Calibration of Medical Imaging Classification Systems with Weight Scaling *

Lior Frenkel and Jacob Goldberger

Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel
{lior.frenkel,jacob.goldberger}@biu.ac.il

Abstract. Calibrating neural networks is crucial in medical analysis applications where the decision making depends on the predicted probabilities. Modern neural networks are not well calibrated and they tend to overestimate probabilities when compared to the expected accuracy. This results in a misleading reliability that corrupts our decision policy. We define a weight scaling calibration method that computes a convex combination of the network output class distribution and the uniform distribution. The weights control the confidence of the calibrated prediction. The most suitable weight is found as a function of the given confidence. We derive an optimization method that is based on a closed form solution for the optimal weight scaling in each bin of a discretized value of the prediction confidence. We report experiments on a variety of medical image datasets and network architectures. This approach achieves state-of-the-art calibration with a guarantee that the classification accuracy is not altered.

Keywords: network calibration · medical decision calibration · network interpretability · temperature scaling · weight scaling

1 Introduction

A classifier is said to be calibrated if the probability values it associates with the class labels match the true probabilities of the correct class assignments. Modern neural networks have been shown to be more overconfident in their predictions than their predecessors even though their generalization accuracy is higher, partly due to the fact that they can overfit on the negative log-likelihood loss without overfitting on the classification error [6,16,9]. In a medical imaging application, we would like to defer images for which the model makes low-confidence predictions to a physician for review. Skipping human review due to confident, but incorrect, predictions, could have disastrous consequences [17]. The lack of connection between the model's predicted probabilities and the model's accuracy is a key obstacle to the application of neural network models to automatic medical diagnosis [2,12,24].

Various confidence calibration methods have recently been proposed in the field of deep learning to overcome the over-confidence issue. Post-hoc scaling approaches to calibration (e.g. Platt scaling [23], isotonic regression [29], and temperature scaling [6]) are widely used. They perform calibration as a post processing step by using hold-out validation data to learn a calibration map that transforms the model's predictions to

* This research was supported by the Ministry of Science & Technology, Israel.

be better calibrated. Temperature scaling is the simplest and most effective calibration method and is the current standard practical calibration method. Guo at el. [6] investigated several scaling models, ranging from single-parameter based temperature scaling to more complex vector/matrix scaling. To avoid overfitting, Kull et al. [14] suggested regularizing matrix scaling with an L_2 loss on the calibration model weights. Gupta et al. [7] built a calibration function by approximating the empirical cumulative distribution using a differentiable function via splines. Most of these calibration methods extend single parameter temperature scaling by making the selected temperature either a linear or a non-linear function of the logits that are computed for the class-set (see e.g. [4,5]). Although network calibration is crucial for producing reliable automatic medical reports, there are only few works that directly address the issue of calibrating medical imaging systems (see e.g. [25,30,3]).

In this study we focus on calibration of neural networks that are applied for medical imaging tasks and propose an alternative to temperature scaling which we dub weight scaling. Weight scaling calibrates the network by computing a suitable convex combination of the original class distribution and the uniform distribution. We show that unlike temperature, vector and matrix scaling [14] and other recently proposed methods (e.g. [7]), we can obtain a closed form solution for the optimal calibration parameters. The proposed calibration does not change the hard classification decision which allows it to be applied on any trained network and guarantees to retain the original classification accuracy in all the tested cases. Unlike previous methods, if a network is more confident on one patient than the other, it remains more confident after the calibration. We evaluated our method against leading calibration approaches on various medical imaging datasets and network architectures using the *expected calibration error* (ECE) [19] calibration measure.

2 Calibration Problem Formulation

Consider a network that classifies an input image x into k pre-defined categories. The last layer of the network architecture is comprised of a vector of k real values $z = (z_1, ..., z_k)$ known as *logits*. Each of these numbers is the score for one of the k possible classes. The logits' vector z is then converted into a soft decision distribution using a *softmax* layer: $p(y = i|x) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$ where x in the input image and y is the image class. The output of the softmax layer has the mathematical form of a distribution. However, the network is not explicitly trained to compute the actual posterior distribution of the classes.

