Medical Image Analysis (2020)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Deep Learning Uncertainty and Confidence Calibration for the Five-class Polyp Classification from Colonoscopy

Gustavo Carneiro[a,*], Leonardo Zorron Cheng Tao Pu[b], Rajvinder Singh[b], Alastair Burt[b]

[a]*Australian Institute for Machine Learning, School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia*
[b]*Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, SA 5005, Australia*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

Polyp classification, deep learning, model calibration, classification uncertainty, Bayesian learning, Bayesian inference

ABSTRACT

There are two challenges associated with the interpretability of deep learning models in medical image analysis applications that need to be addressed: confidence calibration and classification uncertainty. Confidence calibration associates the classification probability with the likelihood that it is actually correct – hence, a sample that is classified with confidence $X\%$ has a chance of $X\%$ of being correctly classified. Classification uncertainty estimates the noise present in the classification process, where such noise estimate can be used to assess the reliability of a particular classification result. Both confidence calibration and classification uncertainty are considered to be helpful in the interpretation of a classification result produced by a deep learning model, but it is unclear how much they affect classification accuracy and calibration, and how they interact. In this paper, we study the roles of confidence calibration (via post-process temperature scaling) and classification uncertainty (computed either from classification entropy or the predicted variance produced by Bayesian methods) in deep learning models. Results suggest that calibration and uncertainty improve classification interpretation and accuracy. This motivates us to propose a new Bayesian deep learning method that relies both on calibration and uncertainty to improve classification accuracy and model interpretability. Experiments are conducted on a recently proposed five-class polyp classification problem, using a data set containing 940 high-quality images of colorectal polyps, and results indicate that our proposed method holds the state-of-the-art results in terms of confidence calibration and classification accuracy.

1. Introduction

In computer-aided diagnosis (CADx) systems, it is important that models not only produce accurate classification results, but they also display well calibrated classification confidence and reliable uncertainty estimates. A well calibrated classification confidence indicates the actual likelihood that the classification is correct, while a solid uncertainty estimate suggests the reliability of a

⋆G.C. acknowledges the support by the Australian Research Council through grant DP180103232 and the Alexander von Humboldt-Stiftung for the renewed research stay sponsorship. All authors acknowledge the grant 2018/7063 - NALHN/THRF - Project Grant - 0006005804 and the TitanXp donated by NVidia.
*Corresponding author: Tel.: +0-000-000-0000; fax: +0-000-000-0000;
e-mail: gustavo.carneiro@adelaide.edu.au (Gustavo Carneiro)

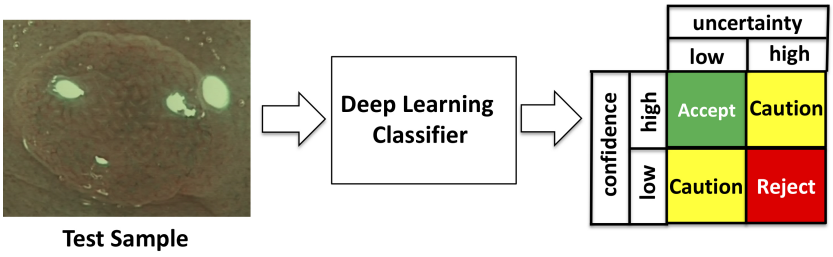

Fig. 1: Interpreting and improving the accuracy of a deep learning classifier using classification uncertainty and confidence. Test samples classified with high confidence and low uncertainty are accepted (green region), low confidence and high uncertainty are rejected (red region), and low confidence low uncertainty or high confidence high uncertainty are classified with caution (yellow regions).

particular classification result. In other words, confidence calibration and classification uncertainty will increase the interpretability of a model decision. For instance, Fig. 1 shows four possible interpretations based on the classification confidence and uncertainty results, which can allow the automatic classification of a test sample to be accepted (green region), rejected (red region) or classified with caution (yellow regions). The development of such model interpretability techniques is becoming important not only for academia (Lipton, 2016), but also for the whole society (Goodman and Flaxman, 2016). An important example of the application of this model interpretability technique would be when the clinician overturns a CADx decision based on its confidence and uncertainty estimates (Jiang et al., 2011). Despite the relevance of this research topic, current literature in medical image analysis, particularly regarding deep learning methods, is relatively sparse.

Deep learning models (LeCun et al., 2015) are now ubiquitous in medical image analysis (Litjens et al., 2017), but as mentioned above, they rarely produce uncertainty estimates and tend to be poorly calibrated (Guo et al., 2017). Guo et al. (2017) recently showed that post-processing calibration methods can be used to calibrate the classification confidence of deep learning models. In deep learning models trained with maximum likelihood estimation, uncertainty can be computed from classification entropy (Settles, 2012; Kendall and Gal, 2017), but more reliable uncertainty estimates can be calculated from the predicted classification variance obtained from Bayesian methods (Gal and Ghahramani, 2016; Gal et al., 2017; Kendall and Gal, 2017). However, Bayesian training and inference tend to be computationally expensive, but Gal et al. (Gal et al., 2017; Kendall and Gal, 2017) have recently proposed tractable Bayesian methods. Although confidence calibration and classification uncertainty have been studied in medical image analysis (Eaton-Rosen et al., 2018; Bullock et al., 2018; Nair et al., 2020), their effect on classification accuracy and model interpretability, and how they interact in a classification process have not been studied, to the best of our knowledge.

In this paper, we study the roles of classification uncertainty (Settles, 2012; Gal and Ghahramani, 2016; Gal et al., 2017; Kendall and Gal, 2017) and post-processing confidence calibration techniques (Guo et al., 2017) in deep learning models applied to medical image classification. Focusing on a recently proposed five-class polyp classification from colonoscopy images (Singh et al., 2013; Pu et al., 2018; Tian et al., 2019) (see Fig. 2), we show that: 1) confidence calibration reduces expected calibration error (ECE) and maximum calibration error (MCE); and 2) rejecting test samples based on high classification uncertainty and low classification confidence improves classification accuracy and mean average precision. Based on this evidence, we propose a new deep learning classifier trained with Bayesian methods that relies on the use of classification uncertainty (Gal and Ghahramani, 2016; Gal et al., 2017; Kendall and Gal, 2017) and post-processing confidence calibration (Guo et al., 2017) to reject unreliable and low-confidence test samples to improve classification accuracy and model interpretation. Experimental results indicate that our proposed deep learning classifier, trained with Bayesian methods and confidence calibration, outperforms the current state of the art (SOTA) in the same data set (Tian et al., 2019), by rejecting samples that present high uncertainty and low confidence.

2. Literature Review

The inference process of deep learning models applied to medical image analysis problems generally disregards the uncertainty present in the classification result. In models trained with maximum likelihood estimation (MLE), a possible way of obtaining this uncertainty is through the classification entropy (Settles, 2012). More reliable uncertainty estimates can be achieved with Bayesian methods (Gal and Ghahramani, 2016; Gal et al., 2017; Kendall and Gal, 2017), where training and inference assumes that both the model and the observations are affected by noise processes, which represent the main sources of classification uncertainty. The major impediment for the use of Bayesian estimation in deep learning models has been the high computational cost of the training and inference algorithms (e.g., Markov-chain Monte-Carlo estimation (Gamerman and Lopes, 2006) or variational methods (Jaakkola and Jordan, 2000)), but this issue has recently been mitigated by Gal et al. (Gal and Ghahramani, 2016; Gal et al., 2017; Kendall and Gal, 2017). In parallel to our work, other papers (Eaton-Rosen et al., 2018; Nair et al., 2020) also proposed the use of Bayesian estimation in deep learning models for estimating the uncertainty in medical image segmentation, showing that the field is acknowledging the importance of this topic.

Modern deep learning models are extremely accurate for certain tasks (Shen et al., 2017), but they tend to be un-calibrated, where their classification confidence does not represent well their expected accuracy (Guo et al., 2017) – in fact, deep learning models generally produce over-confident results. Although confidence calibration has been studied in automated health care (Jiang et al., 2011), it has been largely overlooked by the medical image analysis community, particularly when dealing with deep learning models. This is surprising, given the importance of confidence calibration in clinical settings and also that the solution typically involves a simple post-processing stage (Guo et al., 2017). There are some notable exceptions that use confidence calibration, but provide little insight in its use (Bullock et al., 2018; Eaton-Rosen et al., 2018). Furthermore, it is also interesting to note the lack of papers exploring a combination of classification uncertainty and confidence calibration, which is explored in this paper.

Current automated polyp classification methods typically solve two-class (Hewett et al., 2012; Komeda et al., 2017), three-class (Hayashi et al., 2013; Ribeiro et al., 2017) or four-class (Iwatate et al., 2018) problems, while our approach is one of the first to focus on a recently proposed five-class polyp classification problem (Singh et al., 2013; Pu et al., 2018; Tian et al., 2019). Such classification is argued (Singh et al., 2013; Pu et al., 2018) to be more effective than previous two-class (Hewett et al., 2012; Komeda et al., 2017), three-class (Hayashi et al., 2013; Ribeiro et al., 2017) and four-class (Iwatate et al., 2018) problems because it will enable colonoscopists to make better informed decisions during a colonoscopy. In particular, with a reliable five-class classification, the colonoscopist will assess, using colonoscpy imaging, if a detected polyp is endoscopically resectable (i.e. pre-cancerous or early cancerous lesions – classes *IIo*, *II* and *IIIa*) or not-endoscopically resectable (i.e., benign or invasive cancer - classes *I* or *IIIb*, where for the latter class, the case is referred to surgery). Furthermore, this five-class polyp classification may reduce costs and complications associated with polypectomy because the colonoscopy follow-up interval and method for endoscopic resection can differ depending on the number, size and type of the lesions found during the exam (Levin et al., 2008; Rex et al., 2017). However, the advantages of such five-class polyp classification are counter-balanced with the increased difficulty in dealing with the more challenging problem during the training and inference processes of the classification model.

