Multi-Rater Calibration Error Estimation

Meritxell Riera-Marín[1,2]*, Javier García López[1], Júlia Rodríguez-Comas[1], Miguel A. González Ballester[2,3], Adrian Galdran[4,5]

[1] Sycai Medical, Barcelona, Spain
[2] BCN Medtech, Dept. of Engineering, Universitat Pompeu Fabra, Barcelona, Spain
[3] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[4] Tecnalia, Derio, Spain
[5] Ikerbasque, Basque Foundation for Science, Bilbao, Spain.

Abstract. Calibration, the property of producing predicted probabilities that reflect true likelihoods of outcomes, is a relevant attribute of medical image computing models and a key requirement in clinical decision-making. However, empirical Calibration Error (CE) estimates suffer from instability in data-scarce scenarios. Here, for any existing CE we propose a Multi-Rater version of it (MR-CE), a wrapper over conventional calibration metrics, which provides a new strategy for estimating a CE that effectively addresses this limitation in situations where there are multiple annotations per sample. MR-CEs offer more consistent estimates of calibration errors by leveraging the consensus and disagreement among multiple annotators to generate virtually extended test datasets, more robust to typical binning artifacts. We evaluate a MR version of the popular Expected Calibration Error (ECE), and also of the more recent Kernel Density Estimation-ECE (kdeECE), in a comprehensive set of classification and segmentation problems, demonstrating improved stability compared to their single-rater CE counterparts. Specifically, we show that MR-CEs achieve a reduced variability as the test set size decreases across all analysed datasets. Our findings emphasize the critical role of modelling inter-rater variability not only for training but also for evaluating medical image analysis models, in particular when studying the calibration of modern neural networks.

Keywords: Model Calibration · Uncertainty Quantification · Multi-Rater Modelling.

1 Introduction

Calibration refers to the ability of a model to formulate probabilistic predictions aligned with its own accuracy, ensuring that lower predictive confidence is truly related to less likelihood of being correct. In clinical contexts, where decisions often hinge on a model's probabilistic outputs, poorly calibrated predictions can lead to overconfidence in incorrect classifications or underutilization of accurate

* Corresponding author: m.riera@sycaitechnologies.com

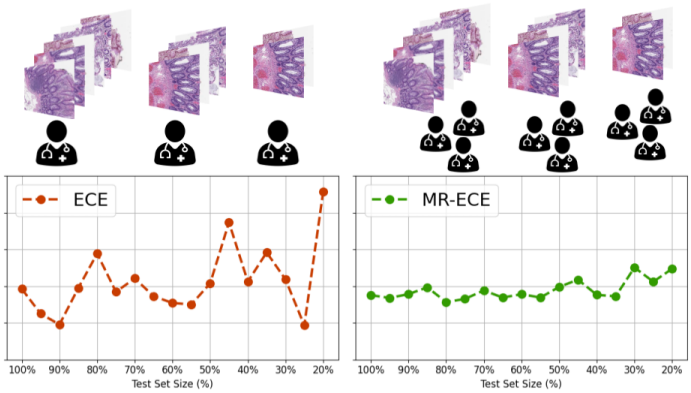

Fig. 1: As the test size decreases, conventional empirical estimates of Calibration Error (ECE, left) become highly unstable. In contrast, adding multi-rater information (MR-ECE, right) makes the estimates robust to decreasing test set size.

models [4]. This is significant in medical image analysis tasks, where data variability and imbalance, or inter-rater differences introduce unique challenges [16].

Calibration is not a binary property of a predictive model, but rather there is a wide variety of miscalibration modes. For example, a model could be extremely accurate with low confidences, exceedingly overconfident but inaccurate, or anywhere in between. Measuring the degree of miscalibration is a far from straightforward process, since we need to estimate its accuracy at each level of confidence, and some regions of the confidence space can be very sparsely populated [13]. In order to improve calibration error estimates, we would need to have large quantities of test data with a representative spectrum of model confidences associated to it, which is often unfeasible. On the other hand, common calibration measures have been shown to be fragile and unstable with respect to the size of test data [6], as illustrated in Fig. 1. Therefore, developing better calibration error metrics is a critical step in order to build more robust uncertainty quantification techniques in data-limited scenarios.