The hard decision predicted class is calculated from the output distribution by $\hat{y} = \arg\max_i p(y = i|x) = \arg\max_i z_i$. The network *confidence* for this sample is defined by $\hat{p} = p(y = \hat{y}|x) = \max_i p(y = i|x)$. The network *accuracy* is defined by the probability that the most probable class $\hat{y}$ is indeed correct. The network is said to be *calibrated* if for each sample the confidence coincides with the accuracy. For example, assume there are hundred images and for each we have a clinical prediction with identical confidence score of 0.9. If the network is well calibrated we expect that the network decision would be correct in ninety cases.

Expected Calibration Error (ECE) [19] is the standard metric used to measure model calibration. It is defined as the expected absolute difference between the model's accuracy and its confidence, i.e., $\mathbb{E}_{x,y}\left|\mathbb{P}(y=\hat{y}|x)-\hat{p}\right|$, where $\mathbb{P}$ is the true probability that the network decision is correct. In practice, we only have a finite amount of validation set samples $(x_1,y_1),...,(x_n,y_n)$, with associated predictions and confidence values $(\hat{y}_1,\hat{p}_1),...,(\hat{y}_n,\hat{p}_n)$. Hence, we cannot directly compute the ECE using this definition. Instead, we divide the unit interval $[0,1]$ into m bins, where the i^{th} bin is the interval $b_i=\left(\frac{i-1}{m},\frac{i}{m}\right]$. Let $B_i=\{t|\hat{p}_t\in b_i\}$ be the set of samples whose confidence values belong to the bin b_i. The accuracy of this bin is computed as $A_i=\frac{1}{|B_i|}\sum_{t\in B_i}\mathbb{1}\left(\hat{y}_t=y_t\right)$, where $\mathbb{1}$ is the indicator function, and y_t and $\hat{y}_t$ are the ground-truth and predicted labels for x_t. A_i is the relative number of correct predictions of instances that were assigned to B_i based on their confidence value. Similarly, the confidence C_i is the average confidence values of all samples in the bin b_i, i.e., $C_i=\frac{1}{|B_i|}\sum_{t\in B_i}\hat{p}_t$. Note that if the network is under-confident at bin b_i then $A_i>C_i$ and vise versa. The ECE can be thus computed as follows:

$$\text{ECE}=\sum_{i=1}^{m}\frac{|B_i|}{n}\left|A_i-C_i\right|. \tag{1}$$

ECE is based on a uniform bin width. If the model is well trained then, hopefully, most of the samples lie within the highest confidence bins. Hence, low confidence bins are almost empty and therefore have no influence on the computed value of the ECE. For this reason, we can consider another metric, Adaptive ECE (adaECE) [20]:

$$\text{adaECE}=\frac{1}{m}\sum_{i=1}^{m}\left|A_i-C_i\right| \tag{2}$$

such that each bin contains $1/m$ of the data points with similar confidence values. Even though the drawbacks of ECE have been pointed out and some improvements have been proposed [15,21,7,31], the ECE is still used as the standard calibration evaluation measure.

3 Weight Scaling based on the Predicted Confidence

Temperature Scaling (TS), is a simple yet highly effective technique for calibrating prediction probabilities [6]. It uses a single scalar parameter $T>0$, where T is the temperature, to rescale logit scores before applying the softmax function to compute the class distribution. The optimal temperature T for a trained model, is found by minimizing the negative log likelihood for a held-out validation dataset. Alternatively, the adaECE measure can be used as the objective score when finding the optimal T. Let A_i and C_i be the accuracy and confidence of the validation-set points in the i-th set B_i. Denote the average confidence in bin i after temperature scaling of all the instances in B_i by a temperature T by $C_i(T)$:

$$C_i(T)=\frac{1}{|B_i|}\sum_{t\in B_i}\max_{j=1}^{k}\frac{\exp(z_{tj}/T)}{\sum_{l=1}^{k}\exp(z_{tl}/T)} \tag{3}$$

s.t. $z_{t1}, ..., z_{tk}$ are the logit values computed by the network that is fed by x_t. The optimal temperature T can be found by minimizing the following adaECE score:

$$L_{\text{TS}}(T) = \frac{1}{m} \sum_{i=1}^{m} |A_i - C_i(T)| \,. \tag{4}$$

The minimization is carried out by a grid search over the possible values of T. Direct minimization of the adaECE measure (2) on the validation set was shown to yield better calibration results than maximizing the likelihood on a validation set [18].

Ji et al. [11] extended TS to a bin-wise setting, denoted Bin-wise Temperature Scaling (BTS), by setting separate temperatures for each bin. BTS is trained by maximizing the log-likelihood function. We can also directly minimize the gap between the confidence and the accuracy in each bin by minimizing the following adaECE score:

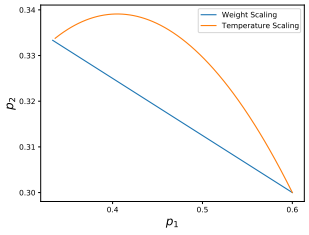
$$L_{\text{CTS}}(T_1, ..., T_m) = \frac{1}{m} \sum_{i=1}^{m} |A_i - C_i(T_i)| \,, \tag{5}$$

We need to apply a grid search to find T_i that satisfies $A_i = C_i(T_i)$. We denote this calibration method Confidence based Temperature Scaling (CTS). Similar to the case of a single temperature, it can be shown that CTS consistently yields better calibration results than BTS. We use CTS as one of the baseline methods that are compared with the calibration method we propose next.

Varying the distribution temperature T from 1 to ∞ induces a continuous path from the original class distribution $p = (p_1, ..., p_k)$ to the uniform distribution $u = (1/k, ..., 1/k)$. The notion of temperature scaling of a distribution originated in statistical physics. There is no intrinsic reason to specifically use a temperature to make the network output distribution smoother. The relevant features of temperature scaling as a smoothing procedure are that the entropy increases monotonically and the confidence decreases monotonically as a function of T, the order of probabilities from smallest to largest is preserved in the smoothing operation and it is a continuous function of T. In this study we put forward a different way to make a distribution smoother. For each weight $\alpha \in [0, 1]$ we define a smooth version of the original distribution p as follow:

$$p_\alpha = \alpha p + (1 - \alpha)u. \tag{6}$$

Varying the weight α from 1 to 0 induces a different path from the class distribution p to the uniform distribution u. We denote the calibration approach based on shifting from p to p_α (6) as Weight Scaling (WS). The figure at the right shows the trajectories of temperature scaling and weight scaling from $p = [0.6, 0.3, 0.1]$ to $u = [1/3, 1/3, 1/3]$.



It can be easily verified that the entropy $H(p_\alpha)$ is a concave function of α and obtains its global maximum at $\alpha = 0$. Hence, as p_α moves from p to u, the entropy of p_α monotonically increases. The confidence after weight scaling by α is simply $\hat{p}_\alpha = \alpha\hat{p} + (1 - \alpha)1/k$ where $\hat{p}$ is the confidence before calibration.

Both temperature scaling and weight scaling preserve the order of the predicted classes and therefore do not change the original hard classification decision. Another

Algorithm 1 Confidence based Weight Scaling (CWS)

training input: A validation dataset $x_1, ..., x_n$ where each x_t is fed into a k-class classifier network to produce class distribution $p_{t1}, ..., p_{tk}$.
Compute the confidence values: $\hat{p}_t = \max_j p_{tj}, \qquad t = 1, ..., n.$
Order the points based on their confidence and divide them into equal size sets $B_1, ..., B_m$.
Compute the average accuracy A_i and confidence C_i based on the points in B_i, and compute the calibration weight:

$$\alpha_i = \max(0, \min(1, \frac{A_i - \frac{1}{k}}{C_i - \frac{1}{k}})), \qquad i = 1, ..., m$$

training output: weights $\alpha_1, ..., \alpha_m$ and a division of the unit interval into m bins.

