Colorectal polyp classification using confidence-calibrated convolutional neural networks

Citation for published version (APA):
Kusters, K., Scheeve, T., Dehghani, N., van der Sommen, F., & de With, P. H. N. (Ed.) (2022). Colorectal polyp classification using confidence-calibrated convolutional neural networks. In K. Drukker, & K. M. Iftekharuddin (Eds.), *Medical Imaging 2022: Computer-Aided Diagnosis* (pp. 442-454). Article 120331P (Proceedings of SPIE; Vol. 12033). SPIE. https://doi.org/10.1117/12.2606801

DOI:
10.1117/12.2606801

Document status and date:
Published: 04/04/2022

Document Version:
Accepted manuscript including changes made at the peer-review stage

Colorectal Polyp classification using Confidence-Calibrated Convolutional Neural Networks

Koen C. Kusters[a], Thom Scheeve[a], Nikoo Dehghani[a], Quirine E.W. van der Zander[b,e], Ramon-Michel Schreuder[c], Ad A.M. Masclee[b], Erik J. Schoon[c,e], Fons van der Sommen[a,d], and Peter H.N. de With[a]

[a]Eindhoven University of Technology, Eindhoven, The Netherlands
[b]Maastricht University Medical Center+, Maastricht, The Netherlands
[c]Catharina Hospital, Eindhoven, The Netherlands
[d]Eindhoven Artificial Intelligence Systems Institute, Eindhoven, The Netherlands
[e]GROW, School for Oncology and Developmental Biology, Maastricht, The Netherlands

ABSTRACT

Computer-Aided Diagnosis (CADx) systems for in-vivo characterization of Colorectal Polyps (CRPs) which are precursor lesions of Colorectal Cancer (CRC), can assist clinicians with diagnosis and better informed decision-making during colonoscopy procedures. Current deep learning-based state-of-the-art solutions achieve a high classification performance, but lack measures to increase the reliability of such systems. In this paper, the reliability of a Convolutional Neural Network (CNN) for characterization of CRPs is specifically addressed by confidence calibration. Well-calibrated models produce classification-confidence scores that reflect the actual correctness likelihood of the model, thereby supporting reliable predictions by trustworthy and informative confidence scores. Two recently proposed trainable calibration methods are explored for CRP classification to calibrate the confidence of the proposed CNN. We show that the confidence-calibration error can be decreased by 33.86% (-0.01648 ± 0.01085), 48.33% (-0.04415 ± 0.01731), 50.57% (-0.11423 ± 0.00680), 61.68% (-0.01553 ± 0.00204) and 48.27% (-0.22074 ± 0.08652) for the Expected Calibration Error (ECE), Average Calibration Error (ACE), Maximum Calibration Error (MCE), Over-Confidence Error (OE) and Cumulative Calibration Error ($CUMU$), respectively. Moreover, the absolute difference between the average entropy and the expected entropy was considerably reduced by 32.00% (-0.04374 ± 0.01238) on average. Furthermore, even a slightly improved classification performance is observed, compared to the uncalibrated equivalent. The obtained results show that the proposed model for CRP classification with confidence calibration produces better calibrated predictions without sacrificing classification performance. This work shows promising points of engagement towards obtaining reliable and well-calibrated CADx systems for in-vivo polyp characterization, to assist clinicians during colonoscopy procedures.

Keywords: Colorectal polyps, Convolutional Neural Networks, Classification, Model Confidence Calibration

1. INTRODUCTION

Colorectal Cancer (CRC) is the third-leading cause of cancer-related deaths for both men and women in the United States, while CRC even ranks second when the cases for men and women are combined.[1] Especially in countries with medium to high Human Development Index (HDI), the incidence of CRC has increased the last couple of years.[2] As with most types of cancer, the survival rate of CRC is heavily correlated with the stage of diagnosis. Precursor lesions of CRC are Colorectal Polyps (CRPs), which can be subdivided into three different polyp types: adenomas (ADs), sessile serrated adenomas/polyps (SSA/Ps) and hyperplastic polyps (HPs). A further subdivision, based on CRC progression risk, can be done into two classes, pre-malignant and benign class. ADs and SSA/Ps belong to the first class since they are able to advance into CRC,[3] while HPs bear no risk of progression into cancer, thus belonging to the latter class. Early detection and removal of these lesions

Contact author: Koen C. Kusters (c.h.j.kusters@tue.nl)

is essential to prevent from further complications and development into CRC, thereby increasing the chance of survival for the patient.

Therefore, bowel cancer screening programs are implemented to identify and treat pre-malignant CRPs at an early stage during colonoscopy. However, current colonoscopy procedures are subject to several complications. Clinical colonoscopy procedures fully rely on pathological diagnosis, since studies show that even experienced and trained endoscopists are not sufficiently able to correctly distinguish endoscopically benign from (pre-)malignant CRPs.[4] Consequently, the current medical protocol dictates that all polyps encountered during colonoscopy procedures should be resected and undergo histopathological evaluation. This approach increases costs for histopathological analysis and induces possible complication risks for patients, by unnecessarily resecting CRPs that bear no risk of progressing into CRC. The above-mentioned challenges illustrate the growing need for in-vivo Computer-Aided Diagnosis (CADx) systems for characterization of polyps.

In order to support diagnosis and resection decision-making of endoscopists as well as adoption of 'resect-and-discard' and 'diagnose-and-leave' strategies[5,6] during clinical assessment, reliable CADx systems are required. Already many CAD systems for automated CRP classification have been developed. Well-established basic machine learning algorithms, such as Support Vector Machines (SVMs) and Random Forests, are employed by providing handcrafted features[7] or features extracted by Convolutional Neural Networks (CNNs).[8–12] More recently, CNN architectures are used as an end-to-end predictor rather than feature extractor, enabled by the increasing amount of publicly available datasets. However, the majority of these datasets are relatively small, since labeling of medical datasets is expensive. Therefore, transfer learning with CNN architectures pre-trained on large datasets as ImageNet[13] is a widely employed strategy.[14–20]

In high-risk medical applications, such as CRP characterization, classification results of such CADx systems should be well-calibrated and reliable to optimally assist clinicians in diagnosis and decision-making. At present, CADx systems are obtaining very promising results towards real-time application,[21] multimodal application[22] and showing feasibility of 'resect and discard' and 'diagnose and leave' strategies[23] by exceeding PIVI thresholds.[5] Unfortunately, metrics for increasing reliability and trustworthiness by model confidence calibration are insufficiently addressed. Calibration of classification confidence is highly desirable, since deep learning models tend to produce over-confident results,[24] thereby limiting reliability and correct interpretation of classification results. Carneiro *et al.*[25] investigated the roles of confidence calibration and classification uncertainty on the accuracy and calibration error for the five-class CRP classification problem. Confidence calibration was obtained by post-processing temperature scaling,[24] while a prediction uncertainty based on the entropy of the probability vector was used for an acceptation/rejection protocol for classification results.