When compared to improving model calibration, the problem of measuring calibration error has traditionally been much less explored. Estimating a Calibration Error (CE) requires comparing model confidence with accuracy by grouping test samples based on confidence levels. The Expected Calibration Error (ECE), detailed in the next section, uses equal-width binning of confidence values and is the most widely used approach due to its simplicity. However, it has been shown to produce biased estimates, leading to various alternative CE measures. Nixon et al. [13] improved error robustness by discarding extreme-confidence samples, while Roeloffs et al. [15] reduced statistical bias by using equal-sized bins instead of equal-width ones. Arrieta-Ibarra et al. [2] reformulated ECE using cumulative probability distributions. Other approaches address binning sensitivity through Proper Scoring Rules [5] or Kernel-Density Estimators [14,24]. A separate family of methods relies on statistical (frequentist) testing [22]. For instance, Tygert

et al. [18] analyzed cumulative differences between sorted confidence scores and corresponding labels, comparing them to Brownian motion. Similar statistical techniques have been developed for machine learning models [2,19].

In this work, we propose to make use of multi-rater annotations to improve CE estimates. Data labelled by multiple annotators is often present when researchers need to control for inter-rater variability [8], or when the training data is suspected to contain noisy labels [3]. In these cases, test data is typically acquired by collecting several opinions, and then forming some sort of consensus that is taken as the gold-standard [1,20]. Instead of using prevalent merging techniques [20], we argue that for a given sample, treating each label independently results in richer, more representative test sets. Our experiments on a wide range of medical image analysis datasets and tasks show that this simple observation allows to formulate a multi-rater version of common calibration errors which is noticeably more robust to the size of the test set, even when data is scarce.

2 Methodology

2.1 Notation, Definitions, and Problem Formulation

We consider a binary classifier $\mathcal{U}$ trained on samples (x, y), where x is an image, and $y \in \{0, 1\}$ is its category. Often, binary classifiers are built as models with a single numerical output, $\mathcal{U}(x) = c \in [0, 1]$. This number is typically referred to as *confidence*, and interpreted as a probability for x to belong to the positive class. However, it is not necessarily the case that $c \approx p(y = 1|x)$, and when this property does not hold we say that the model is miscalibrated.

There is a continuous spectrum of miscalibration degrees a model can suffer, often referred to as its Calibration Error. Unfortunately, measuring the gap between model confidence and actual probabilities is a deceptively complex and brittle process, due to finite-sample limitations. Firstly, (mis)calibration must be measured on held-out data that has not been used for training, to avoid biased estimates. Second, since we cannot estimate $p(y_i = 1|x_i)$ over a single item x_i, we need to group samples. If we use M uniformly distributed bins B_m, we end up with per-bin average confidences and positive proportions defined as:

$$\mathrm{conf}(B_m) = \frac{1}{b_m} \sum_{x \in B_m} c(x), \quad \mathrm{pos}(B_m) = \frac{1}{b_m} \sum_{x \in B_m} \mathbb{1}(x), \tag{1}$$

being $\mathbb{1}(x)$ a characteristic function returning 0 unless $y_x = 1$, in which case it takes a value of 1. These can then be used to define the Expected Calibration Error, the most popular empirical estimate of a model's CE.

$$\mathrm{ECE}(X_{\mathrm{test}}) = \sum_{m=1}^{M} \frac{b_m}{M} \, |\mathrm{conf}(B_m) - \mathrm{pos}(B_m)|. \tag{2}$$