Calibration procedure:
- Given a point x with class distribution $p_1, ..., p_k$, compute the confidence: $\hat{p} = \max_j p_j$.
- Find the index $i \in \{1, ..., m\}$ s.t. $\hat{p}$ is within the borders of i-th bin.
- The calibrated prediction is: $p(y = j|x) = \alpha_i p_j + (1 - \alpha_i)\frac{1}{k}, \qquad j = 1, ..., k$

desired property of a calibration method is preserving the order of clinical decisions in different patients based on the decision confidence. However, as it can easily be verified, a network that is more confident at patient x than at patient y can become less confident at x than y after a temperature scaling calibration using the same temperature in both cases. In contrast, since weight scaling is a monotone function of the confidence, it preserves the ranking of patients based on the clinical decision confidence after weight scaling by the same α.

We next use the adaECE score to learn a calibration procedure based on weight scaling instead of temperature scaling. In the case of weight scaling let

$$C_i(\alpha) = \frac{1}{|B_i|} \sum_{t \in B_i} \max_{j=1}^{k} (\alpha p_{tj} + (1 - \alpha)\frac{1}{k}) = \alpha C_i + (1 - \alpha)\frac{1}{k} \tag{7}$$

be the confidence in bin i after scaling by a weight α where $p_{t1}, ..., p_{tk}$ are the softmax probability values computed by the network that is fed by x_t. In the case of single parameter weight scaling, we look for a weight α that minimizes the following adaECE score:

$$L_{\text{WS}}(\alpha) = \frac{1}{m} \sum_{i=1}^{m} |A_i - C_i(\alpha)| = \frac{1}{m} \sum_{i=1}^{m} |A_i - \alpha C_i - (1 - \alpha)\frac{1}{k}|. \tag{8}$$

Here there is no closed form solution for the optimal α. In a way similar to CTS, we can allow assigning a different weight in each confidence bin. To find the weight set that minimizes the following adaECE score:

$$L_{\text{CWS}}(\alpha_1, ..., \alpha_m) = \frac{1}{m} \sum_{i=1}^{m} |A_i - C_i(\alpha_i)|, \tag{9}$$

we can perform the minimization in each bin separately. To determine the number of bins we compute the adaECE score on the validation set and choose the number of bins

that yields the minimal adaECE score. In the case of weight scaling (unlike temperature scaling) there is a closed form solution to the equation:

$$A_i = C_i(\alpha_i) = \alpha_i C_i - (1 - \alpha_i)\frac{1}{k} \tag{10}$$

which is

$$\alpha_i = \frac{A_i - \frac{1}{k}}{C_i - \frac{1}{k}}. \tag{11}$$

The definition of confidence as the probability of the most likely class implies that always $1/k \leq C_i$. If $1/k \leq A \leq C_i$ then $\alpha_i \in [0, 1]$. In the (rare) case of accuracy less than random, i.e. $A_i < 1/k$, we set $\alpha_i = 0$ and in the (rare) case of under-confidence, i.e. $C_i < A_i$, we set $\alpha_i = 1$. The obtained calibration method is denoted Confidence based Weight Scaling (CWS) and is summarized in Algorithm box 1.

