Evaluating Confidence Calibration in Endoscopic Diagnosis Models

Nikoo Dehghani[1] Ayla Thijssen[2,3] Quirine E. W. van der Zander[2,3] Ramon-Michel Schreuder[3,4]
Erik J. Schoon[3,4] Fons van der Sommen[1,5] Peter H. N. de With[1]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Maastricht University Medical Center+, Maastricht, The Netherlands
[3]GROW Research Institute for Oncology and Reproduction, Maastricht, The Netherlands
[4]Catharina Hospital, Eindhoven, The Netherlands
[5]Eindhoven Artificial Intelligence Systems Institute, Eindhoven, The Netherlands

Abstract

Colorectal polyps are prevalent precursors to colorectal cancer, making their accurate characterization essential for timely intervention and patient outcomes. Deep learning-based computer-aided diagnosis (CADx) systems have shown promising performance in the automated detection and categorization of colorectal polyps (CRP) using endoscopic images. However, alongside the advancement in diagnostic accuracy, the need for reliable and accurate quantification of uncertainty estimates within these systems has become increasingly important. The primary focus of this study is on refining the reliability of computer-aided diagnosis of CRPs within clinical practice. We perform an investigation of widely used model calibration techniques and how they translate into clinical applications, specifically for CRP categorization data. The experiments reveal that the Variational Inference method excels in intra-dataset calibration, but lacks efficiency and inter-dataset generalization. Laplace approximation and temperature scaling methods offer improved calibration across datasets.

1. Introduction

In recent years, the integration of artificial intelligence (AI) into medical imaging has shown remarkable promise, particularly in the realm of colorectal cancer (CRC) diagnosis and management. CRC is a significant global health concern, as it constitutes over 10% of cancer diagnoses and more than 9% of cancer-related deaths worldwide [2]. Early detection and accurate characterization of colorectal polyps, which are precursors to CRC, are pivotal in reducing the related morbidity and mortality rates. Colonoscopy is a cornerstone procedure in CRC screening and offers an opportunity for early diagnosis and prevention [4, 6].

Deep learning-based computer-aided diagnosis (CAD) systems, have emerged as promising tools to enhance the accuracy and efficiency of polyp detection and characterization during colonoscopy. By leveraging vast datasets and complex algorithms, these systems aim to complement human expertise and facilitate more reliable polyp characterization, thereby ultimately improving patient outcomes. However, deep learning-based systems prove to be vulnerable to perturbations and quality degradation [12, 15] and may produce overconfident predictions on outlier data [19]. Additionally in the medical domain, these systems can produce excessively confident predictions, as a result of the lack of calibration, therefore creating harmful biases on physicians' decisions. As an ultimate consequence, this can become life-threatening in a clinical setting, which is detrimental to optimal human-AI collaboration [8]. Moreover, the computational complexity inherent in deep models poses a challenge in obtaining calibrated predictions suitable for real-time integration into clinical systems. [23]

The concept of model calibration refers to the relationship between the accuracy of predictions and their confidence: a well-calibrated model will be less confident when making wrong predictions and more confident when making correct predictions [11]. Confidence calibration measurements provide a more complete understanding of the performance of a model by estimating how closely the confidence matches the accuracy. In the concept of deployment in the clinical setting, the confidence of model predictions becomes a critical component. Patient prognosis may be adversely affected if poorly calibrated models produce confident predictions for incorrect diagnoses. It is important that appropriate methods for model calibration are developed and evaluated for medical imaging applications to facilitate accurate risk assessment, providing clinicians with insights into potential fluctuations in accuracy levels under various conditions. Additionally, it enables model to refrain from making decisions when its confidence levels are low.

Considering these limitations, the main contribution can be summarized as follows. We study the relatively unexplored direction of calibrating modern polyp classification in endoscopic imaging. We perform experiments spanning both intra-dataset and inter-dataset scenarios, aimed at assessing calibration performance. The experiments highlight the strengths and weaknesses of each approach. While SVI demonstrates superior performance in intra-dataset experiments, it falls short in terms of test-time efficiency and inter-dataset generalization. Conversely, Laplace approximation and temperature scaling methods showcase improved calibration across datasets.

2. Background

Calibrating deep neural networks involves ensuring that the predicted probabilities accurately reflect the true likelihoods providing the correct predictions. For this purpose, two main approaches can be identified: (1) post-processing calibration, and (2) train-time calibration, which are further elaborated below.