In this work, we address confidence calibration by two recently proposed trainable methods.[26,27] In contrast to the work of Carneiro *et al.*, in which only a single approach was employed, the calibration techniques in our work do not require an explicit training procedure on the validation set for calibration, thereby offering a more convenient solution and facilitating the calibration process. One of the suggested calibration methods significantly reduces the calibration error, in terms of all five employed metrics, while even slightly increasing the classification performance. Consequently, this enables the proposed method to be reliable and well-calibrated, without sacrificing on polyp distinction capability.

The remainder of this paper is organized as follows. Section 2 gives an overview of the methodology. Section 3 elaborates on the experimental results, followed by an extensive discussion in Section 4. Lastly, the conclusions are presented in Section 5.

2. METHODS

2.1 Data

The data used in this study are prospectively collected at the Maastricht University Medical Center+ (MUMC+), Catharina Hospital Eindhoven (CZE), both in the Netherlands, and the Queen Alexandra Hospital (QA) in Portsmouth, United Kingdom. The dataset includes images acquired with White-Light Endoscopy (WLE) [*], Blue Light Imaging (BLI) [*] and Linked Color Imaging (LCI) [*] modalities, which are collected from CZE and QA.

[*] EG-760 Colonoscope (Fujifilm® Corporation, Tokyo, Japan)

Table 1. Number of polyps per colorectal polyp class.

	Polyp Class		
Hospital	Pre-Malignant (P)	Benign (N)	**Total**
MUMC+	259	30	289
CZE	307	77	384
QA	93	54	147
Total	659	161	820

Table 2. Number of available images per modality per colorectal polyp class

	Polyp Class		
Modality	Pre-Malignant (P)	Benign (N)	**Total**
i-Scan1	258	30	288
i-Scan2	207	27	234
i-Scan3	233	30	263
WLE	395	131	526
BLI	398	133	531
LCI	345	100	445
Total	1836	451	2287

Furthermore, data collection in MUMC+ is done with other acquisition equipment, in the sense that MUMC+ acquires images with i-Scan modality in Modes 1, 2 and 3 [†]. All available modalities ensure visual properties for enhanced visibility of polyps and polyp surface, which can be observed in the examples depicted in Fig. 1.

Included polyp types are HPs, ADs, SSA/Ps and adenocarcinoma. The latter three polyp types are (pre-) malignant and hence considered the positive (P) class in this study, while HPs are considered the negative (N) benign class. The total number of included polyps, subdivided per hospital, are presented in Table 1, while the number of images subdivided per imaging modality are listed in Table 2. From Tables 1 and 2, it can be observed that a maximum of one image per polyp per imaging modality occurs.

The test set used for performance evaluation on unseen data, is restricted to 86 distinct polyps, consisting of respectively 19 benign and 67 pre-malignant polyps. For each polyp, images in WLE, BLI and LCI modalities are contained in the test set, leading to a total number of 258 images. For model training, 80% of the remaining images are used, a total of 316 and 1,308 images in the benign and pre-malignant class, respectively. The remaining 20% is used for the validation set, a total of 78 and 327 images for the benign and pre-malignant class, respectively. In the splits made for testing and training/validation, it is ensured that all polyps from the same patient are contained in a single set to avoid data leakage.

2.2 Data Pre-processing

From raw endoscopic images, the central area is selected as the region of interest. This region ensures a coverage of polyp area as well as the surrounding tissue, such that the black border of the raw endoscopic image is removed. All RGB images are resized to 256 × 256 pixels, while compatibility of the dataset with ImageNet pre-trained networks is ensured by channel-wise subtracting the mean and dividing by the standard deviation of ImageNet data. In order to virtually increase the dataset size for improved generalization of the network, data augmentation techniques are employed. The images are augmented by a combination of horizontal and vertical flipping, rotation by $\theta \in \{0°, 90°, 180°, 270°\}$, Gaussian blurring, contrast/saturation/brightness enhancements, random affine and perspective transforms. To alleviate on the significant class imbalance, as observed in Tables 1 and 2, the described augmentation techniques are employed *per-class*, by ensuring a higher probability of execution for the inferior (N) class compared to the majority (P) class.

[†] EC38-i10F2 Colonoscope (PENTAX® Medical, Hoya Corp., Tokyo, Japan)

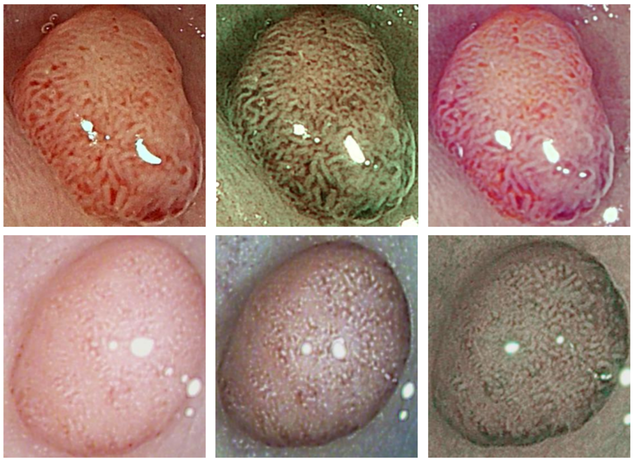

Figure 1. Overview of images in all available modalities (Top) From left to right: WLE, BLI and LCI. (Bottom) From left to right: i-Scan1, i-Scan2 and i-Scan3.