4 Experimental Results

We implemented the proposed calibration methods on various medical imaging classification tasks to evalute their performance. The experimental setup included the following medical datasets:

- **ChestX-ray14** [28]: A huge dataset that contains 112,120 frontal-view X-ray images of 30,805 unique patient of size 1024×1024, individually labeled with up to 14 different thoracic diseases. The original dataset is multi-label. We treated the problem as a multi-class task by choosing the samples contain only one annotated positive label. We used a train/validation/test split of 89,696/11,212/11,212 images.
- **HAM10000** [27]: This dataset contains 10,015 dermatoscopic images of size 800×600. Cases include a representative collection of 7 diagnostic categories in the realm of pigmented lesions. We used a train/validation/test split of 8,013/1,001/1,001 images.
- **COVID-19** [22]: A small dataset of X-ray images obtained from from two different sources. 127 COVID-19 X-ray images was taken from [1] and 500 no-findings and 500 pneumonia frontal chest X-ray images was randomly taken from the ChestX-ray8 database [28]. Here, we used a train/validation/test split of 901/112/112 images.

Each dataset was fine-tuned on pre-trained ResNet-50 [8], DenseNet-121 [10] and VGG-16-BN (with batch normalization) [26] networks. The models were taken from PyTorch site [1]. These network architectures were selected because of their widespread use in classification problems. The last FC layer output size of each of them was adjusted to fit the corresponding number of classes for each dataset. All models were fine-tuned using cross-entropy loss and Adam optimizer [13] with learning rate of 0.0001.

Compared methods. WS and CWS were compared to TS, vector scaling (VS) and matrix scaling (MS) [14]. The optimal TS was found by minimizing the ECE score over a validation set [18]. We also implemented our CTS algorithm, which calculates

[1] https://pytorch.org/vision/stable/models.html

Table 1: ECE for top-1 predictions (in %) using 10 bins (with the lowest in bold and the second lowest underlined) on various medical imaging classification datasets and models with different calibration methods.

Dataset	Architecture	Acc (%)	Uncalibrated	TS	VS	MS	WS	CTS	CWS
ChestX-ray14	ResNet-50	52.7	2.14	2.14	2.00	3.67	2.13	<u>1.81</u>	**1.67**
	DenseNet-121	52.3	4.18	4.18	<u>2.99</u>	3.49	3.65	3.20	**2.75**
	VGG-16-BN	52.3	2.46	2.46	<u>1.81</u>	4.73	1.94	2.49	**1.79**
HAM10000	ResNet-50	88.8	5.17	<u>1.76</u>	5.28	3.08	3.53	1.93	**1.58**
	DenseNet-121	87.4	4.91	2.36	7.59	3.30	3.71	<u>2.00</u>	**1.71**
	VGG-16-BN	89.0	7.73	2.94	3.57	<u>1.78</u>	6.09	1.92	**1.59**
COVID-19	ResNet-50	91.1	6.08	**2.76**	21.39	15.35	4.65	3.99	<u>3.78</u>
	DenseNet-121	90.2	6.25	6.55	29.70	20.78	6.15	<u>5.69</u>	**4.65**
	VGG-16-BN	88.4	8.92	6.82	9.30	8.31	<u>4.27</u>	5.58	**4.09**

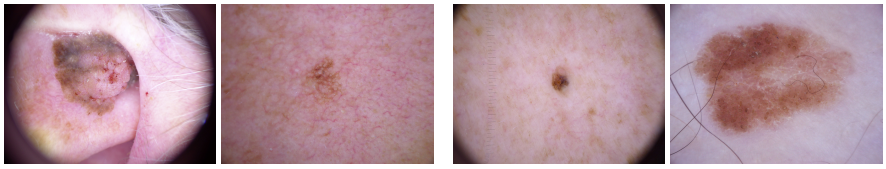
Table 2: adaECE for top-1 predictions (in %) using 10 bins (with the lowest in bold and the second lowest underlined).

Dataset	Architecture	Acc (%)	Uncalibrated	TS	VS	MS	WS	CTS	CWS
ChestX-ray14	ResNet-50	52.7	2.15	2.14	<u>1.95</u>	3.81	2.12	1.99	**1.69**

the optimal temperature in each bin (5). Note that VS and MS may change the model's accuracy and reduce the initial performance, while the other methods preserve it. For each method we evaluated the ECE score (computed using 10 bins) after calibration of the test set. Although adaECE was used as the objective function in our algorithm, ECE is still the standard way to report calibration results, so we used it to compare our calibration results with previous studies.

