_{ece}$ on Dirichlet distributioncan be inferred as:

$$\hat{\boldsymbol{y}} = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\left[\mathbb{E}_{cat(\boldsymbol{y}|\boldsymbol{p})}\boldsymbol{p}\right] = \boldsymbol{\alpha}/S \tag{2}$$

$$L_{ece} = \mathbb{E}_{Dir(\boldsymbol{p}|\boldsymbol{\alpha})}\sum_{i=1}^{K} y_i \log p_i = \sum_{i=1}^{K} y_i(\psi(S) - \psi(\alpha_i)). \tag{3}$$

### 2.2 Evidence Reconcile Block

In case of OOD detection, since the outliers are absent in the training set, the detection is a non-frequentist situation. Referring to the subjective logic [10], when a variable is not governed by a frequentist process, the statical accumulation of supporting evidence would lead to a reduction in uncertainty mass. Therefore, traditional evidence generated on the basis accumulation is inapplicable and would lead to bad uncertainty calibration in OOD detection. Moreover, the higher the similarity between samples, the greater impact of evidence accumulation, which results in a dramatic performance degradation in medical near OOD detection.

To tackle the problem mentioned above, we propose an Evidence Reconcile Block (ERB) that reformulates the representation of original evidence and minimizes the deviation of uncertainty in evidence generation. In the proposed ERB, different pieces of evidence that support different classes are canceled out by transforming them from subjective opinion to epistemic opinion and the theoretical maximum uncertainty mass is obtained.

As shown in Fig. 1, the simplex corresponding to $K$-class opinions has $K$ dimensions corresponding to each category and an additional dimension representing the uncertainty in the evidence, i.e., vacuity in EDL. For a given opinion $\boldsymbol{\omega}$, its projected predictive probability is shown as $\boldsymbol{p}$ with the direction determined by prior $\boldsymbol{a}$. To ensure the consistency of projection probabilities, epistemic opinion $\ddot{\boldsymbol{\omega}}$ should also lie on the direction of projection and satisfy that at least one belief mass of $\ddot{\boldsymbol{\omega}}$ is zero, corresponding to a point on a side of the simplex. Let $\ddot{u}$ denotes the maximum uncertainty, it should satisfy:

$$\ddot{u} = \min_{i}\left[\frac{p_i}{a_i}\right], for\, i \in \{1, \ldots, K\}, \tag{4}$$

Since $\boldsymbol{a}$ is a uniform distribution defined earlier, the transformed belief mass can be calculated as: $\ddot{\boldsymbol{b}} = \boldsymbol{b} - b_{min}$, where $b_{min}$ is the minimum value in the original belief mass $\boldsymbol{b}$. Similarly, the evidence representation $\ddot{\boldsymbol{e}}$ in our ERB, based on epistemic opinion $\ddot{\boldsymbol{\omega}}$, can be formulated as:

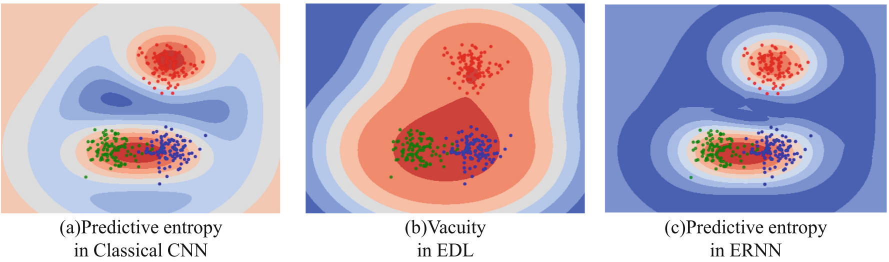
$$\ddot{\boldsymbol{e}} = \boldsymbol{e} - \min_i [\frac{e_i}{a_i}] \boldsymbol{a} = \boldsymbol{e} - \min_i e, for\, i \in \{1, \ldots, K\}. \tag{5}$$

After the transformation by ERB, the parameters $\ddot{\boldsymbol{\alpha}} = \ddot{\boldsymbol{e}} + 1$ of Dirichlet distribution associate with the reconciled evidence can be determined, and the reconciled evidential cross entropy loss $L_{rece}$ can be inferred as (6), in which $\ddot{S} = \sum_{i=1}^{K} \ddot{\alpha}_i$.

$$L_{rece} = \mathbb{E}_{Dir(\boldsymbol{p}|\ddot{\boldsymbol{\alpha}})} \sum_{k=1}^{K} y_i \log p_i = \sum_{i=1}^{K} y_i (\psi(\ddot{S}) - \psi(\ddot{\alpha}_i)). \tag{6}$$

By reconciling the evidence through the transformation of epistemic opinion in subjective logic, this model can effectively reduce errors in evidence generation caused by statistical accumulation. As a result, it can mitigate the poor uncertainty calibration in EDL, leading to better error correction and lower empirical loss in near OOD detection, as analyzed in Sect. 2.3.