There are multiple ways of extending Eq. (2) to a multi-class setting, the most favoured being to consider a model's prediction as the category with maximum probability $\hat{y} = \mathrm{argmax}(c_i)$, define its confidence as that probability, and

compare bin-wise accuracy ($\text{acc}(B_m) = \frac{1}{b_m} \sum_{x \in B_m} \mathbb{1}(y_x = \hat{y}_x)$) to confidence:

$$\text{ECE}(X_{\text{test}}) = \sum_{m=1}^{M} \frac{b_m}{M} \, |\text{conf}(B_m) - \text{acc}(B_m)|. \tag{3}$$

When measuring calibration for segmentation problems, these are often considered as pixel-wise classification tasks and the above definitions remain valid. Unfortunately, and regardless of the considered problem, there is often conflict between increasing M, therefore building narrower bins (which would result in finer but less robust accuracy estimates), and decreasing M, thereby reducing bin resolution, obtaining more robust probability estimates, but paying the price of coarser confidence bands. Several alternative ways of estimating calibration errors have been proposed in the literature, e.g. replacing binning and histograms by kernel density estimates (kdeECE, [14]), or computing a CE class-wise in a one vs rest manner and then averaging the result (cwECE, [10]). However, the limited size of test sets in machine learning problems remains a challenge for any kind of CE estimation. In the next section, we introduce a simple but extremely effective procedure to improve the robustness of calibration error estimates whenever there is data annotated by multiple raters.

2.2 Leveraging Multiple Annotations to Improve Calibration Error Estimates: Multi-Rater Calibration Error

Let us consider a scenario where each sample x in the test set has multi-rater annotations $\{y_1, ..., y_{R_x}\}$, being R_x the number of raters who annotated sample x. In our discussion, R_x can vary with respect to each sample, and we do not assume that we know the identity of each annotator; indeed, different subsets of annotators from a large pool of raters may label the data without affecting our approach. We also do not assume that any annotator has greater skill or expertise than the other, but rather all labels are equally valid. In other words, they can be regarded as samples of a posterior distribution of the true ground-truth, with a higher dispersion reflecting greater aleatoric uncertainty. Finally, we ignore potential dependence structures between samples and multiple annotations. This is a strong assumption, since knowing the label assigned by one rater to a sample can affect the potential label a second rater will attribute to it. However, since this happens in a sample-wise manner, we expect this kind of dependence to affect estimation quality in a similar way across bins, limiting its impact.

The conventional approach to exploit multiple annotations would be to combine labels in a particular way, e.g. using some consensus strategy, or smoothing labels. On the other hand, the most important limitation for estimating the calibration error is the size N of the test set, and merging annotations does not address this issue. Instead, we propose to extend the definition of bin to accommodate multiple, equally valid labels as follows:

$$B_m^{\text{MR}} = \{(x, y_j) \in X_{\text{test}} \mid c(x) \in I_m, 1 \leq j \leq R_x\}, \tag{4}$$

where (x, y_j) denotes a test instance x paired with the label provided by the $j-th$ annotator, X_{test} is the test dataset comprising all such (x, y_j) pairs across all instances and annotaors, $c(x)$ represents the predicted confidence for instance x, I_m is the confidence interval corresponding to bin m, and R_x denotes the number of annotators who labeled instance x.

In words, we do not fuse the multiple annotations associated to x, but rather iterate the presence of x in the test set, each time with a different label. Although "new" samples will share the same model confidence $c(x)$, and therefore lie in the same bin, each of them can have a different label y_j, resulting in useful information to estimate $\text{pos}(B_m^{\text{MR}})$ or $\text{acc}(B_m^{\text{MR}})$ in Eqs. (2) and (3) more robustly. Therefore, a binary Multi-Rater Expected Calibration Error can be written as:

$$\text{MR-ECE}(X_{\text{test}}) = \sum_{m=1}^{M} \frac{b_m^{\text{MR}}}{M} \, |\text{conf}(B_m^{\text{MR}}) - \text{pos}(B_m^{\text{MR}})|, \tag{5}$$

with an analogous formulation for the multi-class MR-ECE.