3. Material and Methods

3.1. Data Set

The experiments conducted in this paper are performed on a recently proposed five-class polyp classification problem (Singh et al., 2013; Pu et al., 2018; Tian et al., 2019), where polyps are labelled according to histology outcomes into the following

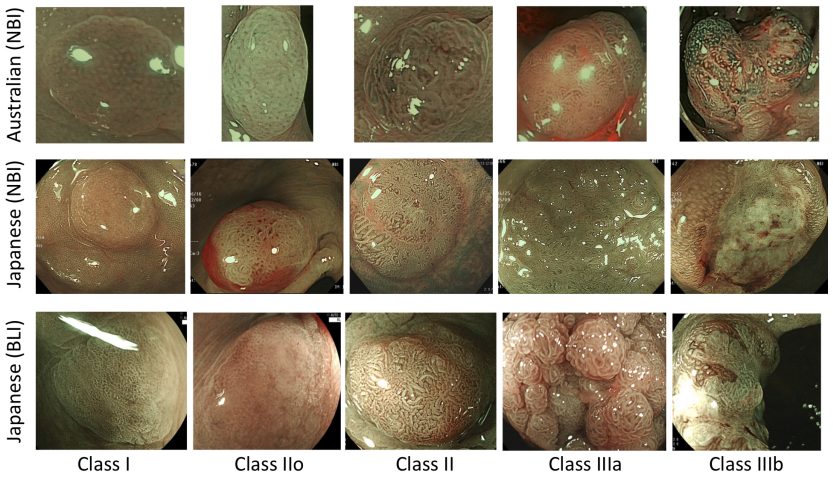

Fig. 2: Five-class polyp classification from colonoscopy images from the Australian and Japanese data sets.

classes (see Fig. 2): hyperplastic polyp (I), sessile serrated adenoma/polyp (IIo), low grade tubular adenoma (II), high grade adenoma/tubulovillous adenoma/superficial cancer ($IIIa$) and invasive cancer ($IIIb$). The data set is formally defined by $\mathcal{D} = \{\mathbf{x}_i, p_i, y_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} : \Omega \rightarrow \mathbb{R}^3$ denotes an RGB image of a polyp (Ω is the image lattice), as shown in Fig. 2, $p_i \in \mathbb{N}$ represents patient identification [1], and $y_i \in \mathcal{Y} = \{I, II, IIo, IIIa, IIIb\}$ denotes the polyp class, obtained from histology.

3.2. Australian and Japanese Data Sets

We test the performance of the proposed method on two data sets that are referred to as the Australian and Japanese data sets. The Australian data set consists of high-quality images of colorectal polyps obtained at a tertiary hospital in South Australia with the Olympus ®190 dual focus colonoscope using the most common advanced endoscopic imaging technology used in colonoscopy: Narrow Band Imaging (NBI). The number of images in the data set is $|\mathcal{D}| = 871$, which were scanned from 218 patients. This data set contains 102 images from 39 patients of class I, 346 images from 93 patients of class II, 281 images from 48 patients of class IIo, 79 images from 25 patients of class $IIIa$ and 63 images from 14 patients of class $IIIb$ – see first row of Fig. 2. The Japanese data set consists of two subsets of high-quality colorectal polyp images retrieved from a tertiary hospital image database: magnified NBI images acquired with the Olympus ®290 series, and magnified Blue Laser Imaging (BLI) images acquired with the Fujifilm ®700 series. The Japanese data set contains 20 NBI images from 20 patients and 49 BLI images from 49 patients. The NBI data set consists of 3 images of class I, 5 images of class II, 2 images of class IIo, 7 images of class $IIIa$ and 3 images of class $IIIb$ – see second row of Fig. 2. The BLI data set has 9 images of class I, 10 images of class II, 10 images of class IIo, 11 images of class $IIIa$ and 9 images of class $IIIb$ – see last row of Fig. 2. All images from the Australian and Japanese data sets were correlated with histology and de-identified into folders according to the MS classification (Singh et al., 2013; Pu et al., 2018; Tian et al., 2019). The Australian data set is used for training and testing the proposed method based on a cross validation experiment, while the Japanese data set is used exclusively for testing the system, enabling us to test the performance of the method on different colonoscopes with the same imaging technology (i.e., Olympus ®190 and 290 series) and on different technologies (i.e., NBI and BLI). The collection of colorectal lesions endoscopy images and clinical information was approved

[1] Note that the data set has been previously de-identified, and p_i is an "artificial" patient identification that is only useful for splitting $\mathcal{D}$ into training, validation and testing sets.

by the Human Research Ethics Committee (TQEH/LMH/MH/2008128 and HREC/16/TQEH/283) in Australia and by the Nagoya University Hospital Ethics Review Committee (2015-0485) in Japan.

3.3. Deep Learning Models

The deep learning models explored in this paper constitute the state-of-the-art models proposed by the fields of machine learning and computer vision. The models are: residual neural network (ResNet) (He et al., 2016) and densely connected networks (DenseNet) (Huang et al., 2017). The main characteristic defining these models is the presence of skip connections that short-cut network layers. ResNets are formed by short-cutting single layers and DenseNets are built with parallel short-cuts over single and multiple layers. Such models show state-of-the-art classification results in several challenges, such as ImageNet (Deng et al., 2009), and Microsoft COCO (Lin et al., 2014).

The deep learning model is formally represented by the classifier $P(y|\mathbf{x}, \mathbf{W})$, where $\mathbf{W}$ denotes the model parameters. For the ResNet, we adopt the ResNet-101 model, which has 101 layers, where each layer is formed by a sequence of convolutional filter, followed by batch normalisation and activation. The short-cut connections are placed over ResNet blocks, where each block consists of a sequence of layers (He et al., 2016). Regarding the DenseNet model, we adopt the DenseNet-121, which contains 121 layers of batch normalisation, followed by activation and convolution, and short-cut connections between layers inside a dense block (Huang et al., 2017). These deep learning classification models are estimated by fitting the training set (formed from $\mathcal{T} \subset \mathcal{D}$), in a maximum likelihood estimation, defined by:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} -\mathbb{E}_{P_{\mathcal{D}}(y|\mathbf{x})}[\log P(y|\mathbf{x}, \mathbf{W}))]. \tag{1}$$

The optimisation in (1) minimises the cross entropy between the empirical data distribution $P_{\mathcal{D}}(y|\mathbf{x})$ and the model output $P(y|\mathbf{x}, \mathbf{W})$ (Bishop, 2006). Model selection to estimate hyper-parameters (e.g., learning rate, number of layers, etc.) is performed with the validation set, represented by $\mathcal{V} \subset \mathcal{D}$, where $\mathcal{T} \bigcap \mathcal{V} = \emptyset$. Then, the inference process of a test sample denoted by $\mathbf{x} \in \mathcal{D} \setminus (\mathcal{T} \bigcup \mathcal{V})$ consists of estimating the class probability distribution $P(y|\mathbf{x}, \mathbf{W}^*)$.

3.4. Bayesian Learning and Inference

The training and inference procedures described in (1) are effective, but they do not provide a straightforward way to estimate the uncertainties present in the model and the observed data. These types of uncertainties can be estimated with Bayesian methods, which compute the class distribution for a sample $\mathbf{x}$ as follows (Gal and Ghahramani, 2016):

$$P(y|\mathbf{x}, \mathcal{T}) = \int_{\mathbf{W}} P(y|\mathbf{x}, \mathbf{W}) P(\mathbf{W}|\mathcal{T}) d\mathbf{W}, \tag{2}$$

where $\mathbf{W}$ is now considered a random variable with prior distribution $P(\mathbf{W})$ and posterior $P(\mathbf{W}|\mathcal{T})$, which is in general intractable. One way to mitigate the intractability of the posterior is to approximate $P(\mathbf{W}|\mathcal{T})$ with a tractable distribution $Q_\theta(\mathbf{W})$, parameterised by $\theta = \{\mathbf{M}_l, p_l\}_{l=1}^L$, with L representing the number of network layers, $\mathbf{M}_l$ denoting the layer-wise mean weight matrices and p_l the dropout probabilities, such that $Q_\theta(\mathbf{W}) = \prod_l \mathbf{M}_l \times \text{diag}(\text{Bernoulli}(1 - p_l)^{K_l})$, where the weight matrix in layer l has dimensions $K_{l+1} \times K_l$ (Gal and Ghahramani, 2016; Gal et al., 2017). This approximation allows (2) to be solved with Monte Carlo (MC) integration:

$$\begin{aligned} Q_\theta(y|\mathbf{x}) &= \frac{1}{N} \sum_{j=1}^{N} P(y|\mathbf{x}, \widehat{\mathbf{W}}_j) \\ &= \frac{1}{N} \sum_{j=1}^{N} \sigma(f^{\widehat{\mathbf{W}}_j}(\mathbf{x})), \end{aligned} \tag{3}$$

where $\widehat{\mathbf{W}}_j \sim Q_\theta(\mathbf{W})$ (for $j \in \{1, ..., N\}$) denotes one of the N samples drawn from $Q_\theta(.)$, $\sigma(.)$ denotes the softmax function, and $f^{\widehat{\mathbf{W}}_j}(\mathbf{x}) \in \mathbb{R}^{|\mathcal{Y}|}$ denotes the logit vector for the final softmax function applied by the classifier

The learning process that estimates θ^* in an optimisation that minimises the Kullback-Leibler (KL) divergence $KL(Q_\theta(\mathbf{W})||P(\mathbf{W}|\mathcal{T})$ is proportional to:

$$\ell(\theta) = -\int Q_\theta(\mathbf{W}) \log \prod_{i=1}^{|\mathcal{T}|} P(y_i|\mathbf{x}_i, \mathbf{W}) d\mathbf{W} + KL(Q_\theta(\mathbf{W})||P(\mathbf{W})), \tag{4}$$

where the integral is approximated with MC integration, as in $\ell(\theta) \approx -\log \prod_{i=1}^{|\mathcal{T}|} P(y_i|\mathbf{x}_i, \widehat{\mathbf{W}}) + KL(Q_\theta(\mathbf{W})||P(\mathbf{W}))$, with $\widehat{\mathbf{W}} \sim Q_\theta(\mathbf{W})$. Note that we assume to have a prior distribution $P(\mathbf{W})$ represented by the discrete quantised Gaussian prior (Gal et al., 2017) that allows an analytically derivation of $KL(Q_\theta(\mathbf{W})||P(\mathbf{W}))$ in (4), as explained in (Gal et al., 2017).