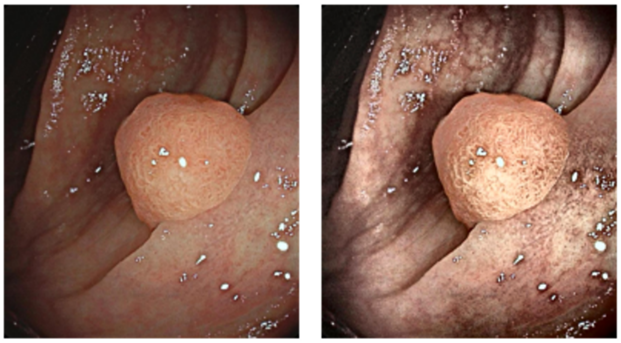

Figure 2. Images with and without CLAHE pre-processing. (Left) Original. (Right) CLAHE (CL=4.0, TS=(3,3)).

Since images originate from several hospitals and acquisition equipment, significant contrast differences are observed among images. To investigate the influence of contrast enhancement on the performance of the proposed CNN, in some experiments Contrast Limited Adaptive Histogram Equalization (CLAHE)[28] is employed prior to all previously stated pre-processing operations. CLAHE divides the image in small tiles, in which the contrast is enhanced by a pre-determined clip limit before equalizing the histogram. Several combinations of tile size (TS) and clip limit (CL) have been explored for our application. The most successful combination for this application is the combination of a value of 4.0 for the clip limit and a tile size of 3×3 pixels (CL=4.0, TS=(3,3)), which is used for conducting further experiments. A comparison of an original image with an image obtained after the CLAHE application is depicted in Fig. 2.

2.3 Network Architecture

In this work, the EfficientNet[29] architecture is serving as the backbone for the proposed classification algorithm. A customized version of EfficientNet-B4 (Eff-B4), has been used for the conducted experiments. In order to make the default network architecture suitable for our two-class CRP classification problem, the classification head with the fully-connected output layer consisting of 1,000 nodes is removed. Instead, a custom classification head is inserted, consisting of a flattened layer, followed by two 1,024-neuron fully-connected layers with ReLU activation, finalized with the 2-neuron fully-connected output layer with SoftMax activation.

The EfficientNet-B4 framework is either initialized with ImageNet pre-trained weights, or with weights obtained by ImageNet pre-training followed by secondary pre-training with an endoscopy-driven dataset. The dataset exploited for secondary endoscopy-driven pre-training is the GastroNet database, as described in the

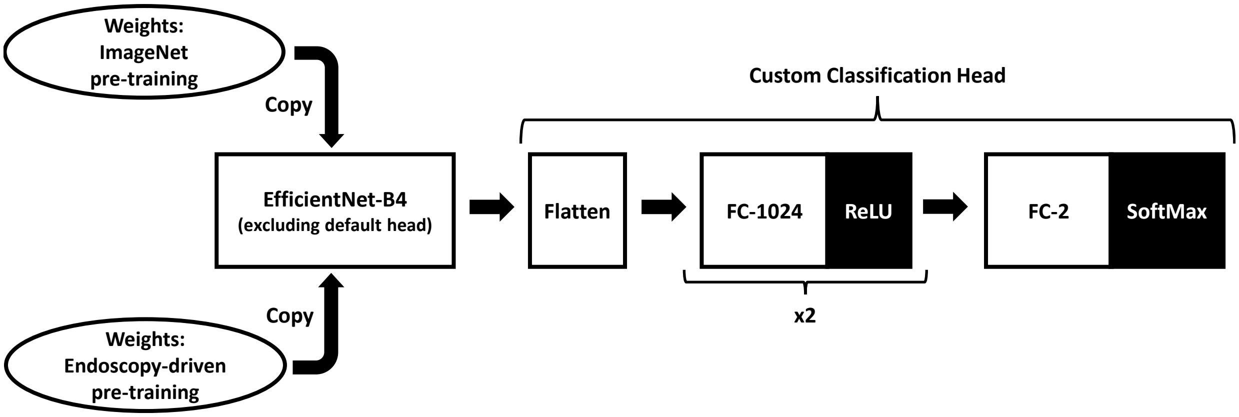

Figure 3. Overview of the employed network architecture used for the two-class CRP classification task.

papers of De Groof *et al.*[30] and Van der Putten *et al.*[31] This dataset consists of $\approx 500,000$ endoscopic images from several organs in the gastro-intestinal track. Pre-training with this set enables the model to gather improved discriminative power for endoscopic imagery. In the remainder of this paper, the dual pre-training variant is adopted, which is referred to as 'PT'. A schematic of the proposed network architecture is depicted in Fig. 3.

2.4 Training Procedure

The proposed network is trained for 75 epochs or until convergence on the validation set, using the Adam optimizer with a learning rate of 1×10^{-5} and $(\beta_1, \beta_2) = (0.9, 0.999)$. Due to the imbalance between the classes, an independent batch generator is constructed, which sub-samples the pre-malignant class, thereby ensuring that each of the two classes is representative during training. Each training iteration, a batch of 8 benign and 10 pre-malignant images is generated followed by a shuffling operation, until all benign images are seen by the network once, which triggers the end of a training epoch. The cross-entropy (CE) loss has been used to evaluate the loss of the model during the training procedure. The proposed methods are implemented in Python using the PyTorch framework and experiments were executed on a GeForce RTX 2080 Ti.

2.5 Confidence Calibration Methods

Model confidence calibration is the problem of matching the predicted confidence scores with the true correctness likelihood of the model. Deep learning models are typically poorly calibrated, because models tend to produce over-confident results.[24] An example of poorly calibrated results is when a model is 80% accurate for a set of predictions, while being 0.99 confident for each of the predictions. Ideally, a model should be 0.80 confident for each of the predictions when being 80% accurate. Over-confidence limits the reliability and correct interpretation of predictions, since there is no clear indication whether the model is likely to be incorrect (i.e. predicting with low confidence when likely to be incorrect). Therefore, well-calibrated models are desired for a high-risk application as CRP classification, to provide clinicians with trustworthy and informative confidence scores.

