How Reliable are the Metrics Used for Assessing Reliability in Medical Imaging?

Mayank Gupta(✉), Soumen Basu, and Chetan Arora

Indian Institute of Technology, Delhi, India
mayank.gupta@cse.iitd.ac.in

Abstract. Deep Neural Networks (DNNs) have been successful in various computer vision tasks, but are known to be uncalibrated, and make overconfident mistakes. This erodes a user's confidence in the model and is a major concern in their applicability for critical tasks like medical imaging. In the last few years, researchers have proposed various metrics to measure miscalibration, and techniques to calibrate DNNs. However, our investigation shows that for small datasets, typical for medical imaging tasks, the common metrics for calibration, have a large bias as well as variance. It makes these metrics highly unreliable, and unusable for medical imaging. Similarly, we show that state-of-the-art (SOTA) calibration techniques while effective on large natural image datasets, are ineffective on small medical imaging datasets. We discover that the reason for failure is large variance in the density estimation using a small sample set. We propose a novel evaluation metric that incorporates the inherent uncertainty in the predicted confidence, and regularizes the density estimation using a parametric prior model. We call our metric, Robust Expected Calibration Error (RECE), which gives a low bias, and low variance estimate of the expected calibration error, even on the small datasets. In addition, we propose a novel auxiliary loss - Robust Calibration Regularization (RCR) which rectifies the above issues to calibrate the model at train time. We demonstrate the effectiveness of our RECE metric as well as the RCR loss on several medical imaging datasets and achieve SOTA calibration results on both standard calibration metrics as well as RECE. We also show the benefits of using our loss on general classification datasets. The source code and all trained models have been released (https://github.com/MayankGupta73/Robust-Calibration).

Keywords: Confidence Calibration · Uncertainty Estimation · Image Classification

1 Introduction

Application of DNNs to critical applications like medical imaging requires that a model is not only accurate, but also well calibrated. Practitioners can trust

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_15.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 149–158, 2023.
https://doi.org/10.1007/978-3-031-43898-1_15

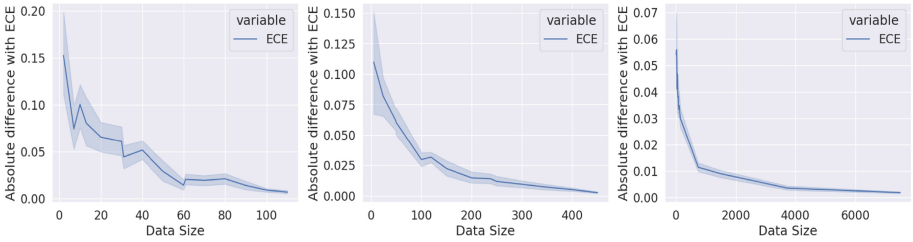

Fig. 1. The graphs show the `ECE` computed on the sample sets of various sizes (x axis) drawn from a given test distribution: (Left) GBCU [2] (Middle) POCUS [3], and (Right) Diabetic Retinopathy [12]. Notice the large bias and the variance, especially for the small sample sets. In this paper we propose a novel calibration metric, Robust `ECE`, and a novel train time calibration loss, `RCR`, especially suited for small datasets, which is a typical scenario for medical image analysis tasks.

deployed models if they are certain that the model will give a highly confident answer when it is correct, and uncertain samples will be labeled as such. In case of uncertainty, the doctors can be asked for a second opinion instead of an automated system giving highly confident but incorrect and potentially disastrous predictions [11]. Researchers have shown that modern `DNNs` are poorly calibrated and overconfident [6,25]. To rectify the problem, various calibration methods have been proposed such as Platt-scaling [26] based post-hoc calibration, or the train-time calibration methods such as `MDCA` [8]. We note that these methods have been mostly tested on large natural image datasets.

Our key observation is that current metrics for calibration are highly unreliable for small datasets. For example, given a particular data distribution, if one measures calibration on various sample sets drawn from the distribution, an ideal metric should give an estimate with low bias and variance. We show that this does not hold true for popular metrics like `ECE` (c.f. Fig. 1). We investigate the reason for such discrepancy. We observe that all the techniques first divide the confidence range into bins, and then estimate the probability of a model predicting confidence in each bin, along with the accuracy of the model in that bin. Such probability estimates become highly unreliable when the sample set is small. Imagine, seeing one correct sample with confidence 0.7, and then declaring the model under-confident in that bin. Armed with the insight, we go on proposing a new metric called Robust `ECE` especially suited for small datasets, and an auxiliary `RCR` loss to calibrate a model at the train time. The proposed loss can be used in addition to an application specific loss to calibrate the model.