3.5. Post-processing Confidence Calibration

A well calibrated classification confidence indicates the actual likelihood that the classification is correct. Such calibration is important because deep learning models have been shown to produce over-confident classification probabilities (Guo et al., 2017) – that is, the classification probability tends to be higher than its correct likelihood. Hence, by calibrating confidence, we can improve the interpretability of the model result, which is a critical step toward the deployment of deep learning models in clinical settings. A recent study (Guo et al., 2017) showed that temperature scaling is an effective post-processing confidence calibration method, which works by modifying the output classification probability computation as follows:

$$P(y|\mathbf{x}, \mathcal{T}) \approx \widetilde{Q}_{\theta^*}(y|\mathbf{x}, s) = \frac{1}{N} \sum_{j=1}^{N} \sigma(\mathbf{f}^{\widehat{\mathbf{W}}_j}(\mathbf{x})/s), \tag{5}$$

where $\widehat{\mathbf{W}}_j \sim Q_{\theta^*}(\mathbf{W})$, and $\mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{x})$ is the logit defined in (3). The confidence for the non-Bayesian classifier can be similarly calibrated with $\widetilde{P}(y|\mathbf{x}, \mathbf{W}^*, s) = \sigma(\mathbf{f}^{\mathbf{W}^*}(\mathbf{x})/s)$, with $\mathbf{f}^{\mathbf{W}^*}(\mathbf{x})$ denoting the logit of the model trained as described in (1). In (5), $s \in \mathbb{R}^+$ is a learnable temperature parameter that smooths out the softmax function $\sigma(.)$ by raising its entropy. This parameter s is learned with stochastic gradient descent using the validation set $\mathcal{V}$ (we cannot use the training set to estimate s because the model tends to have overly optimistic results on the training set $\mathcal{T}$). We only consider temperature scaling because that was the confidence calibration method that produced the best results in (Guo et al., 2017).

3.6. Classification Uncertainty

We consider two ways to compute classification uncertainty. The first way is based on the entropy of the probability vector (Settles, 2012; Kendall and Gal, 2017):

$$H(P(y|\mathbf{x}, \mathcal{T})) = -\sum_{c \in \mathcal{Y}} P(y = c|\mathbf{x}, \mathcal{T}) \log(P(y = c|\mathbf{x}, \mathcal{T})), \tag{6}$$

where $P(y|\mathbf{x}, \mathcal{T})$ represents $P(y|\mathbf{x}, \mathbf{W}^*)$ in (1), $Q_{\theta^*}(y|\mathbf{x})$ in (3) and the calibrated classifiers from (5).

The second way relies on the computation of the predictive variance, approximated as (Kendall and Gal, 2017):

$$V(y|\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} \left(f^{\widehat{\mathbf{W}}_j}(\mathbf{x})\right)^T \left(f^{\widehat{\mathbf{W}}_j}(\mathbf{x})\right) - \left(\frac{1}{N} \sum_{j=1}^{N} f^{\widehat{\mathbf{W}}_j}(\mathbf{x})\right)^T \left(\frac{1}{N} \sum_{j=1}^{N} f^{\widehat{\mathbf{W}}_j}(\mathbf{x})\right). \tag{7}$$

It is not possible to compute the predictive variance for the non-Bayesian classifier $P(y|\mathbf{x}, \mathbf{W}^*)$ using (7) because of the dependence on the N samples drawn from $Q_{\theta^*}(\mathbf{W})$.

3.7. New Classifier that Relies on Uncertainty and Confidence Calibration

The main technical contribution of this paper is a deep learning classifier that uses classification uncertainty and calibrated confidence to reject the classification of test samples, where the goal is to improve the classification accuracy of our model. Such rejection process is based on hyper-parameters $\tau_1^*(Z)$, $\tau_2^*(Z)$, and $\tau_3^*(Z)$, learned from the validation set $\mathcal{V}$ with the goal of rejecting a certain percentage of test samples. More specifically, our proposed deep learning model accepts the classification result of a sample $\mathbf{x}$ based on two conditions:

$$\begin{aligned} &1) \quad P(y|\mathbf{x},\mathcal{T}) > \tau_1^*(Z) \\ &2) \quad H(Q_{\theta^*}(y|\mathbf{x})) < \tau_2^*(Z) \text{ or } V(y|\mathbf{x}) > \tau_3^*(Z), \end{aligned} \tag{8}$$

where in condition 1, $P(y|\mathbf{x},\mathcal{T})$ represents $P(y|\mathbf{x},\mathbf{W}^*)$ in (1), $Q_{\theta^*}(y|\mathbf{x})$ in (3) and the calibrated classifiers from (5), in condition 2, $H(.)$ is the entropy defined in (6) ($Q_{\theta^*}(y|\mathbf{x})$ can be replaced by its calibrated version from (5)) and $V(.)$ is the predictive variance computed from (7), and Z is the percentage of test samples to accept by the classification process. The thresholds are learned with

$$\begin{aligned} \tau_1^*(Z) &= P_{sorted}(Z \times |\mathcal{V}|), \\ \tau_2^*(Z) &= H_{sorted}(Z \times |\mathcal{V}|), \\ \tau_3^*(Z) &= V_{sorted}(Z \times |\mathcal{V}|), \end{aligned} \tag{9}$$

where P_{sorted} contains the values of $\max_{y \in \mathcal{Y}} P(y|\mathbf{x},\mathcal{T})$ sorted in descending order for all elements of the validation set $\mathcal{V}$, H_{sorted} contains the values of $H(Q_{\theta^*}(y|\mathbf{x}))$ sorted in ascending order for all elements of the validation set $\mathcal{V}$, and V_{sorted} contains the values of $V(y|\mathbf{x})$ sorted in ascending order for all elements of the validation set $\mathcal{V}$. We made the decision of using the percentage of test samples Z as a parameter for learning the classification probability, classification entropy and predicted variance thresholds because the actual values of classification probability, predictive variance and classification entropy are meaningless – they become meaningful when associated with Z. We test the proposed classifier with each condition in isolation and both conditions jointly.

4. Experiments

4.1. Experimental Set-up

We present two experiments: one based on training and testing on the Australian data set, and another based on training on the Australian data set and testing on the Japanese data set (see Sec. 3.2). The experiment based on training and testing on the Australian data set uses the results produced from a 5-fold cross validation procedure, where the training set $\mathcal{T} \subset \mathcal{D}$ contains the images from 60% of the patients, the validation set $\mathcal{V} \subset \mathcal{D}$ contains the images from 20% of the patients (where $\mathcal{T} \bigcap \mathcal{V} = \emptyset$), and the testing set $\mathcal{U} \subset \mathcal{D}$ has the remaining 20% of the patients, where $\mathcal{U} = \mathcal{D} \setminus (\mathcal{T} \bigcup \mathcal{V})$. Each one of these subsets are randomly formed to have the same proportion of five classes. The experiment based on training on the Australian data set and testing on the Japanese data set uses the five models learned from the 5-fold cross validation procedure to classify all the images from the Japanese data set. In both experiments, we are able to show mean and standard deviation from the five models.

The models ResNet-101 (He et al., 2016) and DenseNet-121 (He et al., 2016) have been developed in Keras (Chollet, 2015) with Tensorflow (Abadi et al., 2016) backend. Both models have been pre-trained using ImageNet (Deng et al., 2009) – the results from these pre-trained models were better than the ones produced by models trained from scratch (Bar et al., 2015). For fine-tuning the baseline (non-Bayesian) models, we remove the last 1000-node layer from the pre-trained model and replaced it by a softmax activated five-node layer, representing the five classes of the polyp classification problem. For Bayesian learning, we use concrete dropout (Gal et al., 2017), where the 1000-node layer from the original model is replaced by two fully connected layers: one

layer with five nodes activated by a rectified linear unit (ReLU) (Nair and Hinton, 2010) and a second layer with ten nodes (first five nodes activated by softmax, representing the classification probability vector, and the next five nodes denoting the aleatoric uncertainty (Kendall and Gal, 2017)). The parameters of the variational distribution $Q_\theta(\mathbf{W})$, represented by the mean values of the weights and the dropout probabilities (2) are learned only for these two last layers. For all training procedures, we use mini-batches of size 32, 800 training epochs, initial learning rate of 10^{-3}, which is decayed by 0.9 after every 50 training epochs, and 10× data augmentation (i.e., for every training image, we created 10 new images using random translations and scalings). The input image size is $224 \times 224 \times 3$ (the original polyp images acquired from colonoscopy videos are transformed to this size by bicubic interpolation). For the optimisation, we use Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. For Bayesian inference, the number of samples drawn from $Q_\theta(\mathbf{W})$ in (3) is $N = 10$. For training the confidence calibration, we retrain the last layer of the model for 100 epochs, using the validation set $\mathcal{V}$ to estimate s in (5).

The classification results are assessed in two ways: classification accuracy and average precision. Classification accuracy computes the proportion of correctly classified test samples, independently of their classes. This measure is regarded as sample-based because it is averaged over the whole test set, so classes that have more testing samples will have a higher impact on this measure. Another classification performance is provided by the average precision (AP) for each class, obtained by averaging the precision across all recall values between zero and one, and then calculating the mean AP over the five classes. AP is considered to be a class-based measure because it is averaged over the performance for each class, disregarding the imbalanced distribution that exists among the classes.