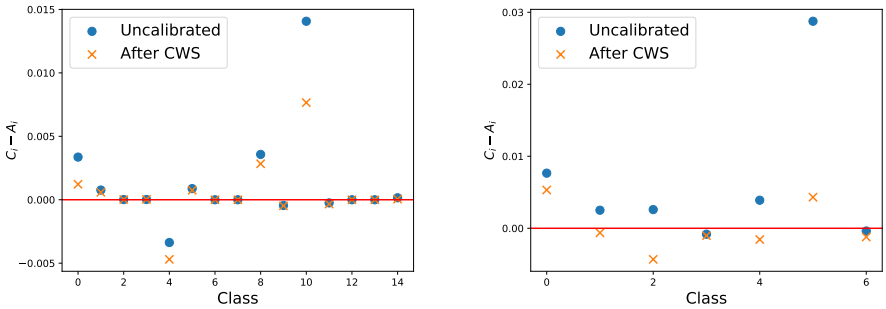
Results. Table 1 shows the calibration results. CWS achieved the best results in almost all cases, except one where it reached the second best result. Moreover, the ECE score after WS calibration was lower than the ECE after TS in more than a half of the cases. ChestX-ray14 is a large dataset with many classes. We can see in this case the advantage of WS over TS that is not calibrating at all (the optimal temperature was $T = 1$). The results also show that vector and matrix scaling collapse when using a small amount of classes, such as the COVID-19 dataset. We next verified that the calibration performance of CWS is still better than the other compared methods when adaECE is used for evaluation. Table 2 shows calibration results in one case evaluated by the adaECE score.

An advantage of WS is that it preserves the order of confidence of two samples, unlike TS that may violate this order. Fig. 1 presents two pairs of samples from HAM10000 with the same label (Benign keratosis). In each pair, the confidence of the first image before calibration was higher and after TS calibration the confidence became lower. Overall, TS changed the confidence order of 3% of the image pairs in the HAM10000 test dataset.



	0.55	$>$	0.53	0.59	$>$	0.53
WS	0.54	$>$	0.51	0.59	$>$	0.51
TS	0.46	$<$	0.49	0.45	$<$	0.51

Fig. 1: Two pairs of samples of benign keratosis taken from the HAM10000 dataset. For each image we show confidence before calibration (top row), after WS calibration (middle row) and after TS calibration (bottom row).



(a) (b)

Fig. 2: Difference between average confidence C_i and average accuracy A_i for each class i of the (a) ChestX-ray14 and (b) HAM10000 datasets trained on ResNet-50.

We investigated the level of confidence of each class in the ChestX-ray14 and HAM10000 datasets. Fig. 2 presents the difference between average confidence and average accuracy of each class before calibration and after applying the CWS algorithm for (a) ChestX-ray14 and (b) HAM10000 trained on ResNet-50. Positive difference symbolizes an over-confident class and negative difference represents an under-confident class. The labels of ChestX-ray14 are *Atelectasis*, *Cardiomegaly*, *Effusion*, *Infiltration*, *Mass*, *Nodule*, *Pneumonia*, *Pneumothorax*, *Consolidation*, *Edema*, *Emphysema*, *Fibrosis*, *PT*, *Hernia* and *No findings*. The labels of HAM10000 are *Actinic Keratoses*, *Basal cell carcinoma*, *Benign keratosis*, *Dermatofibroma*, *Melanoma*, *Melanocytic nevi* and *Vascular skin lesions*. The classes indexes of the model output displayed in Fig. 2 match this order. The results show that there are some classes in each dataset that are less calibrated than the others (like Emphysema in ChestX-ray14 and Melanocytic nevi in HAM10000) and some classes that are already relatively calibrated. They also show that CWS improves calibration for most of the classes and reduce the model's confidence in the few cases where the ECE gets higher.