Contributions: (1) We demonstrate the ineffectiveness of common calibration metrics on medical image datasets with limited number of samples. **(2)** We propose a novel and robust metric to estimate calibration accurately on small datasets. The metric regularizes the probability of predicting a particular confidence value by estimating a parametric density model for each sample. The calibration estimates using the regularized probability estimates have significantly lower bias, and variance. **(3)** Finally, we also propose a train-time auxiliary loss for calibrating models trained on small datasets. We validate the proposed loss on several public medical datasets and achieve `SOTA` calibration results.

2 Related Work

Metrics for Estimating Calibration: Expected Calibration Error (ECE) [21], first proposed in the context of DNNs by [6], divides the predicted confidence values in various bins and then calculates the absolute difference of average confidence and accuracy in that bin. The aggregate over all the bins is outputted as the calibration error. The motivation is to compute the probability of the model outputting a particular confidence, and the accuracy for the samples getting the particular confidence. The expectation of the difference is the calibration error. Since the error relies on accurate computation of probability, the same has large bias and variance when the dataset is small. Static Calibration Error (SCE) [23] extends ECE to multi-class settings by computing ECE for each class and then taking the average. Both ECE and SCE suffer from non-uniform distribution of samples into various bins, resulting in some bins getting small or no allocations. Adaptive binning (AECE) [22] attempts to mitigate the same by adaptively changing the bin sizes according to the given sample set. ECE-KDE [27] uses Dirichlet kernel density estimates for estimating calibration error.

Calibrating DNNs: Calibration techniques typically reshape the output confidence vector so as to minimize a calibration loss. The techniques can be broadly categorized into post-hoc and train-time techniques. Whereas post-hoc techniques [6,10,14] use a validation set to learn parameters to reshape the output probability vector, the train time techniques typically introduce an additional auxiliary loss to aid in calibration [8,15,19,20,24,25,30]. While being more intrusive, such techniques are more popular due to their effectiveness. We also follow a similar approach in this paper. Other strategies for calibration includes learning robust representations leading to calibrated confidences [5,9,16,32].

Calibration in Medical Imaging: Liang *et al.* [17] propose DCA loss which has been used for calibration of medical classification models. The loss aims to minimize the difference between predicted confidence and accuracy. However, the datasets demonstrated are quite large and the technique does not indicate any benefits for smaller datasets. Carneiro *et al.* [4] use MC-Dropout [5] entropy estimation and temperature scaling [6] for calibrating a model trained on colonoscopy polyp classification. Rajaraman *et al.* [28] demonstrate a few calibration methods for classification on class-imbalanced medical datasets.

3 Proposed Methodology

Model Calibration: Let $\mathcal{D}$ be a dataset with N samples: $\mathcal{D} = \{(x_i, y_i^*) \mid i \in 1, \ldots, N\}$. Given an input $x_i \in \mathcal{X}$, a classification model f must predicts its categorical label $y \in \mathcal{Y} = \{1, \ldots, K\}$. Hence, for a sample i, the model outputs a confidence vector, $C_i \in [0,1]^K$, a probability vector denoting the confidence for each class. A prediction $\widehat{y}_i$ is made by selecting the class with maximum confidence in C. A model is said to be calibrated if:

$$\mathbb{P}\left((y^* = \widehat{y}) \;\middle|\; (\widehat{y} = \arg\max_{y_i} C[y_i])\right) = C[\widehat{y}]. \tag{1}$$

Expected Calibration Error (ECE): is computed by bin-wise addition of difference between the average accuracy A_i and average confidence C_i:

$$\texttt{ECE} = \sum_{i=1}^{M} \frac{B_i}{N} |A_i - C_i|, \tag{2}$$

$$A_i = \frac{1}{B_i} \sum_{j \in B_i} \mathbb{I}(y_j^* = \widehat{y}_j), \qquad C_i = \frac{1}{B_i} \sum_{j \in B_j} C_j[\widehat{y}_j] \tag{3}$$

Here, C_j denotes the confidence vector, and $\widehat{y}_j$ predicted label of a sample j. The confidences, $C[\widehat{y}]$, of all the samples being evaluated are split into M equal sized bins with the i^{th} bin having B_i number of samples. C_i represents the average confidence of samples in the i^{th} bin. The basic idea is to compute the probability of outputting a particular confidence and the associated accuracy, and the expression merely substitutes sample mean in place of true probabilities.