Despite its simplicity, we argue that introducing new samples when estimating a CE effectively allow us to control the instability of that CE with respect of the test set size. Note that we could directly build an MR counterpart of any existing CE, since the idea of repeated sampling with different labels is independent of the choice of estimate. It is also important to stress that an inter-rater smoothing of labels would be far from straightforward to implement in this context, as the terms $\text{pos}(B_m)$ in Eq. (2) and $\text{acc}(B_m)$ in Eq. (3) would become ill-defined when using soft labels.

2.3 Evaluating Calibration Errors

In an experimental analysis, we would like to assess if a CE estimate (i.e., an empirical approximation such as the ECE) approximates better the real (unkown) CE in the absence of enough test data. Of course, we only have access to empirical estimates of the real CE, but we can safely assume that such estimates will be better (or at least non-inferior) with more data. Based on this premise, we propose to measure stability with respect to the test set size by observing variation of an estimate as the number of test samples decreases. We achieve this by measuring the Total Variation of the CE estimate: we define a uniform partition $[p_0, p_0 + p, ..., 1 - p, 1]$ of the unit interval, and a random subset of X_{test} of X_{test}^p consisting of a fraction p of the initially available samples (we typically set $p_0 = 0.2$ and $p = 0.05$, see Supp. Material). With this, we compute:

$$\text{TV}(\text{CE}) = \frac{1}{n} \sum_{p} \text{CE}(X_{\text{test}}^p) - \text{CE}(X_{\text{test}}^{p-1}), \tag{6}$$

where X_{test}^p denotes a randomly sampled subset of the test set containing a proportion $p \in [p_0, 1]$ of the original data, and $\text{CE}(X_{test}^p)$ is the calibration error computed over that subset. The sum is taken over a uniform partition of the interval $[p_0, 1]$ with step size Δp, and n corresponds to the number of steps in

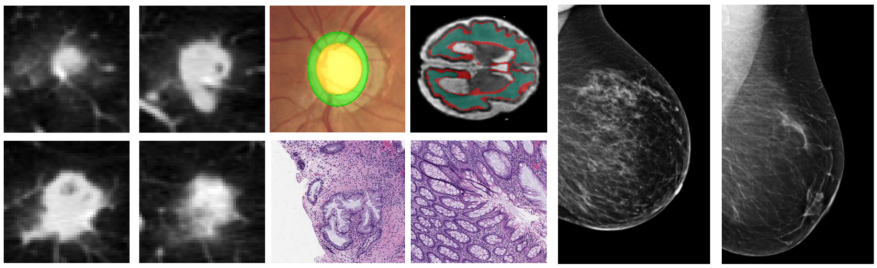

Fig. 2: Samples of the different classification and segmentation datasets employed for validating the proposed MR-CE estimation strategy. (a) LIDC-IDR, (b) CHÁKṢU/QUBIQ Brain Growth (top), MHIST (bottom), (c) CSAW-M.

this partition. TV(CE) gives us an understanding of the stability of a Calibration Error estimate when the test set size decreases. if the CE is robust to smaller quantities of data, its total variation should be less, but for a weaker CE estimate, reducing test set size will produce estimates with greater variability, resulting in a higher TV.

It is worth noting that how to measure calibration of segmentation models is an open research question, with no clear answer in the literature. Simply concatenating all pixels in all images of a test set and computing a classification-like CE is a process bound to quickly saturate computational memory. The obvious alternative is to compute a single CE estimate per image, and then report the average. Then, each test image can be considered as an independent test set, and issues such as class/foreground imbalance may bias estimates, leading us to average apples with oranges in this situation. Despite these considerations, we adopt the latter strategy since it is the most favoured in the literature [9].

3 Experimental Results

3.1 Datasets and Model Training

We conduct our evaluation on several binary/multi-class classification and segmentation tasks over a variety of data, see Fig. 2. Further details are given below.