Post-processing calibration methods: In addressing the challenge of calibration in deep learning-based computer-aided diagnosis (CAD) systems for colorectal polyp characterization, post-processing calibration methods offer a straightforward approach. Temperature scaling (TS), a prominent technique, adjusts the confidence scores output by the CAD system by dividing them by a learned temperature parameter, typically trained on a holdout validation set. Although effective, TS may decrease the confidence of the entire confidence vector, including the correct class, posing limitations in certain scenarios [5, 17]. Guo *et al.* [11] proposed a differentiable approximation of the expected calibration error (ECE) and integrated it into a meta-learning framework to achieve well-calibrated models. Similarly, Islam *et al.* [14] demonstrated class-distribution-aware calibration using a combination of temperature scaling (TS) and label smoothing for long-tailed visual recognition tasks. However, reliance on hold-out validation sets for TS methods can be impractical in real-world scenarios. Additionally, the last-layer Laplace approximation [21] stands out as a notable approach. This method involves approximating the posterior distribution over the last layer weights of the neural network using Laplace approximation. Although computationally intensive, the last-layer Laplace approximation offers a principled approach to refining the calibration of CAD system predictions and enhancing their reliability [16].

Train-time calibration techniques: Train-time calibration techniques aim to enhance the calibration of models during the training phase. Models optimized using negative log-likelihood (NLL) often exhibit a tendency towards overconfident predictions, as they prioritize minimizing errors in predicted probabilities without considering uncertainty. [11] To mitigate this issue, researchers have proposed incorporating probabilistic methods with Bayesian and non-Bayesian formulations for better representation of model parameters. Furthermore, probabilistic methods leveraging Bayesian formalism aim to tackle calibration issues by estimating predictive uncertainty. Approximate inference methods, including stochastic variational inference (SVI) and expectation propagation (EP), present practical avenues for deriving posterior distributions over neural network parameters [10]. Nonetheless, these methods come with the trade-off of augmenting the number of training parameters and incurring additional computational expenses.

In the experiments conducted to evaluate the calibration of CAD systems for colorectal polyp characterization, we employ a comprehensive array of methods spanning both post-processing and train-time calibration categories. By leveraging techniques such as temperature scaling, Laplace approximation, and variational inference approaches, we aim to provide a thorough assessment of calibration performance in this context.

3. Method and metrics

3.1. Confidence calibration methods

To investigate reliable and efficient calibration methods for deep learning-based CADx systems in colorectal polyp characterization, we examine inference time and calibration performance of a train-time approach, namely (A) variational inference, followed by post-processing calibration techniques including (B) Laplace approximation and (c) temperature scaling.

A. Stochastic variational inference approximation is utilized to approximate the posterior distribution over model parameters in Bayesian neural networks (BNNs)[1, 10, 22]. This method involves optimizing a variational objective, known as the Evidence Lower Bound (ELBO), to approximate the complex posterior distribution. The ELBO loss function, formulated as:

$$\text{ELBO} = \mathbb{E}_{q(\theta)}[\log p(y|x,\theta)] - \text{KL}[q(\theta)||p(\theta)], \quad (1)$$

attempts to minimize the conditional output probability when the prior distribution is known. In this equation, $q(\theta)$ represents the approximate posterior distribution, $p(y|x,\theta)$ denotes the likelihood of the data given the model parameters, and $p(\theta)$ signifies the prior distribution over model parameters.

B. Laplace approximation is another posterior approximation method that uses a Gaussian distribution centered at the maximum of the posterior distribution. Mathematically, the Laplace approximation of the posterior distribution $q(\theta)$ is given by:

$$q(\theta) \approx \mathcal{N}(\theta^*, \Sigma^{-1}), \quad (2)$$

where θ^* represents the mode of the posterior distribution, and Σ denotes the Hessian matrix of the negative log posterior evaluated at θ^*. As suggested by Daxberger *et al.* [7], post-hoc applying the Laplace approximation only to the last layer, instead of using all weights, typically yields better performance due to less underfitting and is significantly easier to compute. The Laplace approximation for the posterior distribution of the weights $W^{(L)}$ of the last layer (L) in a neural network given the data D can be expressed by:

$$p(W^{(L)}|D) \approx \mathcal{N}(W^{(L)}|W^{(L)}_{\text{MAP}}, \Sigma^{(L)}). \tag{3}$$

Here, the term $W^{(L)}_{\text{MAP}}$ represents the maximum a-posteriori (MAP) estimate of the weights of the last layer and $\Sigma^{(L)}$ denotes the covariance matrix associated with the Laplace approximation for the last-layer weights.