Static Calibration Error (SCE): extends ECE to a multi-class setting as follows:

$$\texttt{SCE} = \frac{1}{K} \sum_{i=1}^{M} \sum_{k=1}^{K} \frac{B_{i,k}}{N} |A_{i,k} - C_{i,k}|, \tag{4}$$

$$A_{i,k} = \frac{1}{B_{i,k}} \sum_{j \in B_{i,k}, \widehat{y}_j = k} \mathbb{I}(y_j^* = \widehat{y}_j), \qquad C_{i,k} = \frac{1}{B_{i,k}} \sum_{j \in B_{i,k}, \widehat{y}_j = k} C_j[\widehat{y}_j] \tag{5}$$

Here $B_{i,k}$ denotes the number of samples of class k in the i^{th} bin.

3.1 Proposed Metric: Robust Expected Calibration Error (RECE)

We propose a novel metric which gives an estimate of true ECE with low bias, and variance, even when the sample set is small. RECE incorporates the inherent uncertainty in the prediction of a confidence value, by considering the observed value as a sample from a latent distribution. This not only helps avoid overfitting on outliers, but also regularizes the confidence probability estimate corresponding to each confidence bin. We consider two versions of RECE based on the parameterization of the latent distribution.

RECE-G: Here, we assume a Gaussian distribution of fixed variance (σ) as the latent distribution for each confidence sample. We estimate the mean of the latent distribution as the observed sample itself. Formally:

$$\texttt{RECE-G} = \sum_{i=1}^{M} \frac{1}{N} |\widetilde{A}_i - \widetilde{C}_i|, \quad \text{where} \quad \widetilde{C}_i = \sum_{j=1}^{N} c_j * \frac{\mathcal{N}\left(\left[\frac{i-1}{M}, \frac{i}{M}\right]; c_j, \sigma\right)}{\sum_{k=1}^{M} \mathcal{N}\left(\left[\frac{k-1}{M}, \frac{k}{M}\right]; c_j, \sigma\right)},$$

$$\text{and} \quad \widetilde{A}_i = \sum_{j=1}^{N} \mathbb{I}(\widehat{y}_j = y_j^*) * \frac{\mathcal{N}\left(\left[\frac{i-1}{M}, \frac{i}{M}\right]; c_j, \sigma\right)}{\sum_{k=1}^{M} \mathcal{N}\left(\left[\frac{k-1}{M}, \frac{k}{M}\right]; c_j, \sigma\right)}. \tag{6}$$

To prevent notation clutter, we use c_j to denote $C_j[\widehat{y}_j]$. Further, $\mathcal{N}([a, b]; \mu, \sigma)$ denotes the probability of the interval $[a, b]$ for a Gaussian distribution with mean

μ, and variance σ. In the above expression, the range $[\frac{i-1}{M}, \frac{i}{M}]$ corresponds to the range of confidence values corresponding to i^{th} bin. We also normalize the weight values over the set of bins. The value of standard deviation σ is taken as a fixed hyper-parameter. Note that the expression is equivalent to sampling infinitely many confidence values from the distribution $\mathcal{N}(\cdot; c_j, \sigma)$ for each sample j, and then computing the ECE value from thus computed large sampled dataset.

RECE-M: Note that RECE-G assumes fixed uncertainty in confidence prediction for all the samples as indicated by the choice of single σ for all the samples. To incorporate sample specific confidence uncertainty, we propose RECE-M in which we generate multiple confidence observations for a sample using test time augmentation. In our implementation, we generate 10 observations using random horizontal flip and rotation. We use the 10 observed values to estimate a Gaussian Mixture Model (denoted as $\mathcal{G}$) with 3 components. We use θ_j to denote the estimated parameters of mixture model for sample j. Note that, unlike RECE-G, computation of this metric requires additional inference passes through the model. Hence, the computation is more costly, but may lead to more reliable calibration estimates. Formally, RECE-M is computed as:

$$\texttt{RECE-M} = \sum_{i=1}^{M} \frac{1}{N} |\widetilde{A}_i - \widetilde{C}_i|, \quad \text{where} \quad \widetilde{C}_i = \sum_{j=1}^{N} c_j * \frac{\mathcal{G}\left(\left[\frac{i-1}{M}, \frac{i}{M}\right]; \theta_j\right)}{\sum_{k=1}^{M} \mathcal{G}\left(\left[\frac{k-1}{M}, \frac{k}{M}\right]; \theta_j\right)},$$

$$\text{and} \quad \widetilde{A}_i = \sum_{j=1}^{N} \mathbb{I}(\widehat{y}_j = y_j^*) * \frac{\mathcal{G}\left(\left[\frac{i-1}{M}, \frac{i}{M}\right]; \theta_j\right)}{\sum_{k=1}^{M} \mathcal{G}\left(\left[\frac{k-1}{M}, \frac{k}{M}\right]; \theta_j\right)}. \tag{7}$$