1) MHIST[1], a histopathology dataset where seven experts label each image as either Hyperplastic Polyp or Sessile Serrated Adenoma, a binary classification task with significant inter-pathologist variability [21]. **2) CSAW-M**[2], which includes mammograms from over 10,000 individuals, annotated with eight masking categories [17]. Masking, caused by breast tissue density, can obscure tumors in mammography. Test set mammograms are annotated by five experts.

[1] https://bmirds.github.io/MHIST/
[2] https://github.com/yueliukth/CSAW-M

3) LIDC-IDRI[3], containing CT scans with nodule locations, malignancy ratings, and texture ratings from three to seven annotators [1]. In this dataset, 3D nodule textures are distributed into four different classes; we use 1011 lesions extracted with a 48×48×32 bounding cube. **4) QUBIQ**[4] spans multiple subdatasets [12]: Brain Growth (34 multimodal MRI scans annotated by seven raters), Brain Tumor (28 images across three tasks, labeled by three annotators), Kidney (24 CTs with a single task, labeled by three annotators), and Prostate (55 MRI cases, two segmentation tasks, annotated by six raters). We do not use the QUBIQ-Pancreas 3d subdatasets due to inconsistent slice thickness. The remaining QUBIQ subdatasets only contain 2d slices. **5) CHÁKṢU**[5]: a dataset containing 1,345 retinal fundus images with optic disc and cup outlines annotated by five expert ophthalmologists [11]. Cup-to-disc ratios are clinically relevant for glaucoma diagnosis and show significant inter-rater variability [23].

For classification, we use ResNet architectures, with 3D inputs processed via video-based variants where convolutional kernels incorporate the third spatial dimension. For segmentation, we employ encoder-decoder nets with a ResNet50 backbone pretrained on ImageNet and a Feature Pyramid Network decoder [7]. Since our focus is on calibration rather than discriminative performance, we do not optimize hyperparameters extensively. The only dataset-dependent design choice is training length, and we adapt the number of epochs to the size of the dataset. We use the n-adam optimizer to minimize a CE loss, with a batch size of 8 (4 for segmentation), a learning rate of 1e-4, and a cosine-decay schedule. We apply early stopping to select the best model based on validation performance.

3.2 Numerical Analysis of the Stability of CE Estimates

We first conduct a visual evaluation of the stability of calibration error estimates across three classification tasks, comparing the conventional CE with its Multi-Rater version (MR-ECE). We also consider the Kernel Density Estimate-CE and its MR counterpart. After training each classifier, we progressively reduce the test set and compute both measures on the remaining data, expecting that a robust CE estimate remains stable as test samples decrease. We see in Fig. 3 how both MR-ECE and MR-kdeECE achieve this, whereas standard ECE and kdeECE exhibit sharp peaks and fluctuations.

To quantify CE stability with respect to test set size numerically, we compute the Total Variation (TV) of each calibration measure in Table 1. In order to obtain dispersion measures, we bootstrap 100 times the test sets by random sampling with replacement, averaging results across iterations. Our results confirm that MR-ECE and MR-kdeECE consistently achieve lower TV, reducing TV by a factor of 1.5–4. Table 1 highlights MR-CE's superiority, showing reduced variability and stronger stability in low-data scenarios.

For segmentation problems, each test image served as an independent test set, with CE estimates averaged to report mean and standard deviation. Total

[3] https://www.cancerimagingarchive.net/collection/lidc-idri/
[4] https://qubiq21.grand-challenge.org/participation/
[5] https://doi.org/10.6084/m9.figshare.20123135

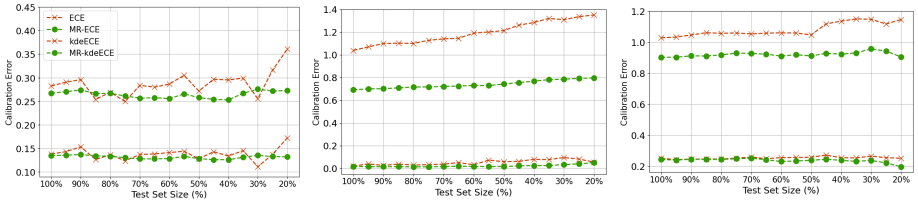

Fig. 3: ECE, kdeECE and their MR-versions computed as the test set size decreases for classification datasets MHIST (left), CSAW-M (middle), LIDC-IDRI.