C. Temperature scaling is a post-processing technique and is employed to refine the calibration of CADx system predictions. Temperature scaling involves scaling the logits output by the CAD system, using a learned scalar temperature parameter. Mathematically, the temperature scaling transformation is expressed as:

$$\text{softmax}(z_i/T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \tag{4}$$

where z_i denotes the logit output corresponding to class i, and T represents the temperature parameter.

Through experimentation and evaluation, we aim to identify the most suitable calibration approach for enhancing the performance of CADx systems in clinical practice. Additionally, we seek to investigate the models' ability to generalize and remain robust when faced with data sourced from a different center, by conducting inter-dataset experimentation.

3.2. Miscalibration metrics

A perfectly calibrated model for a characterization task outputs class confidences that match with the predictive accuracy. If the accuracy is less than the confidence, then the model is overconfident and if it is higher then the model is underconfident. In assessing the calibration performance of CADx systems for colorectal polyp characterization, several calibration metrics provide quantitative measures of the alignment between prediction confidence and accuracy. One commonly used metric is the Expected Calibration Error (ECE) [11, 18], which quantifies the discrepancy between predicted confidence levels and empirical accuracy. Mathematically, the ECE is calculated as the weighted average of the absolute difference between observed accuracy and predicted confidence, where histogram bins are formed based on predicted confidence intervals. Formally, the ECE is expressed as [18]:

$$E_{\text{ECE}} = \frac{1}{N}\sum_{i=1}^{N} N_i \cdot |C_i - A_i|, \tag{5}$$

where N is the total number of samples, N_i represents the number of samples in the i^{th} confidence interval, C_i denotes the predicted confidence in the i^{th} interval, and A_i signifies the empirical accuracy in the i^{th} interval.

Another widely utilized calibration metric is the Brier Score, which measures the mean-squared difference between predicted probabilities and observed outcomes. The Brier Score accounts for both calibration and sharpness of predictions, providing a comprehensive evaluation of model performance[3]. Mathematically, the Brier Score (BS) is calculated as:

$$E_{\text{BS}} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{6}$$

where N is the total number of samples, p_i denotes the predicted probability for the i^{th} sample, and o_i represents the corresponding observed outcome (0 for incorrect prediction, 1 for correct prediction). Since the Brier Score is the difference between the prediction and observed outcome, it can be handled as an error that needs minimization (a perfectly calibrated model yielding a Brier Score of 0).

4. Experiments and Results

4.1. Data

Two databases of endoscopic images of colorectal polyps are employed throughout the experiments. An overview of the databases is available in Table 1. The *CRP* dataset has been curated in-house in collaboration with six hospitals within the Netherlands, which includes several imaging modalities, namely: White-Light Endoscopy (WLE), Blue Light Imaging (BLI), Linked Color Imaging (LCI), Narrow Band Imaging (NBI), as well as the i-Scan modality in Modes 1, 2, and 3. This dataset includes a total of 1,626 distinct polyps. Polyps are labeled according to histology outcomes as adenoma (Ad), sessile serrated lesions (SSL), and hyperplastic (Hp). The latter two polyp types are considered pre-malignant, and Hps are categorized as benign. In this study, experiments are carried out to classify CRPs into benign and pre-malignant classes. The data are partitioned into distinct training and validation sets, with an additional independent test set for evaluation purposes. The number of images in the training, validation, and test set is 2,777, 609, and 757, respectively.

The second employed dataset is the *POLAR* database [13], consisting of endoscopic images of polyps linked with histopathology collected from eight

Dataset	Number of polyps			
	Total	Hp	Ad	SSL
CRP	1,626	278	1,183	165
POLAR	1,339	230	973	136

Table 1. Brief description of the applied datasets.

Dutch hospitals. The images are non-magnified and of NBI modality. The dataset comes with pre-defined data splits. From this dataset, the same three categories of polyps are employed throughout the experiments, labeled into benign and pre-malignant characterization. With this data, the training set incorporates 2,115 images, validation 522 images, and testing 708 images.