3.2 Proposed Robust Calibration Regularization (RCR) Loss

Most train time auxiliary loss functions minimize ECE over a mini-batch. When the mini-batches are smaller, the problem of unreliable probability estimation affects those losses as well. Armed with insights from the proposed RECE metric, we apply similar improvements in state of the art MDCA loss [8]. We call the modified loss function as the Robust Calibration Regularization (RCR) loss:

$$\mathcal{L}_{\texttt{RCR}} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{1}{N_b} \sum_{j=1}^{N_b} z_j^k - \frac{1}{N_b} \sum_{j=1}^{N_b} \mathbb{I}(y_j^* = k) \right| \tag{8}$$

$$z_j^k = \frac{1}{N_b} \sum_{i=1}^{N_b} C_i[k] * \left(\frac{\mathcal{N}(C_j[k]; C_i[k], \sigma)}{\sum_{l=1}^{N_b} \mathcal{N}(C_l[k]; C_i[k], \sigma)} \right). \tag{9}$$

The RCR loss can be used as a regularization term along side any application specific loss function as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{application}} + \beta * \mathcal{L}_{\texttt{RCR}} \tag{10}$$

Here, β is a hyper-parameter for the relative weightage of the calibration. As suggested in the MDCA, we also use focal loss [19] for the $\mathcal{L}_{\text{application}}$. Our RCR

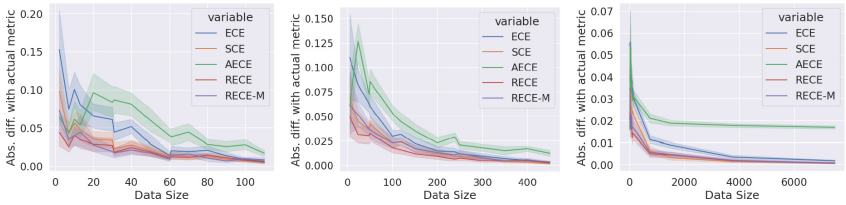

Fig. 2. Comparison on (Left) GBCU, (Middle) POCUS, and (Right) Diabetic Retinopathy datasets. The x-axis denotes the dataset size sampled and the y-axis denotes the absolute difference with the value obtained on the entire dataset. We repeat the process 20 times and plot the 95% confidence interval.

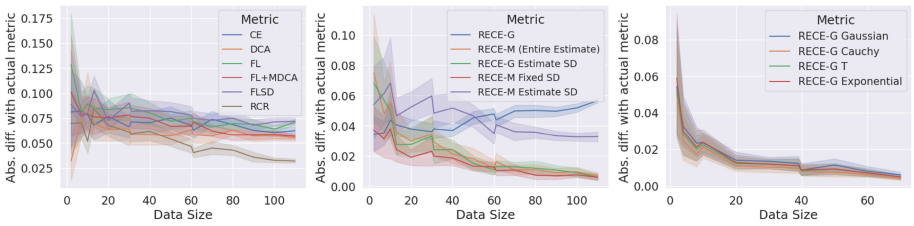

Fig. 3. Ablation experiments for RECE Metric. (Left) Effect of calibration on RECE-M for the GBC-USG Dataset. (Middle) Effect of different standard deviation strategies for the GBC-USG dataset. (Right) Effect of different distributions for the BUSI dataset. We give details in Supplementary A

loss is also independent of binning scheme and differentiable which allows for its application in multiple problem formulations outside of classification (though not the focus of this paper, and hence, not validated through experiments).

4 Experiments and Results

Datasets: We use following publicly available datasets for our experiments to demonstrate variety of input modalities, and disease focus. **GBCU** dataset [2] consists of 1255 ultrasound (US) images used for the classification of gallbladders into normal, benign and malignant. **BUSI** [1] consists of 830 breast US images divided into normal, benign and malignant. **POCUS** [3] is a lung US dataset consisting of 2116 images among healthy, pneumonia and covid-19. **Covid-CT** [33] consists of 746 CT images classified as covid and non-covid. The **Kaggle Diabetic Retinopathy (DR)** dataset [12] consists of 50089 retina images classified into 5 stages of DR severity. The **SIIM-ISIC Melanoma** dataset [29] has 33132 dermoscopic images of skin lesions classified into benign and malignant classes. Wherever the train-test splits have been specified (GBCU, POCUS and Covid-CT), we have used the same. For BUSI and Melanoma we create random stratified splits as none are available. For DR we follow the method of [31] and