Table 1: Total Variation for Single Rater (**SR**) and Multi-Rater (**MR**) Calibration Errors (ECE and kdeECE) for three classification problems.

	MHIST		**CSAW-M**		**LIDC-IDRI**	
	ECE	kdeECE	ECE	kdeECE	ECE	kdeECE
SR	0.25±0.06	0.47±0.01	0.32±0.07	0.41±0.06	0.13±0.03	0.26±0.05
MR	0.04±0.01	0.07±0.02	0.05±0.01	0.13±0.02	0.10±0.02	0.15±0.03

Table 2: Total Variation of the Single Rater (**SR**) and Multi Rater (**MR**) ECE. Optic Disc/Cup (Chákṣu) and several binary segmentation tasks in QUBIQ.

	Chákṣu	**Kidney**	**Prostate T1**	**Prostate T2**
SR	1.15±0.44	0.11±0.09	0.15±0.06	0.08±0.06
MR	0.54±0.20	0.01±0.01	0.01±0.01	0.01±0.01
	Brain Growth	**Brain Tumor T1**	**Brain Tumor T2**	**Brain Tumor T3**
SR	0.85±0.22	1.06±0.58	1.08±0.69	0.84±0.15
MR	0.69±0.21	0.50±0.08	0.94±0.18	0.40±0.02

Variation values for the ECE and the MR-ECE[6] are shown in Table 2. They follow the same trend as in classification, which again verify that the multi-rater version of a CE provides a more reliable calibration measure as test data availability decreases.

4 Conclusions and Future Work

In this work, we have introduced a new mechanism for estimating the Calibration Error (CE) of a model when multiple annotations are available, enabling

[6] In segmentation problems, the kde-ECE was too memory-demanding to compute.

a more nuanced approach to assessing performance of uncertainty estimates in this particular scenario. Our proposed Multi-Rater Calibration Error (MR-CE) virtually expands the size of the test set by independently considering each data point with a different label as a single sample. As a consequence, an MR-CE achieves greater robustness to small test set sizes than its single-rater counterpart. Experiments on an array of medical image classification and segmentation tasks support our hypothesis and confirm the greater reliability of our proposed CE estimates. Another advantage of the proposed approach is that it is orthogonal to the improvement of the adopted base CE empirical estimates. This means that if a practitioner can design a superior CE estimator, they would still be able to benefit from our method whenever they have data with multiple annotations.

With a more reliable CE measure established, future work will focus on training methods that enhance model calibration using multi-label annotations. Additionally, exploring new stability metrics for CE beyond TV will provide deeper insights into their reliability across models and datasets. This research line could refine model evaluation practices, leading to more robust and interpretable assessment methodologies, especially in multi-rater scenarios.

Acknowledgments

This work was supported by the Catalan Government through the Industrial Doctorates program (AGAUR 2021-063), in collaboration with Sycai Technologies SL. It is part of the CPP project (CPP2021-008364), funded by `MCIN/AEI/10.13039/501100011033` and co-financed by the European Union (NextGenerationEU/PRTR). A.G. is supported by the RYC2022-037144-I grant, funded by `MCIN/AEI/10.13039/501100011033` and co-financed by FSE+.