4.2. Experimental setup

A ResNet50 network pretrained with ImageNet [9], serves as the baseline throughout the experiments. For the stochastic variational inference method, the implementation of Dehghani *et al.* [8] is adopted with the Flipout layers [22] implementation, using 10 forward passes during training and 30 forward passes for each inference cycle during testing. As mentioned earlier, another Bayesian approximation is also employed: the last-layer Laplace approximation (LLLA) [7]. During training, various data augmentation methods and weighted sampling are employed for the regularization of the model and to account for the class imbalances. For the temperature scaling (TS) approach, the temperature parameter T is optimized, using a holdout validation set to re-scale the logits of the trained deterministic baseline. For computational efficiency, images are resized to 256×256 pixels. For the training process, the Adam optimizer is employed with a learning-rate scheduler. Throughout the training, a mini-batch size of 16 images is applied, and the data are shuffled after each epoch to enhance the diversity. The experiments are conducted within the PyTorch framework and executed on an RTX 4090 GPU.

4.3. Evaluation metrics

As defined in Section 3.2, we use the ECE and BS errors to measure the miscalibration of the predictions. The uncertainties of the models are measured and quantified using calibration plots, known as Reliability Diagrams [20]. When doing so, we discretize the predicted probabilities of the models into several bins. The resulting plots (as in Figs. 1 and 2) demonstrate the frequency of correctly predicted labels for each bin of the discrete probability values. For performance evaluation, the Area under the curve (AUC) is used. Additionally, as a measure of the test-time efficiency of the models, the inference time per input sample is evaluated. Given the iterative nature of the SVI method, the inference time accumulates as the sum of the total number of forward passes required for each input sample.

4.4. Results of the calibration assessment

The Bayesian approximation methods are thoroughly evaluated along with the temperature scaling approach on a binary polyp categorization task, with the purpose of evaluating their applicability to clinical practice. To this end, the calibration and characterization performances of the models are evaluated, with careful consideration given to their generalization capabilities across different datasets.

A. Intra-dataset experiments: Table 2 shows the categorization and calibration performances of the models when trained and evaluated on the same dataset. The results reveal that the SVI method reduces the ECE by 48.85% and 25.91%, compared to the baseline in the *CRP* and *POLAR* datasets, respectively. Moreover, SVI reduces the BS error by 7.43% and 36.49% for the *CRP* and *POLAR* datasets, respectively. A comparison of the categorization performances demonstrates that the SVI method performs still below the baseline, and below the post-processing approaches that offer the same AUC as the baseline.

Evaluations of the reliability diagrams for the test logits of the models are available in Figs. 1 and 2. The bar plots show that the closest alignment to the perfect calibration belongs to the SVI method. The LLLA approach does not provide a significant improvement, although it reduces the over-confidence and under-confidence gaps. While scaling the predictions of the model using the TS method with optimized temperature parameters, the applied scaling factor may not have a clear intuitive interpretation in the context of the underlying model or dataset. As shown in Fig. 2, the method imposes extra under-confidence or over-confidence gaps.

B. Inter-dataset experiments: Table 2 also presents a performance comparison of the models when evaluated on samples from a dataset, that is distinct from the one used for training. Our analysis reveals a consistent rise in the ECE from the *CRP* to *POLAR* dataset across all calibration metrics. Similarly, when applying the SVI method to the *POLAR*-trained model and testing with the *CRP*, the distribution of weights formed during training diminishes the model's ability to generalize to the testing data.

Notably, the LLLA and TS methods demonstrate superior performance when trained on the POLAR dataset, exhibiting enhanced calibration and predictive accuracy. These findings highlight the significance of dataset compatibility and method selection in achieving optimal model performance across diverse clinical contexts.

C. Efficiency assessment: In a clinical environment where timely decision-making is critical, the computational loads of CAD models during inference can hinder their practical utility. In our investigation of method efficiency,

we provide insights into inference times per sample, as depicted in Fig. 3. The inference time for the SVI is directly proportional to the number of samples drawn from the Gaussian distribution. Consequently, as the number of forward passes increases, so does the inference time.