Model calibration is visualized with reliability diagrams.[32] A reliability diagram summarizes the accuracy versus the predicted confidence of samples, by grouping predictions into bins, based on their predicted confidence. Subsequently, for each bin the accuracy is plotted as a function of the average predicted confidence. The level of miscalibration can be assessed by the gap between the plotted accuracy and the ideal diagonal. The better the bin accuracy aligns with the ideal diagonal, the higher the calibration performance of the model. Several examples of reliability diagrams are depicted in Fig. 4.

In order to calibrate the output of the proposed network, two trainable techniques are considered. Firstly, the auxiliary loss term proposed by Liang *et al.*[26] is added to the default cross-entropy loss during the training procedure. This auxiliary loss term comprises the difference between the average predicted confidence and

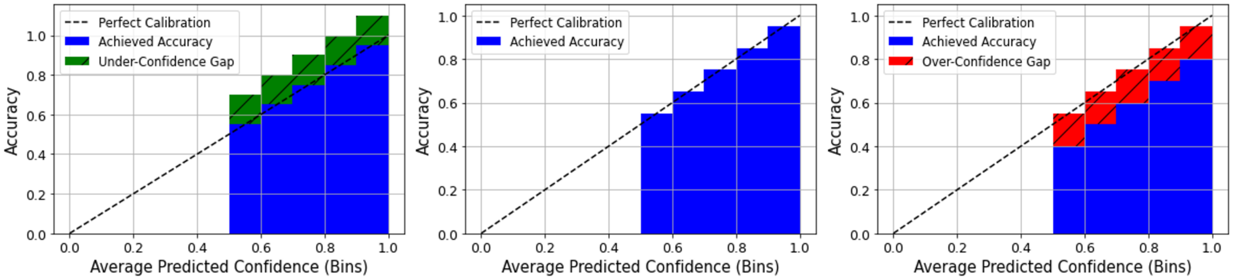

Figure 4. Three examples of reliability diagrams. (Left) Reliability diagram for a model with under-confident predictions, where the accuracy exceeds the average predicted confidence. (Middle) Reliability diagram for a model that is perfectly calibrated, where the accuracy ideally aligns with the average predicted confidence. (Right) Reliability diagram for a model with over-confident predictions, where the accuracy falls short to the average predicted confidence.

accuracy (called DCA), thereby enabling the minimization of the calibration error for each mini-batch directly in the loss. The employed loss ($\mathcal{L}_{\mathrm{DCA}}$) is computed by:

$$\begin{aligned}\mathcal{L}_{\mathrm{DCA}} &= \mathrm{CE} + \beta \cdot \mathrm{DCA} \\ &= \mathrm{CE} + \beta \cdot \left| \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i) - \frac{1}{N}\sum_{i=1}^{N} \hat{p}_i \right|,\end{aligned} \tag{1}$$

where $\hat{y}_i$ and y_i are the predicted and ground-truth class for sample i, respectively. Furthermore, $\hat{p}_i$ is the confidence of predicted label for sample i, parameter N is the number of samples in the mini-batch and β is a weight scalar. The function $\mathbf{1}(\cdot)$ denotes the unity function if the embedded expression holds. In the conducted experiments, an empirically determined value of $\beta = 15$ is used.

In addition to the method proposed Liang *et al.*, the Dynamically Weighted Balanced (DWB) Loss proposed by Fernando *et al.*[27] is employed, instead of the default CE loss function. This DWB loss function is used for comparison and is composed of two terms, a dynamically weighted CE and a regularization component equal to the entropy of the Brier Score. The latter term can be considered as a reliability component that leads to better calibration. The loss ($\mathcal{L}_{\mathrm{DWB}}$) is computed by:

$$\mathcal{L}_{\mathrm{DWB}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c} w_j^{(1-p_{ij})} y_{ij} \log(p_{ij}) - p_{ij}(1-p_{ij}), \tag{2}$$

where w_j is the class weight of Class j, while y_{ij} and p_{ij} are the jth element of the one-hot encoded label and predicted probability for sample i, respectively. Furthermore, N is the number of samples in the mini-batch. The class weights are computed as follows:

$$w_j = \log\left(\frac{\max(n_j | j \in c)}{n_j}\right) + 1, \tag{3}$$

where n_j is the frequency of samples from Class j.

2.6 Evaluation Metrics

2.6.1 Classification

Six metrics are employed for classification performance evaluation. Using the confusion matrix for the two-class classification task illustrated by Table 3, the first five metrics can be defined. Accuracy (Acc) is defined as the total fraction of correct predictions:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \cdot 100\%. \tag{4}$$

Table 3. Confusion matrix for the two-class classification problem.

Predicted Class	**Ground Truth Class** Pre-Malignant (P)	 Benign (N)
Pre-Malignant (P)	TP	FP
Benign (N)	FN	TN

Sensitivity ($Sens$) is the fraction of positive (P) samples that are correctly predicted:

$$Sens = \frac{TP}{TP + FN} \cdot 100\%, \tag{5}$$

while Specificity ($Spec$) is the fraction of negative (N) samples that are correctly predicted:

$$Spec = \frac{TN}{TN + FP} \cdot 100\%. \tag{6}$$

Negative Predictive Value (NPV) is the fraction of correct negative (N) predictions from the total number of negative predictions:

$$NPV = \frac{TN}{TN + FN} \cdot 100\%. \tag{7}$$

Positive Predictive Value (PPV) is the fraction of correct positive (P) predictions from the total number of positive predictions:

$$PPV = \frac{TP}{TP + FP} \cdot 100\%. \tag{8}$$

A receiver operating characteristic curve (ROC) summarizes the relationship between sensitivity and specificity. The area under the curve (AUC) for the ROC concludes the employed classification performance metrics.