Table 1. Comparison of metrics evaluated on different sample set sizes. Similar to Fig 2 we calculate the mean and std dev on different data sizes. The highlighted metric is closest to actual value (on 100% data)

Dataset & Model	Eval Metric	100% Data	1% Data	5% Data	10% Data	25% Data	50% Data
GBCU GBCNet	ECE	0.0802	0.147 ± 0.131	0.155 ± 0.090	0.145 ± 0.069	0.126 ± 0.032	0.101 ± 0.027
	SCE	0.0841	0.092 ± 0.090	0.101 ± 0.046	0.112 ± 0.043	0.113 ± 0.030	0.096 ± 0.015
	AECE	0.0722	0.012 ± 0.025	0.077 ± 0.039	0.088 ± 0.067	0.138 ± 0.039	0.114 ± 0.027
	ECE-KDE	0.3438	0.927 ± 0.236	0.531 ± 0.182	0.479 ± 0.151	0.338 ± 0.075	0.346 ± 0.055
	RECE-G	0.0607	0.049 ± 0.050	$\mathbf{0.062 \pm 0.043}$	$\mathbf{0.065 \pm 0.038}$	$\mathbf{0.066 \pm 0.024}$	$\mathbf{0.060 \pm 0.024}$
	RECE-M	0.0641	$\mathbf{0.066 \pm 0.081}$	0.080 ± 0.057	0.090 ± 0.045	0.075 ± 0.026	0.069 ± 0.019
BUSI ResNet-50	ECE	0.0726	0.113 ± 0.162	0.076 ± 0.092	0.066 ± 0.058	0.071 ± 0.042	0.067 ± 0.024
	SCE	0.0485	0.075 ± 0.109	0.051 ± 0.062	0.044 ± 0.039	0.049 ± 0.028	$\mathbf{0.047 \pm 0.015}$
	AECE	0.0531	0.005 ± 0.007	0.037 ± 0.055	0.045 ± 0.050	0.084 ± 0.045	0.070 ± 0.021
	ECE-KDE	0.1864	0.752 ± 0.401	0.530 ± 0.367	0.292 ± 0.155	$\mathbf{0.186 \pm 0.080}$	0.161 ± 0.064
	RECE-G	0.0301	0.040 ± 0.061	$\mathbf{0.028 \pm 0.034}$	$\mathbf{0.028 \pm 0.027}$	0.034 ± 0.023	0.027 ± 0.014
	RECE-M	0.0239	$\mathbf{0.031 \pm 0.053}$	0.029 ± 0.031	0.035 ± 0.033	0.032 ± 0.019	0.026 ± 0.011
POCUS ResNet-50	ECE	0.0280	0.134 ± 0.082	0.110 ± 0.032	0.089 ± 0.024	0.054 ± 0.016	0.040 ± 0.009
	SCE	0.0444	0.100 ± 0.080	0.091 ± 0.026	0.077 ± 0.016	0.064 ± 0.009	0.053 ± 0.004
	AECE	0.0324	0.078 ± 0.090	0.158 ± 0.045	0.104 ± 0.026	0.074 ± 0.016	0.058 ± 0.011
	ECE-KDE	0.2649	0.661 ± 0.267	0.326 ± 0.092	$\mathbf{0.285 \pm 0.069}$	$\mathbf{0.273 \pm 0.048}$	0.283 ± 0.028
	RECE-G	0.0111	0.046 ± 0.029	$\mathbf{0.042 \pm 0.021}$	0.041 ± 0.016	0.022 ± 0.010	$\mathbf{0.017 \pm 0.008}$
	RECE-M	0.0197	$\mathbf{0.052 \pm 0.046}$	0.059 ± 0.020	0.054 ± 0.017	0.036 ± 0.010	0.030 ± 0.009
Covid-CT ResNet-50	ECE	0.1586	0.184 ± 0.188	0.185 ± 0.110	0.143 ± 0.061	0.200 ± 0.044	0.175 ± 0.027
	SCE	0.1655	$\mathbf{0.184 \pm 0.188}$	0.188 ± 0.109	$\mathbf{0.153 \pm 0.059}$	0.208 ± 0.042	0.183 ± 0.026
	AECE	0.1351	0.089 ± 0.133	$\mathbf{0.123 \pm 0.109}$	0.155 ± 0.063	0.177 ± 0.037	0.164 ± 0.022
	ECE-KDE	0.3186	0.793 ± 0.333	0.423 ± 0.207	0.381 ± 0.128	0.382 ± 0.069	0.360 ± 0.039
	RECE-G	0.1390	0.071 ± 0.076	0.098 ± 0.082	0.073 ± 0.039	$\mathbf{0.146 \pm 0.040}$	$\mathbf{0.145 \pm 0.028}$