## 2.3 Theorical Analysis of ERB


(a)Predictive entropy in Classical CNN

(b)Vacuity in EDL

(c)Predictive entropy in ERNN

**Fig. 2.** Uncertainty estimations for OOD detection in (a) Classical CNN, (b) EDL and (c) proposed ERNN on synthetic gaussian data. The red, green and blue points denote the samples of 3 different classes that follow Gaussian distributions. The blue area represents samples with higher uncertainty (i.e., OOD samples), while the red area represents samples with lower uncertainty (i.e., ID samples). (Color figure online)

As shown in Fig. 2, we utilize samples from three Gaussian distributions to simulate a 3-classification task and generate evidences based on the probability density of each class. When using the traditional CNN to measure the uncertainty of the output with predictive entropy, the model is unable to distinguish far OOD due to the normalization of softmax. While the introduction of evidence representation in the vacuity of EDL allows effective far OOD detections. However, due to the aforementioned impact of evidence accumulation, we observe that the EDL has a tendency to produce small uncertainties for outliers close to in-distribution (ID) samples, thus leading to failures in detecting near OOD samples. Our proposed method combines the benefits of both approaches, the evidence transformed by ERB can output appropriate uncertainty for both near and far OOD samples, leading to better identification performance for both types of outliers.

To further analyze the constraint of the proposed model in OOD detection, we theoretically analyze the difference between the loss functions before and after the evidence transformation, as well as why it can improve the ability of near OOD detection. Detailed provements of following propositons are provided in **Supplements**.

**Proposition 1.** *For a given sample in $K$-classification with the label $c$ and $\sum_{i=1}^{K} \alpha_i = S$, for any $\alpha_c \leq \frac{S}{K}$, $L_{rece} > L_{ece}$ is satisfied.*

The misclassified ID samples with $p_c = \alpha_c/S \leq 1/K$ are often located at the decision boundary of the corresponding categories. Based on Proposition 1, the reconciled evidence can generate a larger loss, which helps the model focus more on the difficult-to-distinguish samples. Therefore, the module can help optimize the decision boundary of ID samples, and promote the ability to detect near OOD.

**Proposition 2.** *For a given sample in $K$-classification with the label $c$ and $\sum_{i=1}^{K} \alpha_i = S$, for any $\alpha_c > \frac{S}{K}$, $L_{rece} < L_{ece}$ is satisfied.*

Due to the lower loss derived from the proposed method, we achieve better classification accuracy and reduce empirical loss, thus the decision boundary can be better represented for detecting outliers.

**Proposition 3.** *For a given sample in $K$-classification and Dirichlet distribution parameter $\boldsymbol{\alpha}$, when all values of $\boldsymbol{\alpha}$ equal to const $\tilde{\alpha}$, $L_{rece} \geq L_{ece}$ is satisfied.*

During the training process, if the prediction $\boldsymbol{p}$ of ID samples is identical to the ideal OOD outputs, the proposed method generates a greater loss to prevent such evidence from occurring. This increases the difference in predictions between ID and OOD samples, thereby enhancing the ability to detect OOD samples using prediction entropy.

In summary, the proposed Evidence Reconciled Neural Network (ERNN) optimizes the decision boundary and enhances the ability to detect near OOD samples. Specifically, our method improves the error-correcting ability when the probability output of the true label is no more than $1/K$, and reduces the empirical loss when the probability output of the true label is greater than $1/K$. Furthermore, the proposed method prevents model from generating same evidence for each classes thus amplifying the difference between ID and OOD samples, resulting in a more effective near OOD detection.

# 3 Experiments

## 3.1 Experimental Setup

**Datasets.** We conduct experiments on ISIC 2019 dataset [3,4,19] and an in-house dataset. ISIC 2019 consists of skin lesion images in JPEG format, which are categorized into NV (12875), MEL (4522), BCC (3323), BKL (2624), AK (867), SCC (628), DF (239) and VASC (253), with a long-tailed distribution of classes. In line with the settings presented in [14,16], we define DF and VASC, for which samples are relatively scarce as the near-OOD classes. The in-house pancreas tumor dataset collected from a cooperative hospital is composed of eight classes: PDAC (302), IPMN (71), NET (43), SCN (37), ASC (33), CP (6), MCN (3), and PanIN (1). For each sequence, CT slices with the largest tumor area are picked for experiment. Similarly, PDAC, IPMN and NET are chosen as ID classes, while the remaining classes are reserved as OOD categories.