2.6.2 Calibration

Five error metrics are used for calibration performance evaluation, all calculated from reliability diagrams.[24] The reliability diagram is a visual tool that summarizes the accuracy versus the average predicted confidence of samples. A reliability diagram is obtained by grouping predictions into M bins of size $1/M$, where B_m is the set of indices whose predicted confidence falls into the interval $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$. The accuracy A for each bin is calculated by:

$$A(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i), \tag{9}$$

where $\hat{y}_i$ and y_i are the predicted and ground-truth class for sample i, respectively. The average predicted confidence C of each bin is calculated by:

$$C(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{10}$$

where $\hat{p}_i$ is the confidence of predicted label for sample i. Using Eqns. (9) and (10), the employed calibration error metrics can be defined. Firstly, the Expected Calibration Error (ECE) is the weighted average of the calibration error across all bins, which is computed as:

$$ECE = \sum_{m=1}^{M} \frac{B_m}{n} \, |A(B_m) - C(B_m)|, \tag{11}$$

where n is the total amount of samples in the test set. Furthermore, the Average Calibration Error (ACE) giving the average error across the bins, is defined as:

$$ACE = \frac{1}{M^+} \sum_{m=1}^{M} \, |A(B_m) - C(B_m)|, \tag{12}$$

where M^+ is the amount of non-empty bins. Moreover, the Maximum Calibration Error (MCE) determines the largest error across the bins and is defined by:

$$MCE = \max_{m \in 1,\ldots,M} |A(B_m) - C(B_m)|. \tag{13}$$

Additionally, the Over-Confidence Error (OE) is the weighted average of the errors across bins where confidence exceeds accuracy, which is computed by:

$$OE = \sum_{m=1}^{M} \frac{B_m}{n} \left[C(B_m) \cdot \max\left(C(B_m) - A(B_m), 0 \right) \right]. \tag{14}$$

Lastly, a reliability diagram can be mapped into a single value, by adding up the calibration errors for all bins. The computation is the same as for ACE, without taking the average by dividing by the number of non-empty bins, which is referred to as the cumulative calibration error ($CUMU$).

2.6.3 Prediction Uncertainty

The uncertainty attached to a prediction can be expressed by the entropy of the predicted probability vector. The average entropy ($\bar{H}(\mathcal{X})$) for test set $\mathcal{X}$, based on the predicted probability vectors, is calculated as:

$$\bar{H}(\mathcal{X}) = -\frac{1}{N} \sum_{x \in \mathcal{X}} \sum_{c \in y} \hat{p}(y = c|x) \log\left(\hat{p}(y = c|x) \right), \tag{15}$$

where $\hat{p}(y = c|x)$ is the predicted probability for Class c given Sample x and N denotes the total number of samples in test set $\mathcal{X}$. Some reference values for single samples are: $H([0.05, 0.95]) \approx 0.199$, $H([0.1, 0.9]) \approx 0.325$ and $H([0.15, 0.85]) \approx 0.423$. As can be observed from the reference values, the more confident the predictions, the lower the resulting entropy. A low entropy value is only desired when the accuracy of the model matches the average predicted confidence for the samples, otherwise the model is considered over-confident and inherently being poorly calibrated. To separate this issue from the entropy calculation, we adopt an alternative calculation.

We propose to use $\Delta\bar{H}(\mathcal{X})$ as a metric to measure the calibration performance of the model, by capturing the absolute difference between achieved average entropy and the expected entropy. The expected entropy is calculated with Eqn. (15), by assuming perfect calibration. This implies that the average predicted confidence matches the achieved accuracy, e.g. with an accuracy of 85% the average predicted confidence for the predicted class should be 0.85. As such, the average predicted vector should be $[0.15, 0.85]$ or $[0.85, 0.15]$ for pre-malignant and benign samples, respectively. Consequently, an expected entropy of ≈ 0.423 is obtained when using Eqn. (15). The closer $\Delta\bar{H}(\mathcal{X})$ is to zero, the better the average predicted confidence reflects the actual correctness likelihood, inherently pointing to a better calibration.

3. EXPERIMENTAL RESULTS

The classification and calibration results for models with and without a calibration method are presented in Tables 4 and 5, respectively, while the resulting reliability diagrams are depicted in Fig. 5.

DCA Auxiliary Loss: Models trained using DCA auxiliary loss with a value of $\beta = 15$, gain on average 0.26% and 1.49% for accuracy and sensitivity, respectively, while losing 4.09% for specificity. The calibration metric results show that application of DCA auxiliary loss significantly reduces the error metrics ECE, ACE, MCE, OE and $CUMU$ on average by 33.86% (-0.01648 ± 0.01085), 48.33% (-0.04415 ± 0.01731), 50.57% (-0.11423 ± 0.00680), 61.68% (-0.01553 ± 0.00204) and 48.27% (-0.22074 ± 0.08652), respectively. Furthermore, $\bar{H}(\mathcal{X})$ is slightly increased on average by 14.23% (0.03914 ± 0.00366), while an average reduction of 32.00% (-0.04374 ± 0.01238) for $\Delta\bar{H}(\mathcal{X})$ is obtained. All above-mentioned results follow from the comparison with their uncalibrated equivalents. Moreover, the reliability diagrams from models trained with DCA auxiliary loss (middle column) show that overall the achieved accuracies of the bins are closer to the ideal diagonal, compared to the diagrams of the uncalibrated models (left column).

Table 4. Classification results for the proposed network with and without DCA auxiliary loss ($\beta = 15$) or DWB loss.

Model	**CLAHE**	**PT**	**Calib.**	Acc	$Sens$	$Spec$	NPV	PPV	AUC
Eff-B4	×	×	×	85.27%	86.07%	82.46%	62.67%	94.54%	0.91
Eff-B4	✓	×	×	86.05%	87.06%	82.46%	64.38%	94.59%	0.91
Eff-B4	✓	✓	×	85.27%	86.57%	80.70%	63.01%	94.05%	0.91
Eff-B4	×	×	DCA	86.05%	88.06%	78.95%	65.22%	93.65%	0.91
Eff-B4	✓	×	DCA	85.66%	87.56%	78.95%	64.29%	93.62%	0.91
Eff-B4	✓	✓	DCA	85.66%	88.56%	75.44%	65.15%	92.71%	0.91
Eff-B4	×	×	DWB	82.17%	84.08%	75.44%	57.33%	92.35%	0.89
Eff-B4	✓	×	DWB	82.95%	84.58%	77.19%	58.67%	92.90%	0.90
Eff-B4	✓	✓	DWB	84.88%	85.57%	82.47%	61.84%	94.51%	0.91

Table 5. Calibration results for the proposed network with and without DCA auxiliary loss ($\beta = 15$) or DWB loss.