5. Conclusion and Discussion

Confidence calibration of CAD systems ensures that the predictive probabilities accurately reflect the true likelihoods of providing correct output, which is of utmost importance in a clinical setting. We investigate the relatively unexplored direction of calibration performance of polyp characterization systems in endoscopic imaging. In the field of confidence calibration, along with post-processing techniques such as temperature scaling, Bayesian approximation methods including variational inference and Laplace approximation have gained significant attention. Through multiple intra-dataset and inter-dataset experiments, we explore the robustness of the calibration performances of these methods and their applicability to CRP categorization models. The results demonstrate that the SVI method, while providing a superior calibration performance for intra-dataset experiments, falls behind in test-time efficiency and inter-dataset generalization. The Laplace approximation and the temperature scaling methods provide better calibration through cross-dataset examination. However, utilizing the LLLA does not provide a significant improvement to the calibration performance, compared to the baseline in an intra-dataset assessment. However, the TS method requires accurate optimization of the temperature, while still being prone to introducing extra under-confidence or over-confidence alignments. The importance of implementing reliable CAD models within the clinical setting is evident, but the efficiency and real-time applicability of such models cannot be ignored for their design. This work serves as a step forward in discovering more accurate methods that can seamlessly integrate into medical practice, offering robust AI prediction support for clinical workflows.

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 2

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018. 1

[3] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. 3

[4] Rafael Cardoso, Feng Guo, Thomas Heisser, Monika Hackl, Petra Ihle, Harlinde De Schutter, Nancy Van Damme, et al. Colorectal cancer incidence, mortality, and stage distribution in european countries in the colorectal cancer screening era: an international population-based study. *The Lancet Oncology*, 22(7):1002–1013, 2021. 1

[5] Gustavo Carneiro, Leonardo Zorron Cheng Tao Pu, Rajvinder Singh, and Alastair Burt. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical image analysis*, 62:101653, 2020. 2

[6] Douglas A Corley, Christopher D Jensen, Amy R Marks, Wei K Zhao, Jeffrey K Lee, Chyke A Doubeni, Ann G Zauber, et al. Adenoma detection rate and risk of colorectal cancer and death. *New england journal of medicine*, 370 (14):1298–1306, 2014. 1

[7] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Info. Processing Sys.*, 34:20089–20103, 2021. 3, 4

[8] Nikoo Dehghani, Thom Scheeve, Quirine EW van der Zander, Ayla Thijssen, Ramon-Michel Schreuder, Ad AM Masclee, Erik J Schoon, Fons van der Sommen, and Peter HN de With. Robust colorectal polyp characterization using a hybrid bayesian neural network. In *MICCAI Workshop on Cancer Prevention through Early Detection*, pages 108–117. Springer, 2022. 1, 4

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[10] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011. 2

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conf. on machine learning*, pages 1321–1330. PMLR, 2017. 1, 2, 3

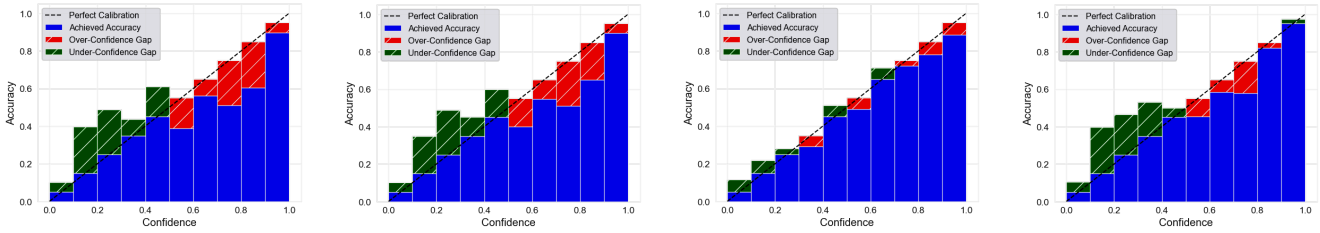
[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[13] Britt BSL Houwen, Yark Hazewinkel, Ioannis Giotis, Jasper LA Vleugels, Nahid S Mostafavi, Paul van Putten, Paul Fockens, Evelien Dekker, POLAR Study Group, et al. Computer-aided diagnosis for optical diagnosis of diminutive colorectal polyps including sessile serrated lesions: a real-time comparison with screening endoscopists. *Endoscopy*, 55(08):756–765, 2023. 3

[14] Mobarakol Islam, Lalithkumar Seenivasan, Hongliang Ren, and Ben Glocker. Class-distribution-aware calibration for long-tailed visual recognition. *arXiv preprint arXiv:2109.05263*, 2021. 2