The calibration results are evaluated with the expected calibration error (ECE) and maximum calibration error (MCE). Both measures are computed from the reliability diagram, which plots sample accuracy as a function of confidence (Guo et al., 2017). The accuracy is measured by grouping predictions into M bins of size $1/M$ and computing the accuracy in each bin (in the experiments below, $M = 10$). Assuming that $\mathcal{B}_m$ represents the set of sample indices, whose confidence is in $\left(\frac{m-1}{M}, \frac{m}{M}\right]$, then $\text{acc}(\mathcal{B}_m) = \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \mathbf{1}(\hat{y}_i = y_i)$, where $\mathbf{1}(.)$ is the indicator function and $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} P(y|\mathbf{x}, \mathbf{W}^*)$ (for the Bayesian classifier, $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} Q^*_\theta(y|\mathbf{x})$). The confidence for each bin $\mathcal{B}_m$ is then defined by $\text{conf}(\mathcal{B}_m) = \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} P(y_i|\mathbf{x}_i, \mathbf{W}^*)$ (replace $P(.)$ by $Q_{\theta^*}(y|\mathbf{x})$ for the Bayesian classifier). The expected calibration error is measured with

$$ECE = \sum_{m=1}^{M} \frac{|\mathcal{B}_m|}{|\mathcal{U}|} \left|\text{acc}(\mathcal{B}_m) - \text{conf}(\mathcal{B}_m)\right|, \tag{10}$$

where $|\mathcal{U}|$ represents the number of samples in the test set; and the maximum calibration error is computed with

$$MCE = \max_{m \in \{1,...,M\}} \left|\text{acc}(\mathcal{B}_m) - \text{conf}(\mathcal{B}_m)\right|. \tag{11}$$

We test our proposed classifier from Sec. 3.7 by training the hyper-parameters $\tau^*_1(Z)$, $\tau^*_2(Z)$, and $\tau^*_3(Z)$ using the validation set $\mathcal{V}$, where Z defined in (9) is set as $Z \in \{0.5, 0.6, ..., 0.9, 1.0\}$ (for the training and testing on the Australian data set) and $Z \in \{0.7, 0.8, 0.9, 1.0\}$ (for the training on the Australian and testing on the Japanese data set – the range for this experiment is smaller because of the smaller size of the Japanese data set). In practice, our proposed classifier rejects high-uncertainty and low-confidence testing samples using the conditions in (8) , where the uncertainty is computed either from the classification entropy (6) or the predicted variance (7), and confidence is calculated by $P(y|\mathbf{x}, \mathbf{W}^*)$ from (1), $Q_{\theta^*}(y|\mathbf{x})$ from (3) or the calibrated classifiers from (5). We run experiments where each condition is applied in isolation (i.e., high-uncertainty or low-confidence), and both conditions are applied jointly.

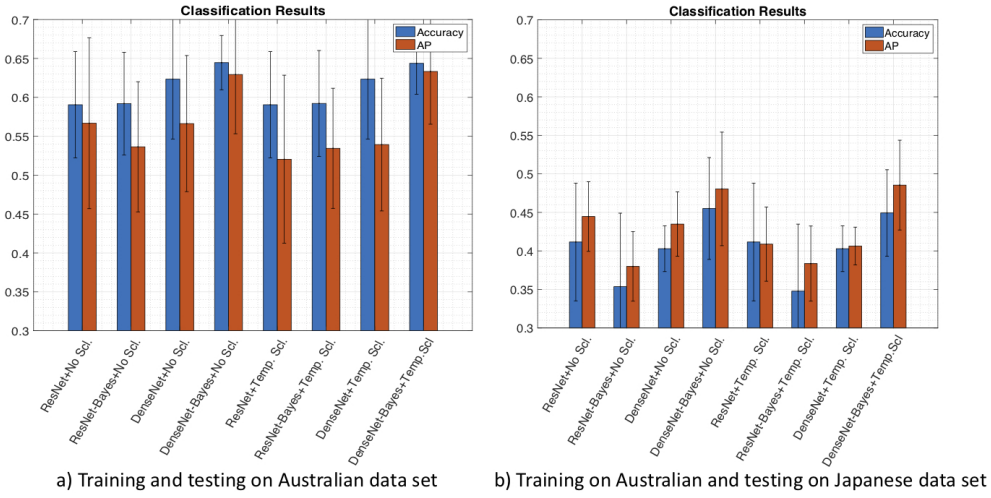

Fig. 3: Mean and standard deviation of the accuracy (ACC) and mean average precision (AP) results over the 5-fold cross validation experiment for (a) training and testing on Australian data set, and (b) training on Australian and testing on Japanese data set.

4.2. Experimental Results

In this section, models are labelled as follows: 1) model name (**ResNet** or **DenseNet**), 2) models trained with Bayesian learning and inference in (2) are labelled with **-Bayes** and models trained with maximum likelihood estimation and non-Bayesian inference have no label, and 3) models trained with confidence calibration are labelled as **+Temp. Scl.** and without calibration as **+No Scl.**. Therefore, the combinations above produce eight models to be analysed.

Figure 3 displays the accuracy and average precision for the eight models on the experiment using the training and testing Australian data set (left) and the training on Australian and testing on Japanese data sets (right). As explained in (Guo et al., 2017), notice that the classification accuracy results with and without temperature scaling do not change (minor differences are due to numerical reasons). The expected calibration error (10) and maximum calibration error (11) results are displayed in Fig. 4 – top shows the experiment with the training and testing Australian data set and bottom shows the training on Australian and testing on Japanese data sets. In general, calibrated methods produces smaller ECE and MCE (but notice that the MCE on the Japanese data set decreases only for the DenseNet models), which is another expected result (Guo et al., 2017). Figure 5 and 6 show the entropy (6) for all models and predictive variance for the Bayesian methods (7), respectively, for the two experiments. It is clear that entropy increases for methods calibrated with temperature scaling (Guo et al., 2017), and variance decreases for the calibrated Bayesian methods.

The experiments to test the proposed model from Sec. 3.7 (i.e., the model that rejects high uncertainty and low confidence samples) are shown in Figure 7–9. In particular, Figure 7 shows the results for Bayesian models, assuming predictive variance as uncertainty, while Fig. 8 displays the results for Bayesian models, assuming entropy as uncertainty, and Fig. 9 depicts results for non-Bayesian models, assuming entropy as uncertainty. In each of these figures, we show classification accuracy (first row) and average precision (second row) as a function of the predicted proportion of test set (i.e., the Z value in (9)) being rejected by high values of uncertainty (i.e., entropy or predictive variance – labelled as **uncertainty**), low values of confidence (labelled as **confidence**), or both (labelled as **uncert.+conf.**). The third row in these figures shows the actual proportion of test sets rejected as a function of Z (the horizontal axis, labelled as "Percentage of testing samples", represents Z). The fourth row of those figures shows a scatter plot between confidence and uncertainty to assess the negative correlation between these two measures. To measure

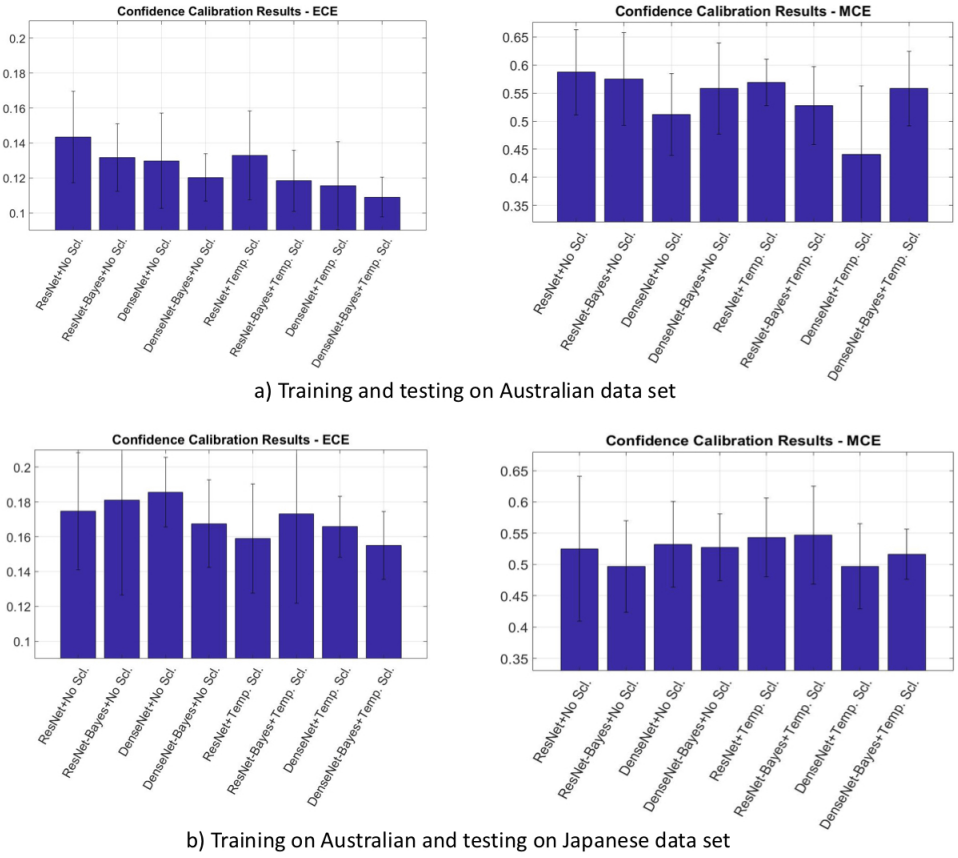

Fig. 4: Mean and standard deviation of the expected calibration error (ECE) (10) and maximum calibration error (MCE) (11) results over the 5-fold cross validation experiment for (a) training and testing on Australian data set, and (b) training on Australian and testing on Japanese data set.

the correlation between the two uncertainty measures for the Bayesian methods, we also show a scatter plot between entropy and predictive variance in Fig. 10.

We provide a comparison with our implementation of the state of the art method in the five-class polyp data set (Tian et al., 2019) in Fig. 11 and Tables 1-2 [2], where we show the results of our proposed method based on Bayesian training with temperature scaling calibration and with the rejection of samples based on confidence and uncertainty (we relied on the entropy of the probability vector for these results). The samples for Tian et al.'s method (Tian et al., 2019) are rejected based solely on the first condition in Eq. 8 (i.e., classification probability), where for both methods, results vary as a function of Z, i.e., the proportion of rejected test samples. We also show a few qualitative results in Fig. 12 produced by our proposed approach, depicting correctly and incorrectly classified test samples that presented high/low confidence and uncertainty. For these results we arbitrarily set the classification probability threshold to 90% and the predictive variance to 0.01 – such specific thresholds enable a classification accuracy of around 0.76.

5. Discussion

We present results of DenseNet and ResNet for two reasons: 1) show that similar results can be obtained independently of the model, and 2) show that confidence calibration and uncertainty work well with state-of-the-art models. In general, from Figures 3–6,

[2] The results by Pu et al. (Pu et al., 2018) were obtained using a different experimental set-up and are not directly comparable to the ones in this paper and in (Tian et al., 2019). Our implementation reached an accuracy of 0.59 ± 0.07 and an AP of 0.57 ± 0.11, while the original method in (Tian et al., 2019) had accuracy of 0.60 ± 0.05 and AP of 0.56 ± 0.04 – this shows that the two implementations produce similar classification results.