	RECE-M	0.1325	0.099 ± 0.115	0.116 ± 0.087	0.087 ± 0.046	0.155 ± 0.041	0.143 ± 0.026
Diabetic Retinopathy ResNet-50	ECE	0.0019	0.032 ± 0.008	0.013 ± 0.004	0.011 ± 0.003	0.005 ± 0.002	0.004 ± 0.001
	SCE	0.0228	0.045 ± 0.011	$\mathbf{0.028 \pm 0.006}$	$\mathbf{0.023 \pm 0.004}$	$\mathbf{0.023 \pm 0.002}$	$\mathbf{0.023 \pm 0.001}$
	AECE	0.0028	0.034 ± 0.007	0.024 ± 0.005	0.022 ± 0.003	0.020 ± 0.002	0.020 ± 0.002
	ECE-KDE	0.0458	0.102 ± 0.019	0.064 ± 0.009	0.053 ± 0.004	0.046 ± 0.004	$\mathbf{0.046 \pm 0.002}$
	RECE-G	0.0008	$\mathbf{0.015 \pm 0.007}$	$\mathbf{0.006 \pm 0.003}$	0.005 ± 0.003	0.003 ± 0.001	$\mathbf{0.001 \pm 0.001}$
	RECE-M	0.0017	0.018 ± 0.008	0.007 ± 0.004	0.006 ± 0.003	0.003 ± 0.001	0.002 ± 0.001
SIIM-ISIC Melanoma ResNet-50	ECE	0.1586	0.015 ± 0.013	0.010 ± 0.006	$\mathbf{0.013 \pm 0.005}$	$\mathbf{0.013 \pm 0.002}$	$\mathbf{0.013 \pm 0.002}$
	SCE	0.1655	0.015 ± 0.013	0.010 ± 0.006	0.013 ± 0.004	0.013 ± 0.002	0.013 ± 0.002
	AECE	0.1351	0.021 ± 0.014	0.017 ± 0.005	0.018 ± 0.004	0.015 ± 0.002	0.014 ± 0.002
	ECE-KDE	1.9551	1.965 ± 0.026	1.959 ± 0.011	1.958 ± 0.012	$\mathbf{1.955 \pm 0.005}$	$\mathbf{1.955 \pm 0.003}$
	RECE-G	0.1390	$\mathbf{0.007 \pm 0.006}$	$\mathbf{0.005 \pm 0.003}$	0.006 ± 0.002	0.006 ± 0.001	0.006 ± 0.001
	RECE-M	0.0129	0.015 ± 0.013	0.010 ± 0.006	$\mathbf{0.013 \pm 0.005}$	$\mathbf{0.013 \pm 0.002}$	$\mathbf{0.013 \pm 0.002}$
CIFAR-10 ResNet-50	ECE	0.0295	0.056 ± 0.016	0.040 ± 0.008	0.033 ± 0.005	0.032 ± 0.004	0.031 ± 0.002
	SCE	0.0070	0.015 ± 0.003	0.012 ± 0.002	0.010 ± 0.001	0.009 ± 0.001	0.008 ± 0.000
	AECE	0.0287	0.036 ± 0.014	0.030 ± 0.010	0.024 ± 0.006	0.024 ± 0.004	0.023 ± 0.002
	ECE-KDE	0.1232	$\mathbf{0.119 \pm 0.024}$	$\mathbf{0.123 \pm 0.012}$	0.125 ± 0.012	0.125 ± 0.007	$\mathbf{0.123 \pm 0.005}$
	RECE-G	0.0288	0.034 ± 0.018	0.033 ± 0.010	$\mathbf{0.028 \pm 0.005}$	$\mathbf{0.029 \pm 0.004}$	$\mathbf{0.029 \pm 0.002}$
	RECE-M	0.0287	$\mathbf{0.033 \pm 0.016}$	0.032 ± 0.010	0.026 ± 0.006	0.029 ± 0.004	0.029 ± 0.002
CIFAR-100 ResNet-50	ECE	0.0857	0.118 ± 0.021	0.090 ± 0.016	0.088 ± 0.011	0.089 ± 0.007	$\mathbf{0.086 \pm 0.004}$
	SCE	0.0025	0.006 ± 0.001	$\mathbf{0.005 \pm 0.000}$	0.005 ± 0.000	$\mathbf{0.004 \pm 0.000}$	0.003 ± 0.000
	AECE	0.0855	0.108 ± 0.015	0.076 ± 0.013	0.073 ± 0.007	0.071 ± 0.008	0.070 ± 0.005
	ECE-KDE	0.5161	1.120 ± 0.072	0.527 ± 0.042	0.496 ± 0.016	0.495 ± 0.012	0.506 ± 0.010
	RECE-G	0.0854	0.082 ± 0.021	0.081 ± 0.017	$\mathbf{0.086 \pm 0.011}$	0.088 ± 0.007	0.085 ± 0.004
	RECE-M	0.0854	$\mathbf{0.085 \pm 0.022}$	0.081 ± 0.017	$\mathbf{0.086 \pm 0.011}$	0.088 ± 0.007	0.085 ± 0.004