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

References

1. Armato III, S.G., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Medical Physics **38**(2), 915–931 (2011). https://doi.org/10.1118/1.3528204
2. Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M., Xu, C.: Metrics of Calibration for Probabilistic Predictions. Journal of Machine Learning Research **23**(351), 1–54 (2022)
3. Bucarelli, M.S., Cassano, L., Siciliano, F., Mantrach, A., Silvestri, F.: Leveraging Inter-Rater Agreement for Classification in the Presence of Noisy Labels. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3439–3448 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00335, iSSN: 2575-7075

4. Chua, M., Kim, D., Choi, J., Lee, N.G., Deshpande, V., Schwab, J., Lev, M.H., Gonzalez, R.G., Gee, M.S., Do, S.: Tackling prediction uncertainty in machine learning for healthcare. Nature Biomedical Engineering **7**(6), 711–718 (Jun 2023). https://doi.org/10.1038/s41551-022-00988-x
5. Gruber, S.G., Buettner, F.: Better Uncertainty Calibration via Proper Scores for Classification and Beyond. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)
6. Huang, J., Brennan, D., Sattler, L., Alderman, J., Lane, B., O'Mathuna, C.: A comparison of calibration methods based on calibration data size and robustness. Chemometrics and Intelligent Laboratory Systems **62**(1), 25–35 (2002). https://doi.org/https://doi.org/10.1016/S0169-7439(01)00211-8
7. Iakubovskii, P.: Segmentation Models Pytorch (2019), https://github.com/qubvel/segmentation_models.pytorch, publication Title: GitHub repository
8. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12336–12346 (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01216
9. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Frontiers in Neuroscience **14** (Apr 2020). https://doi.org/10.3389/fnins.2020.00282
10. 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)
11. Kumar, J.R.H., Seelamantula, C.S., Gagan, J.H., Kamath, Y.S., Kuzhuppilly, N.I.R., Vivekanand, U., Gupta, P., Patil, S.: Chákṣu: A glaucoma specific fundus image database. Scientific Data **10**(1), 70 (Feb 2023). https://doi.org/10.1038/s41597-023-01943-4
12. Li, H.B., et al.: QUBIQ: Uncertainty Quantification for Biomedical Image Segmentation Challenge (Jun 2024). https://doi.org/10.48550/arXiv.2405.18435
13. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring Calibration in Deep Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (Jun 2019)
14. Popordanoska, T., Sayer, R., Blaschko, M.B.: A consistent and differentiable Lp canonical calibration error estimator. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 7933–7946 (Apr 2022)
15. Roelofs, R., Cain, N., Shlens, J., Mozer, M.C.: Mitigating Bias in Calibration Error Estimation. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. pp. 4036–4054. PMLR (May 2022)
16. Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P.: Classifier calibration: a survey on how to assess and improve predicted class probabilities. Machine Learning **112**(9), 3211–3260 (Sep 2023). https://doi.org/10.1007/s10994-023-06336-7
17. Sorkhei, M., Liu, Y., Azizpour, H., Azavedo, E., Dembrower, K., Ntoula, D., Zouzos, A., Strand, F., Smith, K.: CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)

18. Tygert, M.: Calibration of P-values for calibration and for deviation of a subpopulation from the full population. Advances in Computational Mathematics **49**(5), 70 (Sep 2023). https://doi.org/10.1007/s10444-023-10068-6
19. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., Schön, T.: Evaluating model calibration in classification. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 89, pp. 3459–3467. PMLR (Apr 2019)
20. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging **23**(7), 903–921 (Jul 2004). https://doi.org/10.1109/TMI.2004.828354
21. Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Tomita, N., Torresani, L., others: A Petri Dish for Histopathology Image Analysis. In: International Conference on Artificial Intelligence in Medicine. pp. 11–24 (2021)
22. Widmann, D., Lindsten, F., Zachariah, D.: Calibration tests in multi-class classification: A unifying framework. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
23. Wundram, A.M., Fischer, P., Wunderlich, S., Faber, H., Koch, L.M., Berens, P., Baumgartner, C.F.: Leveraging Probabilistic Segmentation Models for Improved Glaucoma Diagnosis: A Clinical Pipeline Approach. In: Medical Imaging with Deep Learning (2024)
24. Zhang, J., Kailkhura, B., Han, T.Y.J.: Mix-n-Match : Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning. In: Proceedings of the 37th International Conference on Machine Learning. pp. 11117–11128. PMLR (Nov 2020)