Model	**CLAHE**	**PT**	**Calib.**	ECE	ACE	MCE	OE	$CUMU$	$\bar{H}(\mathcal{X})$	$\Delta\bar{H}(\mathcal{X})$
Eff-B4	×	×	×	0.02848	0.05866	0.22590	0.01926	0.29330	0.28367	0.13433
Eff-B4	✓	×	×	0.06098	0.12100	0.27534	0.02677	0.60499	0.29675	0.10730
Eff-B4	✓	✓	×	0.03976	0.08531	0.18740	0.03089	0.42654	0.25090	0.16710
Eff-B4	×	×	DCA	0.02479	0.03623	0.10737	0.00634	0.18113	0.32696	0.07709
Eff-B4	✓	×	DCA	0.03075	0.05622	0.15581	0.00888	0.28110	0.33113	0.07996
Eff-B4	✓	✓	DCA	0.02423	0.04007	0.08276	0.01510	0.20037	0.29064	0.12045
Eff-B4	×	×	DWB	0.10629	0.18139	0.35330	0.09373	0.90695	0.17555	0.29326
Eff-B4	✓	×	DWB	0.10909	0.16173	0.29044	0.09979	0.80867	0.15149	0.30519
Eff-B4	✓	✓	DWB	0.09022	0.13655	0.21728	0.07781	0.68276	0.16765	0.25713

DWB Loss: Models trained with DWB loss lose on average 2.20%, 1.82% and 3.51% for accuracy, sensitivity and specificity, respectively. The calibration metric results show that the application of DWB loss significantly increases error metrics ECE, ACE, MCE, OE and $CUMU$ on average by 159.67% (0.05879 ± 0.01348), 100.98% (0.07157 ± 0.03643), 25.94% (0.05746 ± 0.04982), 270.44% (0.06480 ± 0.01266) and 100.98% (0.11524 ± 0.47507), respectively. Furthermore, $\bar{H}(\mathcal{X})$ is significantly decreased on average by 40.08% (-0.11221 ± 0.02548), while an average increase of 118.87% (0.14895 ± 0.04459) for $\Delta\bar{H}(\mathcal{X})$ is obtained. All above-mentioned results follow from the comparison with their uncalibrated equivalents. Moreover, the reliability diagrams from models trained with DWB loss (right column) show that overall the achieved accuracies of the bins have larger discrepancies from the ideal diagonal, compared to the diagrams of the uncalibrated models (left column).

4. DISCUSSION

The DCA auxiliary loss is fruitful in decreasing the calibration error without or slightly hurting the classification performance. The improved calibration performance is demonstrated by the considerably reduced values for all five employed calibration-error metrics, the decrease in the absolute difference between the achieved and the expected average entropy ($\Delta\bar{H}(\mathcal{X})$), as well as the diminished distance between the obtained accuracy of the bins and the ideal diagonal in the reliability diagrams. The increase of average prediction uncertainty ($\bar{H}(\mathcal{X})$), in terms of the average entropy of predictions, is a logical consequence of proper confidence calibration, since over-confidence is corrected and illustrated by the amplitude reduction (red bars) of over-confidence gaps in the reliability diagrams. The obtained results show that the trainable DCA auxiliary loss term, proposed by Liang *et al.*,[26] minimizes the calibration error in the loss directly and is extremely successful for performance and facilitation of calibration.

The DWB loss function is unsuccessful in decreasing the calibration error. On the contrary, the error is significantly increased while also a decreased classification performance is observed. The augmented miscalibration

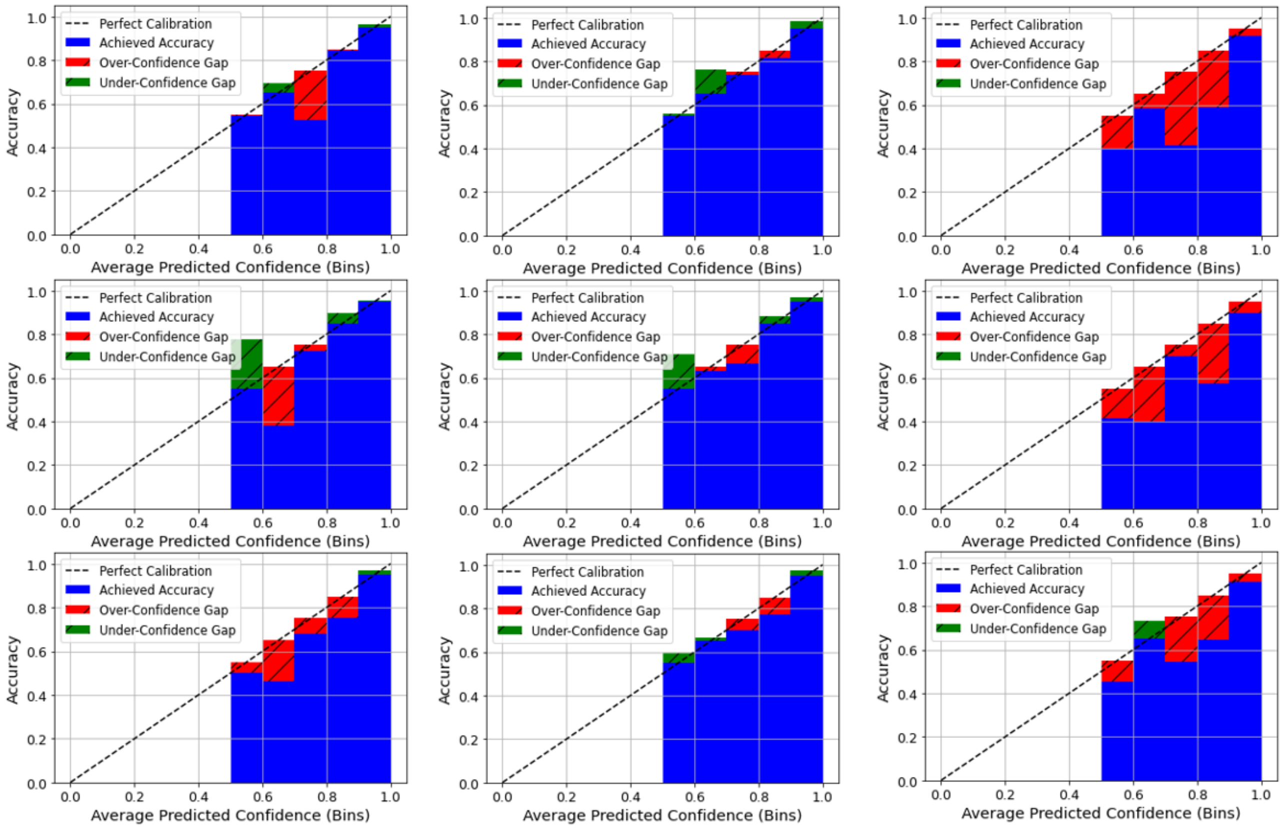

Figure 5. Overview of reliability diagrams. (Top row) Eff-B4 without CLAHE application and without PT. (Middle row) Eff-B4 with CLAHE application and without PT. (Bottom row) Eff-B4 with CLAHE application and with PT.
From left to right: No Calibration, DCA and DWB.