split the dataset into a binary classification problem. We show the generality of our method on natural image datasets with **CIFAR10 and CIFAR100** [13].

Experimental Setup: We use a ResNet-50 [7] model as a baseline for most of our experiments. The GBCU dataset is trained on the GBCNet architecture [2]. Both are intialized with ImageNet weights. We use SGD optimizer with weight decay $5e{-}4$, momentum 0.9 and step-wise LR decay with factor 0.1. We use LR 0.003 for GBCU and 0.01 for DR while the rest use 0.005. We train the models for 160 epochs with batch size 128. Horizontal flip is the only train-time augmentation used. For the `RECE` metric, we use $\sigma = 0.1$ and $M = 15$ bins for evaluation. For `RCR` loss we use $\beta = 1$ and focal loss as $\mathcal{L}_{\text{application}}$ with $\gamma = 1$.

Table 2. Comparison of calibration methods on medical datasets. On the more reliable RECE-G and RECE-M metric, calibrating with our regularizing loss term consistently achieves SOTA results.

Dataset & Model	Method	Acc.	ECE	SCE	RECE-G	RECE-M	Brier Score
GBC-USG GBCNet	Cross-Entropy	**0.9016**	**0.0802**	0.0841	0.0607	0.0610	0.2005
	FL [18]	0.8934	0.0913	0.0911	0.0644	0.0702	0.2084
	FL+MDCA [8]	0.8934	0.0810	**0.0811**	0.0508	0.0604	**0.1929**
	FLSD [19]	**0.9016**	0.1036	0.0909	0.0660	0.0581	0.1967
	DCA [17]	0.8934	0.0869	0.0905	0.0603	0.0433	0.1957
	RCR	0.8852	0.0810	0.0863	**0.0355**	**0.0372**	0.1957
BUSI ResNet-50	Cross-Entropy	0.9487	0.0726	**0.0485**	0.0301	0.0329	0.1000
	FL [18]	0.9487	0.0761	0.0613	0.0276	0.0415	0.1065
	FL+MDCA [8]	**0.9615**	0.0683	0.0505	0.0378	0.0358	**0.0977**
	FLSD [19]	**0.9615**	0.0939	0.0530	0.0493	0.0568	0.1028
	DCA [17]	0.9487	0.0669	0.0489	0.0337	0.0308	0.0988
	RCR	0.9231	**0.0608**	0.0680	**0.0201**	**0.0235**	0.1277
POCUS ResNet-50	Cross-Entropy	0.8845	0.0456	**0.0443**	0.0368	0.0452	0.1594
	FL [18]	0.8866	0.0488	0.0649	0.0301	0.0408	0.1584
	FL+MDCA [8]	0.8761	0.0646	0.0685	0.0598	0.0595	0.2014
	FLSD [19]	**0.9349**	0.0455	0.0541	0.0394	0.0444	**0.1085**
	DCA [17]	0.8782	0.0689	0.0567	0.0583	0.0607	0.1853
	RCR	0.8908	**0.0387**	0.0625	**0.0205**	**0.0339**	0.1522
Diabetic Retinopathy ResNet-50	Cross-Entropy	0.8641	0.0464	0.0889	0.0456	0.0458	0.2007
	FL [18]	0.8765	0.0369	0.0884	0.0365	0.0369	0.1813
	FL+MDCA [8]	**0.9351**	0.0442	**0.0752**	0.0439	0.0442	**0.1023**
	FLSD [19]	0.8620	0.0359	0.0896	0.0277	0.0216	0.1947
	DCA [17]	0.8332	0.0669	0.1171	0.0665	0.0669	0.2474
	RCR	0.8297	**0.0316**	0.0951	**0.0250**	**0.0206**	0.2400