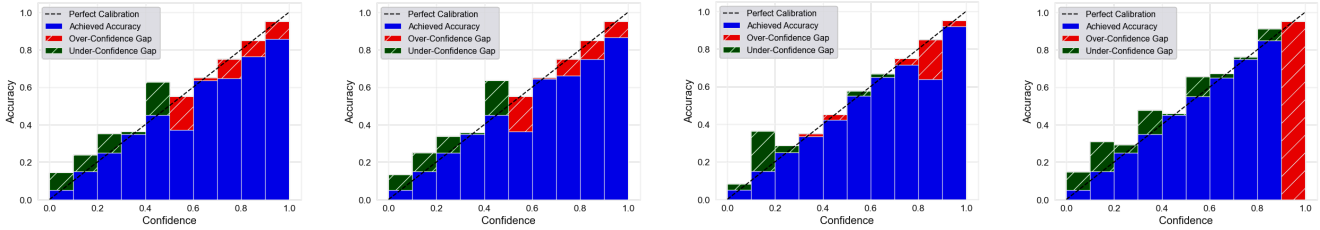
[15] Tim JM Jaspers, Tim GW Boers, Carolus HJ Kusters, Martijn R Jong, Jelmer B Jukema, Albert J de Groof, Jacques J Bergman, Peter HN de With, and Fons van der Sommen. Investigating the impact of image quality on endoscopic ai model performance. In *International Workshop on Applications of Medical AI*, pages 32–41. Springer, 2023. 1

[16] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020. 2



(a) Baseline (b) LLLA (c) SVI (d) TS

Figure 1. Reliability diagrams related to the CRP dataset from left to right: the baseline, LLLA, SVI, and TS methods.



(a) Baseline (b) LLLA (c) SVI (d) TS

Figure 2. Reliability diagrams related to the POLAR dataset from left to right: the baseline, LLLA, SVI, and TS methods.

	CRP (intra)			*POLAR* (intra)			*CRP* → *POLAR* (inter)			*POLAR* → *CRP* (inter)		
Method	AUC ↑	ECE ↓	BS ↓	AUC ↑	ECE ↓	BS ↓	AUC ↑	ECE ↓	BS ↓	AUC ↑	ECE ↓	BS ↓
Baseline	**0.87**	0.1092	0.1359	**0.78**	0.1043	0.1688	**0.80**	**0.0866**	**0.1246**	0.75	0.1186	0.2108
LLLA	0.87	0.1021	0.1341	0.78	0.0976	0.1673	0.75	0.1156	0.2092	**0.80**	**0.0791**	**0.1228**
SVI	0.83	**0.0558**	**0.1258**	0.76	**0.0773**	**0.1072**	0.72	0.1519	0.1773	0.78	0.1043	0.1688
TS	0.87	0.0922	0.1319	0.78	0.0833	0.1671	0.75	0.1009	0.2063	**0.80**	0.0922	0.1246

Table 2. Calibration and categorization performances evaluated for the employed methods in an intra-dataset task (left-half) and an inter-dataset task (right-half), where the arrow symbol points to the dataset used for testing.

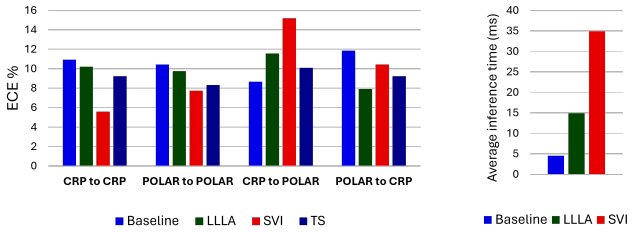


Figure 3. Measured ECE (left) and inference time (right) for different calibration approaches.

[17] Koen C Kusters, Thom Scheeve, Nikoo Dehghani, Quirine EW van der Zander, Ramon-Michel Schreuder, Ad AM Masclee, Erik J Schoon, Fons van der Sommen, et al. Colorectal polyp classification using confidence-calibrated convolutional neural networks. In *Medical Imaging 2022: Computer-Aided Diagnosis*, pages 456–468. SPIE, 2022. 2

[18] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 3

[19] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of IEEE conf. on computer vision and pattern recog.*, pages 427–436, 2015. 1

[20] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. 4

[21] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, et al. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180. PMLR, 2015. 2

[22] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, et al. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arX preprint arXiv:1803.04386*, 2018. 2, 4

[23] Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, page 100003, 2023. 1