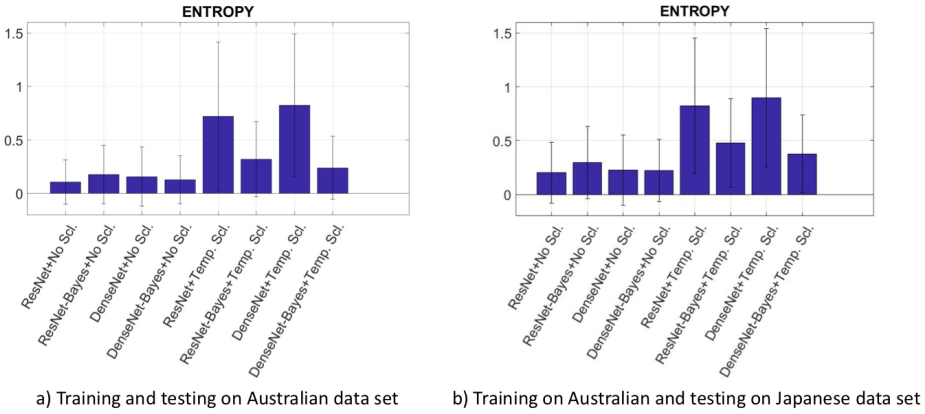

Fig. 5: Mean and standard deviation of the classification entropy (6) over the 5-fold cross validation experiment for (a) training and testing on Australian data set, and (b) training on Australian and testing on Japanese data set..

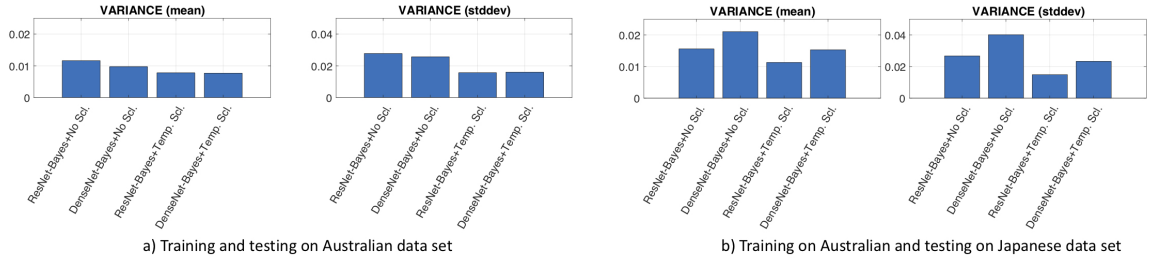

Fig. 6: Mean and standard deviation of the predictive variance (7) of the Bayesian methods over the 5-fold cross validation experiment for (a) training and testing on Australian data set, and (b) training on Australian and testing on Japanese data set.

it is clear that DenseNet tends to perform better than ResNet, confirming a similar result observed in other challenges (Deng et al., 2009) – this result is consistent across the two experiments (training and testing on Australian data set, and training on Australian and testing on Japanese. The Bayesian DenseNet (with and without calibration) is the top performer among all models presented in this paper. Also, confidence calibration reduces the calibration errors (ECE and MCE), a result that is extremely important for improving the interpretability of deep learning models in medical image analysis problems. There are two additional consequences of the use of confidence calibration: 1) increase of classification entropy (Fig. 5), as stated in Sec. 3.5; and 2) decrease of the standard deviation of the predicted variance (Fig. 6), indicating a more stable prediction of classification uncertainty from (7). From Fig. 3, it is clear that accuracy and AP results reduce significantly for the second experiment (training on Australian and testing on Japanese data set), indicating that further studies are necessary to improve the generalisation of the proposed method. We also

Table 1: Table of the mean and standard deviation (in brackets) of the accuracy (ACC) and AP as a function of Z (i.e., proportion of rejected samples) for all the models tested in the comparison with SOTA on Fig. 11 for **training and testing on Australian data set**. The best result per column is highlighted.

	Z=1.0		Z=0.9		Z=0.8		Z=0.7		Z=0.6		Z=0.5	
	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP	ACC	AP
ResNet+No Scl.	.59(.06)	.57(.11)	.61(.06)	.58(.11)	.64(.07)	.59(.12)	.68(.06)	.60(.12)	.69(.07)	.61(.13)	.71(.08)	.62(.12)
ResNet-Bayes+No Scl.	.59(.06)	.54(.08)	.62(.07)	.55(.08)	.63(.08)	.56(.07)	.65(.07)	.57(.07)	.67(.08)	.59(.07)	.69(.09)	.62(.09)
DenseNet+No Scl.	.62(.08)	.57(.09)	.65(.08	.56(.08)	.68(.08)	.56(.10)	.68(.09)	.50(.10)	.69(.09)	.52(.11)	.69(.09)	.51(.11)
DenseNet-Bayes+No Scl.	**.64(.03)**	**.63(.08)**	**.68(.04)**	**.65(.08)**	**.70(.04)**	.64(.07)	**.74(.05)**	.65(.08)	**.76(.04)**	.66(.09)	.78(.05)	.66(.11)
ResNet+Temp. Scl.	.59(.06)	.52(.11)	.61(.06)	.53(.11)	.63(.07)	.54(.12)	.64(.06)	.55(.12)	.63(.06)	.57(.11)	.64(.05)	.61(.12)
ResNet-Bayes+Temp. Scl.	.59(.06)	.53(.08)	.63(.07)	.55(.07)	.64(.07)	.56(.07)	.65(.07)	.57(.06)	.66(.08)	.59(.08)	.68(.07)	.62(.10)
DenseNet+Temp. Scl.	.62(.08)	.54(.08)	.65(.08)	.54(.08)	.65(.08)	.54(.08)	.65(.08)	.55(.08)	.65(.08)	.57(.08)	.65(.08)	.57(.08)
DenseNet-Bayes+Temp. Scl.	**.64(.03)**	**.63(.07)**	**.68(.04)**	**.65(.07)**	**.70(.04)**	**.65(.06)**	.73(.04)	**.66(.07)**	**.76(.03)**	**.67(.09)**	**.79(.05)**	**.68(.14)**
Tian et al. ISBI'19	.59(.07)	.57(.11)	.61(.07)	.58(.11)	.64(.07)	.59(.12)	.68(.07)	.60(.12)	.69(.07)	.61(.13)	.71(.08)	.62(.12)

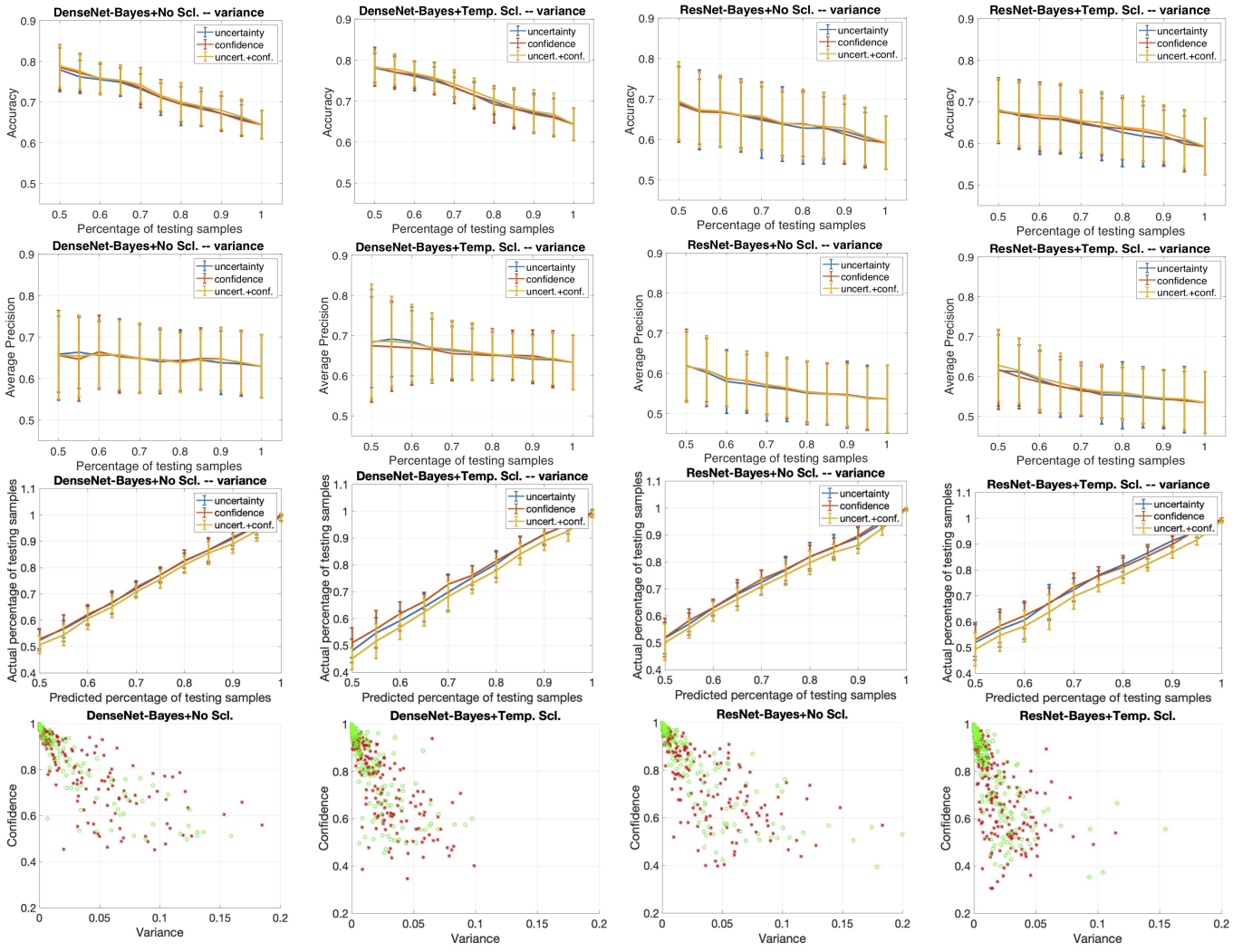

Fig. 7: Accuracy (row 1) and average precision (row 2) of the **Bayesian models** trained **with calibration** (columns 2 and 4) and **without calibration** (columns 1 and 3) as a function of the predicted proportion of test set (i.e., the Z value in (9)) being rejected by **predicted variance values (labelled as uncertainty)**, **classification probability (labelled as confidence)**, or **both (labelled as uncert.+conf.)**. Row 3 shows the actual proportion of test samples rejected as a function of Z (labelled as "Percentage of testing samples" in the horizontal axis), and Row 4 shows a scatter plot between confidence and uncertainty, where green points represent the correctly classified test samples and red, the incorrect ones.