Comparison between `RECE-G`, `RECE-M` and Other Metrics: Figure 2 and Table 1 give the comparison of different metrics computed over increasingly larger sample sets. For these experiments, we randomly sample the required sample size from the test set and compute metrics on them. The process is repeated 20 times and the average value is plotted with 95% confidence intervals. We plot the absolute difference with the baseline being the metric evaluated on the entire dataset. The results show that `RECE-G` and `RECE-M` outperform other metrics and are able to converge to the value computed from the whole dataset, using the smallest amount of data. The results for natural datasets are also shown.

Evaluation of Calibration Methods: We compare our `RCR` loss with other `SOTA` calibration techniques in Table 2. The results show that `RCR` is able to not only minimize our `RECE` metric but also other common metrics.

5 Conclusion

We demonstrated the ineffectiveness of existing calibration metrics for medical datasets with limited samples and propose a robust calibration metric to accurately estimates calibration independent of dataset size. We also proposed a novel loss to calibrate models using proposed calibration metric.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
2. Basu, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Surpassing the human accuracy: detecting gallbladder cancer from USG images with curriculum learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20886–20896 (2022)
3. Born, J., et al.: Accelerating detection of lung pathologies with explainable ultrasound image analysis. Appl. Sci. **11**(2), 672 (2021)
4. Carneiro, G., Pu, L.Z.C.T., Singh, R., Burt, A.: Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med. Image Anal. **62**, 101653 (2020)
5. 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)
6. 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)
7. 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)
8. Hebbalaguppe, R., Prakash, J., Madan, N., Arora, C.: A stitch in time saves nine: a train-time regularizing loss for improved neural network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16081–16090 (2022)
9. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: AugMix: a simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
10. Islam, M., Seenivasan, L., Ren, H., Glocker, B.: Class-distribution-aware calibration for long-tailed visual recognition. arXiv preprint arXiv:2109.05263 (2021)
11. Jiang, X., Osl, M., Kim, J., Ohno-Machado, L.: Calibrating predictive model estimates to support personalized medicine. J. Am. Med. Inform. Assoc. **19**(2), 263–274 (2012)
12. Kaggle, EyePacs: kaggle diabetic retinopathy detection, July 2015. https://www.kaggle.com/c/diabetic-retinopathy-detection/data
13. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
14. Kull, M., Perello Nieto, M., 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, vol. 32 (2019)
15. Kumar, A., Sarawagi, S., Jain, U.: Trainable calibration measures for neural networks from kernel mean embeddings. In: International Conference on Machine Learning, pp. 2805–2814. PMLR (2018)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Liang, G., Zhang, Y., Wang, X., Jacobs, N.: Improved trainable calibration method for neural networks on medical imaging classification. arXiv preprint arXiv:2009.04057 (2020)

18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
19. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Adv. Neural. Inf. Process. Syst. **33**, 15288–15299 (2020)
20. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)
22. Nguyen, K., O'Connor, B.: Posterior calibration and exploratory analysis for natural language processing models. arXiv preprint arXiv:1508.05154 (2015)
23. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops, vol. 2 (2019)
24. Patra, R., Hebbalaguppe, R., Dash, T., Shroff, G., Vig, L.: Calibrating deep neural networks using explicit regularisation and dynamic data pruning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1541–1549 (2023)
25. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)
26. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. **10**(3), 61–74 (1999)
27. Popordanoska, T., Sayer, R., Blaschko, M.: A consistent and differentiable LP canonical calibration error estimator. In: Advances in Neural Information Processing Systems, vol. 35, pp. 7933–7946 (2022)
28. Rajaraman, S., Ganesan, P., Antani, S.: Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS ONE **17**(1), e0262838 (2022)
29. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci. Data **8**(1), 34 (2021)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
31. Toledo-Cortés, S., de la Pava, M., Perdomo, O., González, F.A.: Hybrid deep learning gaussian process for diabetic retinopathy diagnosis and uncertainty quantification. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) OMIA 2020. LNCS, vol. 12069, pp. 206–215. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63419-3_21
32. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
33. Zhao, J., Zhang, Y., He, X., Xie, P.: COVID-CT-dataset: a CT scan dataset about COVID-19. arXiv preprint arXiv:2003.13865 (2020)