**Implementations and Evaluation Metrics.** To ensure fairness, we used pre-trained Resnet34 [9] as backbone for all methods. During our training process, the images were first resized to 224 × 224 pixels and normalized, then horizontal and vertical flips were applied for augmentation. The training was performed using one GeForce RTX 3090 with a batch size of 256 for 100 epochs using the AdamW optimizer with an initial learning rate of 1e-4 along with exponential decay. Note that we employed five-fold cross-validation on all methods, without using any additional OOD samples during training. Furthermore, we selected the precision(pre), recall(rec), and f1-score(f1) as the evaluation metrics for ID samples, and used the Area Under Receiver Operator Characteristic (AUROC) as OOD evaluation metric, in line with the work of [6].

## 3.2 Comparison with the Methods

In the experiment, we compare the OOD detection performance of ERNN to several uncertainty-based approaches:

- *Prototype Network* described in [22], where the prototypes of classes are introduced and the distance is utilized for uncertainty estimation.
- *Prior Network* described in [12], in which the second order dirichlet distribution is utlized to estimate uncertainty.
- *Evidential Deep Learning* described in [17], introduces evidence representation and estimates uncertainty through subjective logic.
- *Posterior Network* described in [2], where density estimators are used for generating the parameters of dirichlet distributions.

Inspired by [14], we further compare the proposed method with Mixup-based methods:

- *Mixup*: As described in [23], mix up is applied to all samples.
- *MT-mixup*: Mix up is only applied to mid-class and tail-class samples.

- *MTMX-Prototype*: On the basis of MT mixup, prototype network is also applied to estimate uncertainty.

The results on two datasets are shown in Table 1. We can clearly observe that ERNN consistently achieves better OOD detection performance than other uncertainty-based methods without additional data augmentation. Even with using Mixup, ERNN exhibits near performance with the best method (MTMX-Prototype) on ISIC 2019 and outperforms the other methods on in-house datasets. All of the experimental results verify that our ERNN method improves OOD detection performance while maintaining the results of ID classification even without any changes to the existing architecture.

**Table 1.** Comparison with other methods. ID metrics -Precision (pre), Recall (rec), and f1-score(f1); OOD metric - AUROC(%).

| | ISIC2019 | | | | In-house Pancreas tumor | | | |
|---|---|---|---|---|---|---|---|---|
| | ID(pre) | ID(rec) | ID(f1) | OOD(AUROC) | ID(pre) | ID(rec) | ID(f1) | OOD(AUROC) |
| Baseline [9] | 0.86 | 0.86 | 0.86 | 68.15 | 0.76 | 0.78 | 0.76 | 54.39 |
| Prototype Networks [22] | 0.85 | 0.86 | 0.86 | 72.84 | 0.75 | 0.75 | 0.74 | 52.86 |
| Prior Networks [12] | 0.87 | 0.87 | 0.87 | 74.54 | 0.72 | 0.72 | 0.71 | 49.79 |
| EDL [17] | **0.88** | **0.88** | **0.88** | 72.51 | **0.78** | **0.79** | **0.78** | 55.39 |
| Posterior Networks [2] | 0.36 | 0.51 | 0.35 | 57.50 | 0.74 | 0.72 | 0.73 | 47.25 |
| **ERNN(ours)** | **0.88** | **0.88** | **0.88** | **75.11** | **0.78** | 0.77 | 0.77 | **57.63** |
| Mixup [23] | **0.87** | **0.88** | **0.88** | 71.72 | **0.78** | **0.79** | **0.76** | 56.32 |
| MT-mixup [14] | 0.85 | 0.85 | 0.86 | 73.86 | 0.76 | 0.76 | 0.74 | 58.50 |
| MTMX-Prototype [14] | 0.85 | 0.86 | 0.86 | **76.37** | 0.75 | 0.77 | 0.74 | 54.18 |
| **MX-ERNN(ours)** | 0.86 | 0.87 | 0.86 | **76.34** | **0.78** | 0.74 | 0.71 | **60.21** |

**Table 2.** Ablation study. We present f1 metric for ID validation and AUROC metric for OOD detection on both ISIC 2019 and in-house Pancreas tumor dataset. "✓" means ERNN with the corresponding component, "-" means "not applied".

| Method | | | ISIC2019 | | In-house | |
|---|---|---|---|---|---|---|
| Backbone | EH | ERB | ID(f1) | OOD(AUROC) | ID(f1) | OOD(AUROC) |
| ✓ | - | - | 0.86 | 68.15 | 0.76 | 54.39 |
| ✓ | ✓ | | **0.88** | 74.15 | **0.78** | 55.39 |
| ✓ | - | ✓ | - | - | - | - |
| ✓ | ✓ | ✓ | **0.88** | **75.11** | 0.77 | **57.63** |