note a good generalisation in terms of the ECE and MCE results for the second experiment in Fig. 4, except for the MCE results for the ResNet models trained with calibration. Another important observation is the discrepancy in the improvement observed for accuracy, compared with the improvement for average precision. Notice from Sec. 3.2 that the training set is highly imbalanced, with around 40% of the training samples belonging to class *II* and more than 20% to class *IIo*, which tends to bias the classification probability towards these majority classes. Such effect can explain this better accuracy improvement, particularly for the training and testing on Australian set experiment. The usual methods to fix this issue are to under-sample the majority classes, or over-sample the minority classes, or to re-weight training samples based on the proportion of samples of their classes. We tried all these approaches with little differences in final results, so this is an issue that needs further investigation.

The experiments in Figures 7–9 show that rejecting samples based on uncertainty and classification confidence improves classification accuracy and average precision for all Bayesian methods, where DenseNet-Bayes+Temp.Scl. shows the best overall improvements. For the non-Bayesian methods, it is possible to see some improvement for the methods trained without confidence calibration. We investigated this issue by looking at the confidence versus entropy scatter plot (row 4 in Fig. 9) and found that

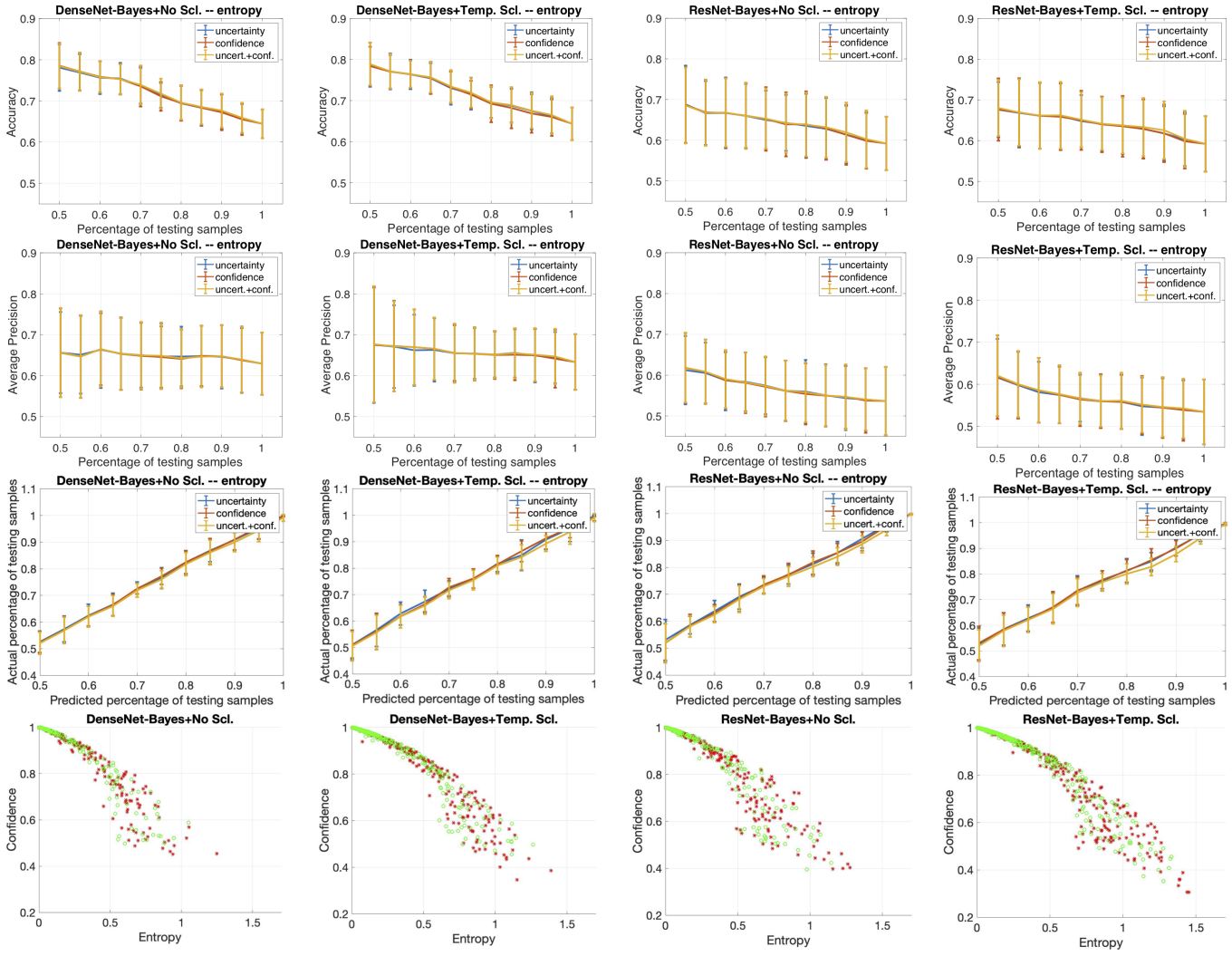

Fig. 8: Accuracy (row 1) and average precision (row 2) of the **Bayesian models** trained **with calibration** (columns 2 and 4) and **without calibration** (columns 1 and 3) as a function of the predicted proportion of test set (i.e., the Z value in (9)) being rejected by **entropy values (labelled as uncertainty)**, **classification probability (labelled as confidence)**, or **both (labelled as uncert.+conf.)**. Row 3 shows the actual proportion of test samples rejected as a function of Z (labelled as "Percentage of testing samples" in the horizontal axis), and Row 4 shows a scatter plot between confidence and uncertainty, where green points represent the correctly classified test samples and red, the incorrect ones.

this is related to a result produced by temperature scaling that distributes the classification over the range [0.2,1.0], placing a large proportion of correctly classified points in the bottom-right part of the graph that are likely to be rejected. This results in lack of classification performance improvement, as shown by the figure. Another interesting observation from Figures 7–9 is that the combination of uncertainty and confidence for rejecting samples provides an upper-bound performance for each condition in isolation, indicating that there is some complementarity in these two conditions. Reinforcing this argument, even though row 4 of Figures 7–9 shows a negative correlation between confidence and uncertainty, it is also possible to notice some scattering for low confidence, high uncertainty samples. The scatter plots between entropy and predictive variance in Fig. 10 for the Bayesian methods show that these two measures are indeed correlated, particularly for low entropy and low variance cases. This fact is noticeable from Figures 7 and 8 that show no significant difference between the use of predicted variance and entropy as uncertainty measures. Another important observation is the high correlation between the actual percentage of testing samples and the predicted percentage of testing samples Z, shown in row 3 of Figures 7–9. This demonstrates the reliability of the training to estimate the values of $\tau_1^*(Z)$, $\tau_2^*(Z)$, and $\tau_3^*(Z)$ in (9).