is demonstrated by the significantly increased calibration-error metrics, the increase in the absolute difference between the achieved and the expected entropy ($\Delta\bar{H}(\mathcal{X})$), as well as the enlarged distance between the obtained accuracy of the bins and the ideal diagonal in the reliability diagrams. The significant decrease of average prediction uncertainty ($\bar{H}(\mathcal{X})$), in terms of the average entropy of predictions, is a consequence of the predictions being more over-confident, illustrated by the extremely deteriorated OE metric and the larger over-confident gaps in the reliability diagrams.

An explanation for the unsatisfactory results may be that the DWB loss is mainly focused on handling class-imbalances, while to a lesser extent confidence calibration. The results in the work of Fernando *et al.*[27] show that the DWB loss is able to achieve improved calibration performance for multi-class (> 2) highly imbalanced datasets, compared to other loss functions such as Weighted Cross-Entropy and Focal Loss. However, the calibration performance of the proposed method was not compared to methods purely designed for model-confidence calibration. Furthermore, the dataset in our work is not multi-class, while also not being highly imbalanced due to sub-sampling measures on the majority class during model training. Altogether, these factors may cause the disappointing calibration performance of the DWB loss in our case.

5. CONCLUSIONS

CADx systems for in-vivo characterization of colorectal polyps are clinically valuable to avoid unnecessary removal of polyps and to enable adoption of new medical protocols, thereby reducing complication risks for patients and costs attached to histopathological analysis. Current research towards deep learning-based CADx systems lack methods to improve reliability in order to provide optimal assistance for clinicians during clinical procedures.

In this paper, reliability improvement of a Convolutional Neural Network for the two-class colorectal polyp classification is addressed by confidence calibration. Confidence calibration is desirable because deep learning models tend to produce over-confident results, thereby limiting reliability and correct interpretation of classification results, which can be misleading for clinicians when being assisted by such an algorithm. Well-calibrated models produce classification confidence scores that reflect the actual correctness likelihood of the model, thereby providing reliable predictions by trustworthy and informative confidence scores. In our work, two recently proposed trainable calibration methods are employed for the purpose of model confidence calibration. More specifically, we have compared DCA Auxiliary Loss and the DWB Loss, adopted for generic medical imaging and tuned in this paper for colorectal polyp classification. These techniques are considered to give valuable guidance to the clinicians for polyp classification. The method by Liang *et al.*[26] significantly reduces the calibration error of the proposed CNN, in terms of all five employed metrics (ECE, ACE, MCE, $CUMU$ and OE) while even slightly increasing the classification performance.

A limitation of this study is the relatively small size of the test set used for performance evaluation on unseen data, with only a classification improvement of 0.5% and 1.75% to each sample for sensitivity and specificity, respectively. However, only limited conclusions can be drawn about generalization on unseen data. Therefore, for future research a larger and more balanced dataset is desired, which can significantly increase the already feasible results. In summary, this work shows that the proposed confidence-calibration method can provide support to clinicians with in-vivo characterization of CRPs, by trustworthy and informative confidence scores. Furthermore, promising points of engagement are shown towards further research into reliable, well-calibrated deep learning-based CADx systems for CRP characterization.

ACKNOWLEDGMENTS

We thankfully acknowledge the Dutch Cancer Society for funding the COMET-OPTICAL project (Project No. 12639), which guided the presented study. The Dutch Cancer Society did not contribute to the study protocol, data analysis or manuscript writing.

REFERENCES

[1] Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., and Jemal, A., "Colorectal cancer statistics, 2020," *CA: A Cancer Journal for Clinicians* **70**(3), 145–164 (2020).

[2] Wong, M. C., Huang, J., Lok, V., Wang, J., Fung, F., Ding, H., and Zheng, Z.-J., "Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location," *Clinical Gastroenterology and Hepatology* (2020).

[3] Jass, J., "Classification of colorectal cancer based on correlation of clinical, morphological and molecular features," *Histopathology* **50**, 113–30 (02 2007).

[4] Vleugels, J. L., Dijkgraaf, M. G., Hazewinkel, Y., Wanders, L. K., Fockens, P., and Dekker, E., "Effects of training and feedback on accuracy of predicting rectosigmoid neoplastic lesions and selection of surveillance intervals by endoscopists performing optical diagnosis of diminutive polyps," *Gastroenterology* **154**(6), 1682 – 1693.e1 (2018).

[5] Rex, D., Kahi, C., O'Brien, M., Levin, T., Pohl, H., Rastogi, A., Burgart, L., Imperiale, T., Ladabaum, U., Cohen, J., and Lieberman, D., "The american society for gastrointestinal endoscopy pivi (preservation and incorporation of valuable endoscopic innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps," *Gastrointestinal endoscopy* **73**, 419–22 (03 2011).

[6] Abu Dayyeh, B. K., Thosani, N., Konda, V., Wallace, M. B., Rex, D. K., Chauhan, S. S., Hwang, J. H., Komanduri, S., Manfredi, M., Maple, J. T., Murad, F. M., Siddiqui, U. D., and Banerjee, S., "Asge technology committee systematic review and meta-analysis assessing the asge pivi thresholds for adopting real-time endoscopic assessment of the histology of diminutive colorectal polyps," *Gastrointestinal Endoscopy* **81**(3), 502.e1–502.e16 (2015).