### 3.3 Ablation Study

In this section, we conduct a detailed ablation study to clearly demonstrate the effectiveness of our major technical components, which consist of evaluation of evidential head, evaluation of the proposed Evidence Reconcile Block on both

ISIC 2019 dataset and our in-house pancreas tumor dataset. Since the Evidence Reconcile Block is based on the evidential head, thus there are four combinations, but only three experimental results were obtained. As shown in Table 2, It is clear that a network with an evidential head can improve the OOD detection capability by 6% and 1% on ISIC dataset and in-house pancreas tumor dataset respectively. Furthermore, introducing ERB further improves the OOD detection performance of ERNN by 1% on ISIC dataset. And on the more challenging in-house dataset, which has more similarities in samples, the proposed method improves the AUROC by 2.3%, demonstrating the effectiveness and robustness of our model on more challenging tasks.

## 4 Conclusion

In this work, we propose a simple and effective network named Evidence Reconciled Nueral Network for medical OOD detection with uncertainty estimation, which can measure the confidence in model prediction. Our method addresses the failure in uncertainty calibration of existing methods due to the similarity of near OOD with ID samples. With the evidence reformation in the proposed Evidence Reconcile Block, the error brought by accumulative evidence generation can be mitigated. Compared to existing state-of-the-art methods, our method can achieve competitive performance in near OOD detection with less loss of accuracy in ID classification. Furthermore, the proposed plug-and-play method can be easily applied without any changes of network, resulting in less computation cost in identifying outliers. The experimental results validate the effectiveness and robustness of our method in the medical near OOD detection problem.

**Acknowledgements.** This work was supported by the National Natural Science Foundation of China (No. 62173252, 61976134), and Natural Science Foundation of Shanghai (No. 21ZR1423900).

## References

1. Berger, C., Paschali, M., Glocker, B., Kamnitsas, K.: Confidence-Based Out-of-Distribution Detection: A Comparative Study and Analysis. In: Sudre, C.H., et al. (eds.) UNSURE/PIPPI -2021. LNCS, vol. 12959, pp. 122–132. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87735-4_12
2. Charpentier, B., Zügner, D., Günnemann, S.: Posterior network: uncertainty estimation without OOD samples via density-based pseudo-counts. Adv. Neural Inf. Process. Syst. **33**, 1356–1367 (2020)
3. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International skin Imaging Collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 168–172. IEEE (2018)
4. Combalia, M., et al.: Bcn20000: dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)

5. Deng, J., et al.: ImageNet: a large-scale hierarchical image database. In: CVPR09 (2009)
6. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3614–3631 (2020)
7. Ghafoorian, M., et al.: Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 516–524. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_59
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Jøsang, A.: Subjective logic, vol. 4. Springer (2016). https://doi.org/10.1007/978-3-319-42337-1
11. Liu, W., Yue, X., Chen, Y., Denoeux, T.: Trusted multi-view deep learning with opinion aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 7585–7593 (2022)
12. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. Adv. Neural Inf. Process. Syst. **31** (2018)
13. Malinin, A., Gales, M.: Reverse kl-divergence training of prior networks: improved uncertainty and adversarial robustness. Adv. Neural Information Process. Syst. **32** (2019)
14. Mehta, D., Gal, Y., Bowling, A., Bonnington, P., Ge, Z.: Out-of-distribution detection for long-tailed and fine-grained skin lesion images. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I. pp. 732–742. Springer (2022). https://doi.org/10.1007/978-3-031-16431-6_69
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Roy, A.G., et al.: Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. Med. Image Anal. **75**, 102274 (2022)
17. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Adv. Neural Information Process. Syst. **31** (2018)
18. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: improved calibration and predictive uncertainty for deep neural networks. Adv. Neural Information Process. Syst. **32** (2019)
19. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
20. Ulmer, D., Meijerink, L., Cinà, G.: Trust issues: uncertainty estimation does not enable reliable OOD detection on medical tabular data. In: Machine Learning for Health. pp. 341–354. PMLR (2020)
21. Winkens, J., et al.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020)
22. Yang, H.M., Zhang, X.Y., Yin, F., Liu, C.L.: Robust classification with convolutional prototype learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3474–3482 (2018)

23. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
24. Zhao, X., Ou, Y., Kaplan, L., Chen, F., Cho, J.H.: Quantifying classification uncertainty using regularized evidential neural networks. arXiv preprint arXiv:1910.06864 (2019)