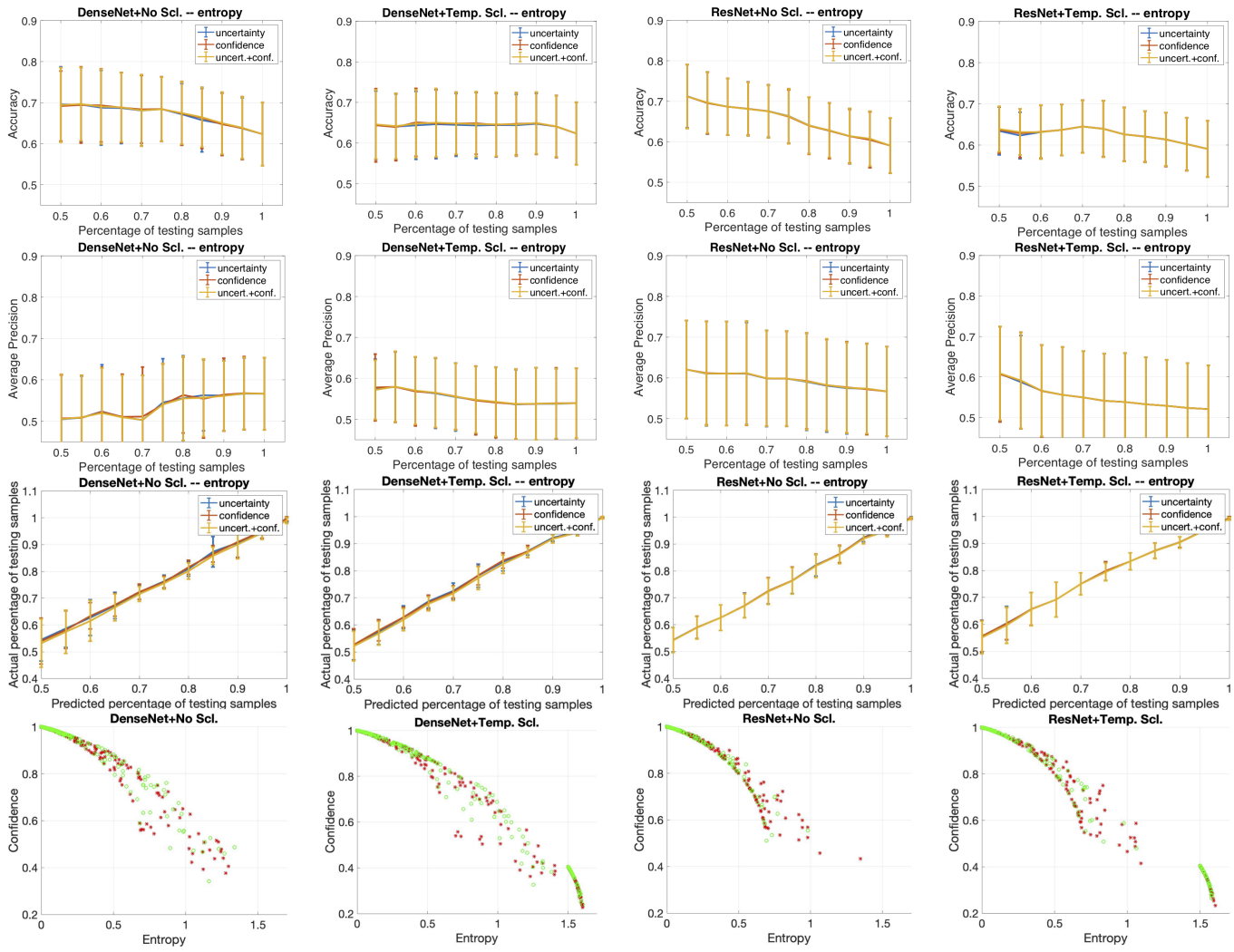

Fig. 9: Accuracy (row 1) and average precision (row 2) of the **non-Bayesian models** trained **with calibration** (columns 2 and 4) and **without calibration** (columns 1 and 3) as a function of the predicted proportion of test set (i.e., the Z value in (9)) being rejected by **entropy values (labelled as uncertainty)**, **classification probability (labelled as confidence)**, or **both (labelled as uncert.+conf.)**. Row 3 shows the actual proportion of test samples rejected as a function of Z (labelled as "Percentage of testing samples" in the horizontal axis), and Row 4 shows a scatter plot between confidence and uncertainty, where green points represent the correctly classified test samples and red, the incorrect ones.

Comparing our results with the state of the art in the five-class polyp classification problem for the first experiment (training and testing on Australian data set) in Fig. 11(a) and Tab. 1, our proposed DenseNet-Bayes methods without rejecting samples have classification accuracy around 64% and average precision 63%, which are slightly superior to the current state of the art (SOTA) (Tian et al., 2019) that uses the same experimental set-up and obtains classification accuracy around 59% and average precision around 57%, with an un-calibrated ResNet-50 model. When rejecting samples based on uncertainty and calibrated confidence, then our approach provides a substantial improvement over the SOTA results, where the classification accuracy for DenseNet-Bayes methods reaches around 70% when rejecting around 20% of the testing samples and close to 80% when rejecting 50% of the testing samples. The average precision also shows improvements, where we reach around 64% when rejecting around 20% of the testing samples and around 68% when rejecting 50% of the testing samples. The rejection process for the SOTA (Tian et al., 2019), based on uncalibrated confidence, also shows improvements, where it reaches 71% accuracy and 62% average precision when rejecting 50% of the testing samples. However, such SOTA improvements are not competitive to the results produced by our proposed method. For the second experiment with the training on Australian and testing on Japanese data sets, shown in Fig. 11(b) and

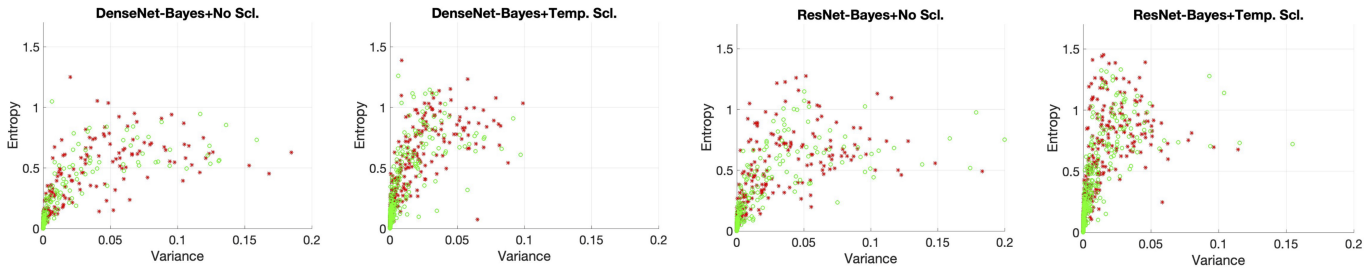

Fig. 10: Scatter plot between the two uncertainty measures entropy and variance for Bayesian methods, trained with (columns 2 and 4) and without (columns 1 and 3) calibration, where green points represent the correctly classified test samples and red, the incorrect ones.

Table 2: Table of the mean and standard deviation (in brackets) of the accuracy (ACC) and AP as a function of Z (i.e., proportion of rejected samples) for all the models tested in the comparison with SOTA on Fig. 11 for **training on Australian and testing on Japanese data set**. The best result per column is highlighted.

	Z=1.0		Z=0.9		Z=0.8		Z=0.7	
	ACC	AP	ACC	AP	ACC	AP	ACC	AP
ResNet+No Scl.	.41(.07)	.44(.05)	.44(.09)	.46(.07)	.48(.08)	.48(.06)	.49(.09)	.48(.05)
ResNet-Bayes+No Scl.	.35(.09)	.38(.05)	.37(.08)	.40(.04)	.38(.09)	.41(.06)	.38(.08)	.42(.06)
DenseNet+No Scl.	.40(.03)	.44(.04)	.43(.04)	.44(.06)	.46(.07)	.44(.05)	.47(.08)	.44(.04)
DenseNet-Bayes+No Scl.	**.45(.07)**	**.48(.07)**	.47(.06)	**.49(.06)**	**.48(.06)**	.48(.07)	.49(.07)	.47(.06)
ResNet+Temp. Scl.	.41(.07)	.41(.05)	.43(.05)	.42(.06)	.42(.05)	.42(.05)	.40(.07)	.43(.06)
ResNet-Bayes+Temp. Scl.	.35(.09)	.38(.05)	.35(.07)	.41(.05)	.37(.07)	.41(.05)	.39(.09)	.43(.06)
DenseNet+Temp. Scl.	.40(.03)	.41(.02)	.44(.05)	.40(.01)	.43(.04)	.41(.03)	.42(.04)	.42(.03)
DenseNet-Bayes+Temp. Scl.	**.45(.06)**	**.48(.06)**	**.48(.06)**	.48(.05)	**.48(.06)**	**.49(.07)**	**.51(.06)**	**.48(.05)**
Tian et al. ISBI'19	.41(.08)	.44(.05)	.44(.09)	.46(.07)	**.48(.09)**	.47(.06)	.49(.09)	**.48(.05)**

Tab. 2, we notice that the classification accuracy of the DenseNet-Bayes methods (in particular the one with calibration) starts at around 45% and reaches 51% when rejecting around 30% of the testing samples. This compares favourably with the SOTA (Tian et al., 2019) method that starts with accuracy around 41% and reaches 49% when rejecting around 30% of the testing samples. Regarding AP, results of the DenseNet-Bayes methods are stable at around 48% with the rejection of testing samples, while for the SOTA (Tian et al., 2019), results improve from 44% to around 48% when rejecting around 30% of the testing samples. Such results may suggest that rejecting samples based on uncertainty and calibrated confidence from the DenseNet-Bayes models is not more effective than rejecting samples based on uncalibrated confidence from the SOTA model given that the improvement for both models is similar. However, such conclusion is unwarranted because the baseline classification results (i.e., with no samples being rejected) for DenseNet-Bayes models is more accurate than for the SOTA model, which means that the identification of uncertain and low confident samples is significantly more challenging for DenseNet-Bayes models.

6. Conclusions and Future Work

In this paper, we studied the interaction and the roles of confidence calibration (via post-process temperature scaling) and classification uncertainty (through classification entropy or predictive variance from Bayesian methods) on classification accuracy and calibration errors. The main conclusions were: 1) confidence calibration reduces calibration errors; and 2) rejecting test samples based on high classification uncertainty and low classification confidence improves classification accuracy and average precision for Bayesian methods. These results motivated us to develop a new Bayesian deep learning model trained with post-processing confidence calibration that produces highly interpretable classification uncertainty and calibrated confidence that holds the current state-of-the-art classification accuracy for the five-class polyp classification (Tian et al., 2019), after rejecting samples with low confidence and high uncertainty.

The method presented in this paper is a proof of concept that can be potentially used in a clinical setting with the colour labels