To conclude, calibrated confidence estimates of predictions are critical to increase our trust in the using of neural networks for clinical decisions. As interest grows in deploying neural networks in medical decision making systems, the predictable behavior of the model will be a necessity. In this work, we introduced a simple and effective calibration method based on weight scaling of the prediction confidence. Most calibration methods are trained by optimizing the cross entropy score. CWS function learning can be done by explicitly optimizing the ECE measure. We compared our CWS method to various state-of-the-art methods and showed that it was on par in term of the ECE measure. We believe that it can be used in place of the standard temperature scaling method. In general, a calibrated prediction has a concrete probabilistic interpretation that hopefully enables practitioners build better trust on AI systems.

References

1. Cohen, J.P., Morrison, P., Dao, L.: Covid-19 image data collection. arXiv 2003.11597 (2020), https://github.com/ieee8023/covid-chestxray-dataset
2. Crowson, C.S., Atkinson, E.J., Therneau, T.M.: Assessing calibration of prognostic risk scores. Statistical Methods in Medical research **25**(4), 1692–1706 (2016)
3. Fernando, K.R.M., Tsokos, C.P.: Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems (2021)
4. Frenkel, L., Goldberger, J.: Network calibration by class-based temperature scaling. In: The European Signal Processing Conference (EUSIPCO) (2021)
5. Frenkel, L., Goldberger, J.: Network calibration by temperature scaling based on the predicted confidence. In: The European Signal Processing Conference (EUSIPCO) (2022)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML) (2017)
7. Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., Hartley, R.: Calibration of neural networks using splines. In: International Conference on Learning Representations (ICLR) (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Ji, B., Jung, H., Yoon, J., Kim, K., Shin, Y.: Bin-wise temperature scaling (BTS): Improvement in confidence calibration performance through simple scaling techniques. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) (2019)
12. Jiang, X., Osl, M., Kim, J., Ohno-Machado, L.: Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association **192**(2), 263–274 (2011)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kull, M., Nieto, M.P., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In: Advances in Neural Information Processing Systems (NeurIPs) (2019)

15. Kumar, A., Liang, P.S., Ma, T.: Verified uncertainty calibration. In: Advances in Neural Information Processing Systems (NeurIPs) (2019)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems (NeurIPs) (2017)
17. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems (NeurIPs) (2021)
18. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H., Dokania, P.K.: Calibrating deep neural networks using focal loss. In: Advances in Neural Information Processing Systems (NeurIPs) (2020)
19. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: AAAI Conference on Artificial Intelligence (2015)
20. Nguyen, K., O'Connor, B.: Posterior calibration and exploratory analysis for natural language processing models. arXiv preprint arXiv:1508.05154 (2015)
21. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops (2019)
22. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in Biology and Medicine **121**, 103792 (2020)
23. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers **10**(3), 61–74 (1999)
24. Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., Kleinberg, J.: Direct uncertainty prediction for medical second opinions. In: International Conference on Machine Learning (ICML) (2019)
25. Rousseau, A.J., Becker, T., Bertels, J., Blaschko, M.B., Valkenborg, D.: Post training uncertainty calibration of deep networks for medical image segmentation. In: International Symposium on Biomedical Imaging (ISBI) (2021)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
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), 1–9 (2018)
28. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
29. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: International Conference on Knowledge Discovery and Data Mining (KDD) (2002)
30. Zhang, F., Dvornek, N., Yang, J., Chapiro, J., Duncan, J.: Layer embedding analysis in convolutional neural networks for improved probability calibration and classification. IEEE Transactions on Medical Imaging **39**(11), 3331–3342 (2020)
31. Zhang, J., Kailkhura, B., Han, T.Y.J.: Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In: International Conference on Machine Learning (ICML) (2020)