[7] Scheeve, T., Schreuder, R.-M., Van der Sommen, F., Ijspeert, J., Dekker, E., Schoon, E., and With, P., "Computer-aided classification of colorectal polyps using blue-light and linked-color imaging," 37 (03 2019).

[8] Ribeiro, E., Uhl, A., Wimmer, G., and Häfner, M., "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and Mathematical Methods in Medicine* **2016** (2016).

[9] Zhang, R., Zheng, Y., Mak, T. W. C., Yu, R., Wong, S. H., Lau, J. Y. W., and Poon, C. C. Y., "Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain," *IEEE Journal of Biomedical and Health Informatics* **21**(1), 41–47 (2017).

[10] Fonollá, R., v. d. Sommen, F., Schreuder, R. M., Schoon, E. J., and de With, P. H. N., "Multi-modal classification of polyp malignancy using cnn features with balanced class augmentation," in [*2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*], 74–78 (2019).

[11] Usami, H., Iwahori, Y., Adachi, Y., Bhuyan, M., Wang, A., Inoue, S., Ebi, M., Ogasawara, N., and Kasugai, K., "Colorectal polyp classification based on latent sharing features domain from multiple endoscopy images," *Procedia Computer Science* **176**, 2507 – 2514 (2020). Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.

[12] Fonollà, R., Smyl, M., van der Sommen, F., Schreuder, R. M., Schoon, E. J., and de With, P. H. N., "Triplet network for classification of benign and pre-malignant polyps," in [*Medical Imaging 2021: Computer-Aided Diagnosis*], Mazurowski, M. A. and Drukker, K., eds., **11597**, 648 – 654, International Society for Optics and Photonics, SPIE (2021).

[13] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "ImageNet: A Large-Scale Hierarchical Image Database," in [*CVPR09*], (2009).

[14] Bour, A., Castillo-Olea, C., Garcia-Zapirain, B., and Zahia, S., "Automatic colon polyp classification using convolutional neural network: A case study at basque country," in [*2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*], 1–5 (2019).

[15] Sierra, F., Gutiérrez, Y., and Martínez, F., "An online deep convolutional polyp lesion prediction over narrow band imaging (nbi)," in [*2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*], 2412–2415 (2020).

[16] Tanwar, S., Goel, P., Johri, P., and Diván, M., "Classification of benign and malignant colorectal polyps using pit pattern classification," *SSRN Electronic Journal* (01 2020).

[17] Patino-Barrientos, S., Sierra-Sosa, D., Zapirain, B., Castillo, C., and Elmaghraby, A., "Kudo's classification for colon polyps assessment using a deep learning approach," *Applied Sciences* **10**, 501 (01 2020).

[18] Patel, K., Li, K., Tao, K., Wang, Q., Bansal, A., Rastogi, A., and Wang, G., "A comparative study on polyp classification using convolutional neural networks," (07 2020).

[19] Chen, P.-J., Lin, M.-C., Lai, M.-J., Lin, J.-C., Lu, H. H.-S., and Tseng, V. S., "Accurate classification of diminutive colorectal polyps using computer-aided analysis," *Gastroenterology* **154**(3), 568 – 575 (2018).

[20] Song, E., Park, B., Ha, C.-A., Hwang, S. W., Park, s. h., Yang, D.-H., Ye, B., Myung, S.-J., Yang, S.-K., Kim, N., and Leong, R., "Endoscopic diagnosis and treatment planning for colorectal polyps using a deep-learning model," *Scientific Reports* **10**, 30 (01 2020).

[21] Byrne, M. F., Chapados, N., Soudan, F., Oertel, C., Linares Pérez, M., Kelly, R., Iqbal, N., Chandelier, F., and Rex, D. K., "Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model," *Gut* **68**(1), 94–100 (2019).

[22] Fonolla Navarro, R., Zander, Q., Schreuder, R.-M., Masclee, A., Schoon, E., Van der Sommen, F., and With, P., "A cnn cadx system for multimodal classification of colorectal polyps combining wl, bli, and lci modalities," *Applied Sciences* **10**, 5040 (07 2020).

[23] Zachariah, R., Samarasena, J., Luba, D., Duh, E., Dao, T., Ninh, A., and Karnes, W., "Prediction of polyp pathology using convolutional neural networks achieves "resect and discard" thresholds," *The American Journal of Gastroenterology* **115**, 1 (10 2019).

[24] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., "On calibration of modern neural networks," *CoRR* **abs/1706.04599** (2017).

[25] Carneiro, G., Zorron Cheng Tao Pu, L., Singh, R., and Burt, A., "Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy," *Medical Image Analysis* **62**, 101653 (2020).

[26] Liang, G., Zhang, Y., Wang, X., and Jacobs, N., "Improved trainable calibration method for neural networks on medical imaging classification," (2020).

[27] Fernando, K. R. M. and Tsokos, C. P., "Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems* , 1–12 (2021).

[28] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K., "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing* **39**(3), 355–368 (1987).

[29] Tan, M. and Le, Q. V., "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR* **abs/1905.11946** (2019).

[30] de Groof, A. J., Struyvenberg, M. R., van der Putten, J., van der Sommen, F., Fockens, K. N., Curvers, W. L., Zinger, S., Pouw, R. E., Coron, E., Baldaque-Silva, F., Pech, O., Weusten, B., Meining, A., Neuhaus, H., Bisschops, R., Dent, J., Schoon, E. J., de With, P. H., and Bergman, J. J., "Deep-learning system detects neoplasia in patients with barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking," *Gastroenterology* **158**(4), 915–929.e4 (2020).

[31] van der Putten, J., de Groof, J., van der Sommen, F., Struyvenberg, M., Zinger, S., Curvers, W., Schoon, E., Bergman, J., and de With, P. H. N., "Pseudo-labeled bootstrapping and multi-stage transfer learning for the classification and localization of dysplasia in barrett's esophagus," in [*Machine Learning in Medical Imaging*], Suk, H.-I., Liu, M., Yan, P., and Lian, C., eds., 169–177, Springer International Publishing, Cham (2019).

[32] Niculescu-Mizil, A. and Caruana, R., "Predicting good probabilities with supervised learning," *Proceedings of the 22nd international conference on Machine learning* (2005).