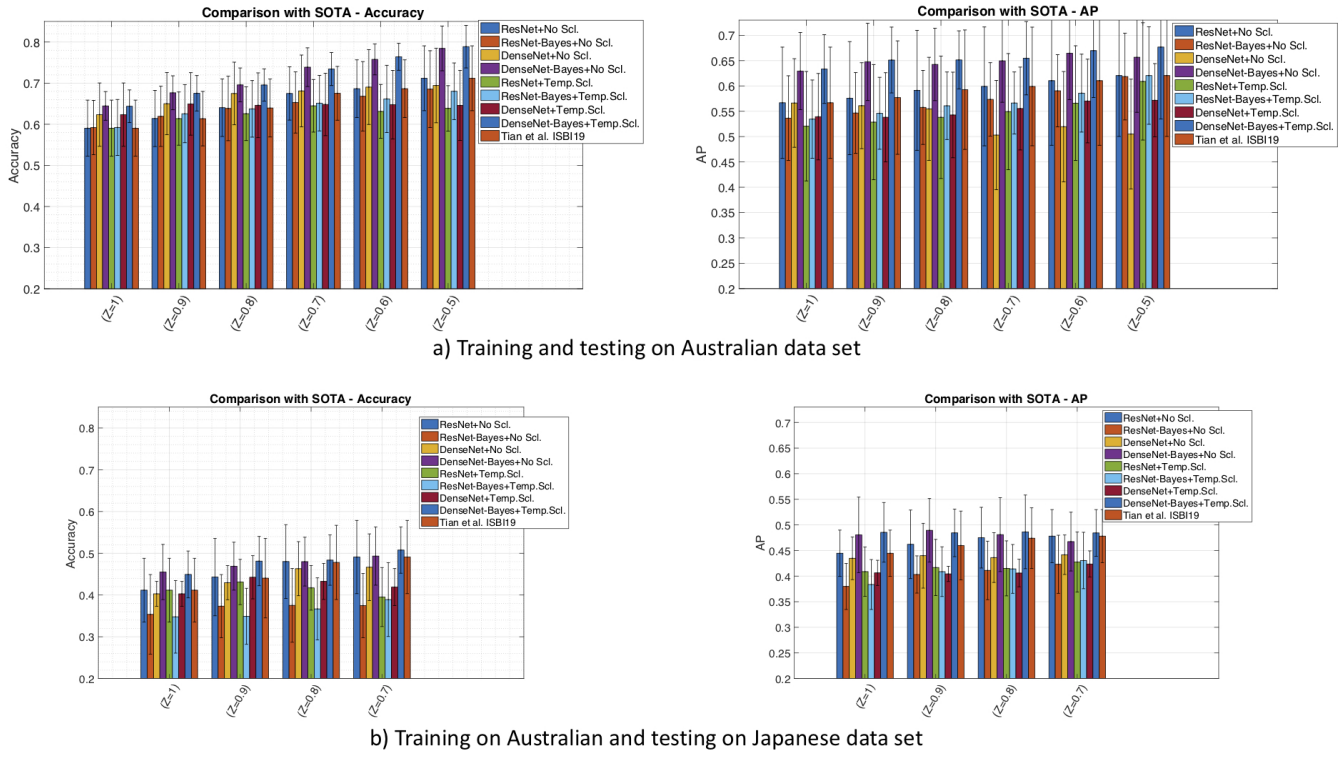

Fig. 11: Accuracy and AP comparison with the state of the art in the five-class polyp data set (Tian et al., 2019) as a function of Z (i.e., proportion of rejected samples) for (a) training and testing on Australian data set (numerical results in Tab. 1), and (b) training on Australian and testing on Japanese data set (numerical results in Tab. 2).

depicted in Fig. 1, where for example, when the system has high confidence and low uncertainty, it shows a green flag to the clinician. This green flag indicates that the clinician can be biased towards accepting the result produced by the system. On the other hand, for the cases of low confidence and high uncertainty, then the system shows the yellow or red flags, indicating that the clinician should be careful with the result produced by the method. Therefore, the system and clinician will have to work together for reaching a diagnosis. The actual testing of the system in a clinical setting is left for future work with our clinical collaborators. Also, the testing results with the Japanese data set indicates an important point for further investigation: how to generalise better to different domains in terms of classification and calibration results. Furthermore, imbalanced training is another important point that needs to be further studied to make more effective use of the training sets currently available to model these systems.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H., 2015. Chest pathology detection using deep learning with non-medical training, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 294–297.

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Bullock, J., Cuesta-Lazaro, C., Quera-Bofarull, A., 2018. Xnet: A convolutional neural network (cnn) implementation for medical x-ray image segmentation suitable for small datasets. arXiv preprint arXiv:1812.00548 .

Chollet, F., 2015. keras. `[https://github.com/fchollet/keras`.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 248–255.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J., 2018. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 691–699.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, pp. 1050–1059.

Gal, Y., Hron, J., Kendall, A., 2017. Concrete dropout, in: Advances in Neural Information Processing Systems, pp. 3581–3590.

Gamerman, D., Lopes, H.F., 2006. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman and Hall/CRC.

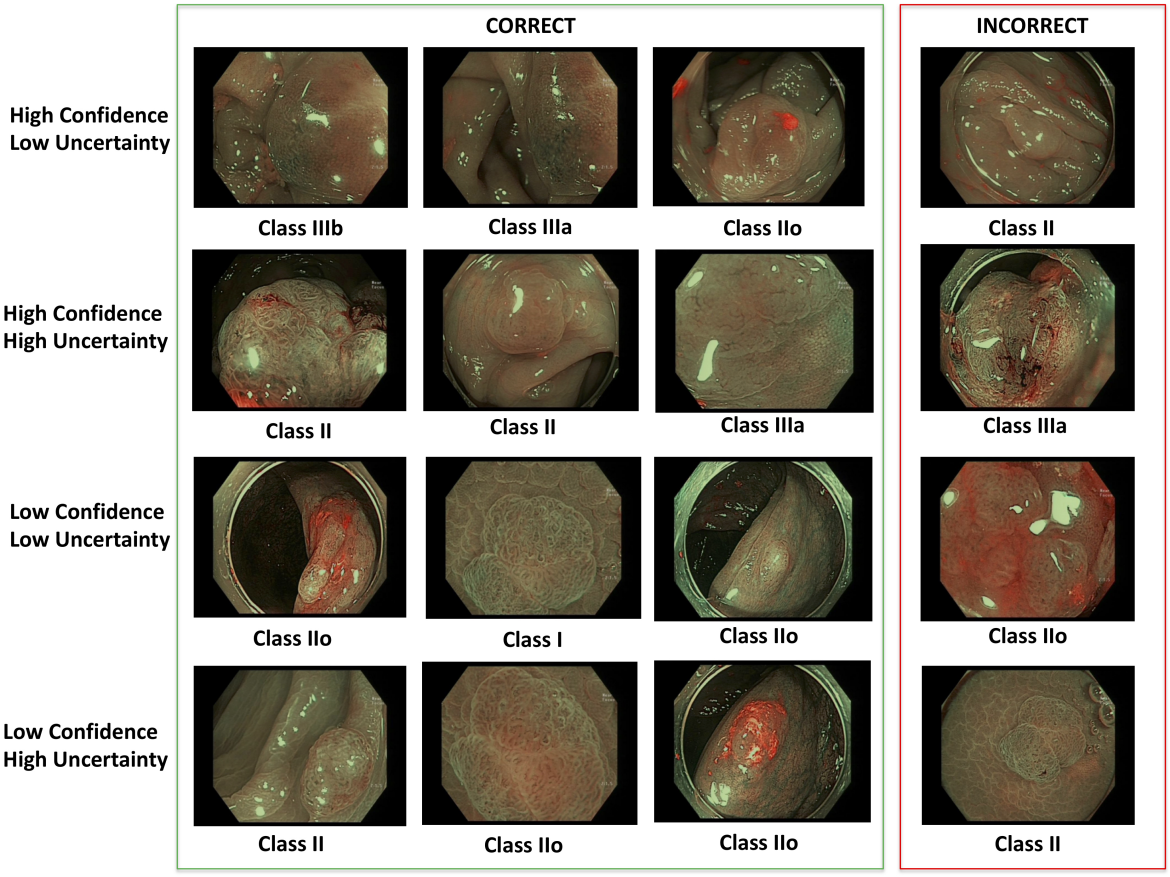

Fig. 12: Examples of results produced by DenseNet model based on Bayesian training with temperature scaling calibration, where high confidence are results where the classification probability is above 90% (consequently, low confidence is below 90%), and low uncertainty has predictive variance below 0.01 (and high uncertainty has predictive variance above 0.01).

Goodman, B., Flaxman, S., 2016. European union regulations on algorithmic decision-making and a" right to explanation". arXiv .

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. arXiv preprint arXiv:1706.04599 .

Hayashi, N., Tanaka, S., Hewett, D.G., Kaltenbach, T.R., Sano, Y., Ponchon, T., Saunders, B.P., Rex, D.K., Soetikno, R.M., 2013. Endoscopic prediction of deep submucosal invasive carcinoma: validation of the narrow-band imaging international colorectal endoscopic (nice) classification. Gastrointestinal endoscopy 78, 625–632.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hewett, D.G., Kaltenbach, T., Sano, Y., Tanaka, S., Saunders, B.P., Ponchon, T., Soetikno, R., Rex, D.K., 2012. Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. Gastroenterology 143, 599–607.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Iwatate, M., Sano, Y., Tanaka, S., Kudo, S.e., Saito, S., Matsuda, T., Wada, Y., Fujii, T., Ikematsu, H., Uraoka, T., et al., 2018. Validation study for development of the japan nbi expert team classification of colorectal lesions. Digestive Endoscopy 30, 642–651.

Jaakkola, T.S., Jordan, M.I., 2000. Bayesian parameter estimation via variational methods. Statistics and Computing 10, 25–37.

Jiang, X., Osl, M., Kim, J., Ohno-Machado, L., 2011. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association 19, 263–274.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, pp. 5574–5584.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Komeda, Y., Handa, H., Watanabe, T., Nomura, T., Kitahashi, M., Sakurai, T., Okamoto, A., Minami, T., Kono, M., Arizumi, T., et al., 2017. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology 93, 30–34.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

Levin, B., Lieberman, D.A., McFarland, B., Smith, R.A., Brooks, D., Andrews, K.S., Dash, C., Giardiello, F.M., Glick, S., Levin, T.R., et al., 2008. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology. CA: a cancer journal for clinicians 58, 130–160.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

Lipton, Z.C., 2016. The mythos of model interpretability. arXiv .

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation.

Medical image analysis 59, 101557.
Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807–814.
Pu, L.Z.C.T., Campbell, B., Burt, A.D., Carneiro, G., Singh, R., 2018. Computer-aided diagnosis for charaterising colorectal lesions: Interim results of a newly developed software. Gastrointestinal Endoscopy 87, AB245.
Rex, D.K., Boland, C.R., Dominitz, J.A., Giardiello, F.M., Johnson, D.A., Kaltenbach, T., Levin, T.R., Lieberman, D., Robertson, D.J., 2017. Colorectal cancer screening: recommendations for physicians and patients from the us multi-society task force on colorectal cancer. The American journal of gastroenterology 112, 1016.
Ribeiro, E., Häfner, M., Wimmer, G., Tamaki, T., Tischendorf, J., Yoshida, S., Tanaka, S., Uhl, A., 2017. Exploring texture transfer learning for colonic polyp classification via convolutional neural networks, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE. pp. 1044–1048.
Settles, B., 2012. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6, 1–114.
Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.
Singh, R., Jayanna, M., Navadgi, S., Ruszkiewicz, A., Saito, Y., Uedo, N., 2013. Narrow-band imaging with dual focus magnification in differentiating colorectal neoplasia. Digestive Endoscopy 25, 16–20.
Tian, Y., Pu, L., Singh, R.B., Alastair Carneiro, G., 2019. One-stage five-class polyp detection and classification, in: Biomedical Imaging (ISBI 2019), 2017 IEEE 16th International Symposium on, IEEE.