Robust Colorectal Polyp Characterization Using a Hybrid Bayesian Neural Network

Robust Colorectal Polyp Characterization Using a Hybrid Bayesian Neural Network

Nikoo Dehghani[1(✉)], Thom Scheeve[1], Quirine E. W. van der Zander[2,5], Ayla Thijssen[2,5], Ramon-Michel Schreuder[3], Ad A. M. Masclee[2], Erik J. Schoon[3,5], Fons van der Sommen[1,4], and Peter H. N. de With[1]

[1] Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands
n.dehghani@tue.nl
[2] Maastricht University Medical Center+, 6229 HX Maastricht, The Netherlands
[3] Catharina Hospital, 5623 EJ Eindhoven, The Netherlands
[4] Eindhoven Artificial Intelligence Systems Institute, 5612 AZ Eindhoven, Netherlands
[5] GROW - School for Oncology and Reproduction, 6211 LK Maastricht, Netherlands

Abstract. Computer-Aided Diagnosis (CADx) systems can play a crucial role as a second opinion for endoscopists to improve the overall optical diagnostic performance of colonoscopies. While such supportive systems hold great potential, optimal clinical implementation is currently impeded, since deep neural network-based systems often tend to overestimate the confidence about their decisions. In other words, these systems are poorly calibrated, and, hence, may assign high prediction scores to samples associated with incorrect model predictions. For the optimal clinical workflow integration and physician-AI collaboration, a reliable CADx system should provide accurate and well-calibrated classification confidence. An important application of these models is characterization of Colorectal polyps (CRPs), that are potential precursor lesions of Colorectal cancer (CRC). An improved optical diagnosis of CRPs during the colonoscopy procedure is essential for an appropriate treatment strategy. In this paper, we incorporate Bayesian variational inference and investigate the performance of a hybrid Bayesian neural network-based CADx system for the characterization of CRPs. Results of conducted experiments demonstrate that this Bayesian variational inference-based approach is capable of quantifying model uncertainty along with calibration confidence. This framework is able to obtain classification accuracy comparable to the deterministic version of the network, while achieving a 24.65% and 9.14% lower Expected Calibration Error (ECE) compared to the uncalibrated and calibrated deterministic network using a post-processing calibration technique, respectively.

Keywords: Colorectal polyp characterization · Bayesian inference · Model calibration · Classification uncertainty

S. Ali et al. (Eds.): CaPTion 2022, LNCS 13581, pp. 108–117, 2022.
https://doi.org/10.1007/978-3-031-17979-2_11

1 Introduction

Colorectal cancer (CRC) ranks third in terms of most diagnosed cancer and appears as the second cause of cancer deaths in the world [1]. Colorectal polyps (CRPs) are precursor lesions of CRC and can be divided into two major categories, non-neoplastic and neoplastic. Non-neoplastic polyps, including Hyperplastic polyps (HP), are considered as benign polyps. In contrast, neoplastic polyps are consisting of the Adenomas (ADs) and Sessile Serrated Lesions (SSLs) and can harbor a malignant potential. It is possible to prevent CRC if these polyps are detected and removed at an early stage of the disease [2]. Colonoscopy is the most common procedure for screening and characterization of CRPs. Computer-aided diagnosis (CADx) systems can assist physicians with a more reliable diagnosis, by characterizing CRPs using optical methods.

With the advancement of deep neural networks, excellent results obtained by different CADx systems have been reported in literature for detection [3,4], segmentation [3,5,6], or classification [4,7–9] of CRPs. However, despite their recent success, these systems have not been extensively adopted in the clinical pilot studies so far. An important reason for the slow adoption of these systems is that neural networks are often over-confident in their decisions and fail to express the uncertainty over their predictions [10]. Thus, these systems may produce high class probabilities for incorrect predictions. These high-probability predictions can create harmful biases on physicians' decisions and become life-threatening in a clinical setting. Therefore, it is important that a model is capable of producing well-calibrated classification confidence along with its predictions.

Research on confidence calibration and the estimation of classification uncertainty, in the field of CRP characterization, has been limited. In [11,12], the authors investigated the roles of confidence calibration in CRP characterization via extra post-processing steps. As an alternative, alleviating the need for such additional training stages, Bayesian models have been widely adopted in different applications due to their ability to capture reliable uncertainty measures over the decision of the network during the training process, as evidenced by work of Krishnan *et al.* [13] for activity recognition. Bayesian neural networks (BNNs) [14,15] offer a probabilistic interpretation of deep learning models, by placing distributions over the model parameters and thereby learning from an ensemble of possible distributions of weights. Conversely, conventional Deep Neural Networks (DNNs) tend to disregard uncertainty around the model parameters by obtaining maximum likelihood estimates, which, in combination with most common loss functions, leads to overconfident decisions.

In this work, we propose a CADx system based on Bayesian variational inference [16] for characterization of CRPs. The system offers confidence calibration during the training procedure, in contrast to earlier studies on this topic [11,12], which require an extra post-processing step for the same purpose. Our results demonstrate that the proposed approach is not only competitive in terms of classification accuracy with respect to a Deterministic version of the model, but it is also able to provide reliable confidence measures. To the best of our knowledge, this is the first research study deploying a Bayesian variational inference

framework for characterization of CRPs and expressing confidence-calibrated classification results.

2 Methodology

2.1 Dataset

The experiments conducted in this study are performed on data collected at the Catharina Hospital Eindhoven (CHE) and the Maastricht University Medical Center+ (MUMC+), in the Netherlands, and the Queen Alexandra Hospital (QA) in Portsmouth, United Kingdom. The dataset includes images with White-Light Endoscopy (WLE), Blue Light Imaging (BLI), and Linked Color Imaging (LCI)[1] modalities acquired from CHE and QA. Images collected at the MUMC+ have i-Scan modality in Modes 1, 2, and 3[2]. Several different polyp types are included, namely: HPs, ADs, SSLs and adenocarcinomas. The latter three polyp types are considered pre-malignant, and HPs are categorized as benign. In this study, experiments are carried out to classify CRPs into benign and (pre)malignant classes. To assess the classification performance of the proposed method, a total number of 2,287 images were used, including 1,836 pre-malignant polyps and 451 benign polyps. To evaluate the performance on unseen data, an independent test set is constructed, comprising 86 distinct polyps (258 images), of which 19 are benign and 67 pre-malignant. For each polyp, images from all three modalities are contained in the test set. The remainder of the data is split with 80/20% ratio for training and validation process, respectively, resulting into a training set of 316 benign and 1,308 pre-malignant images, while the validation set has 78 benign and 327 pre-malignant images. To prevent data leakage, all polyps from the same patient are kept together in one set (separation on patient basis).

2.2 Bayesian Neural Networks

Bayesian neural networks (BNNs) offer a probabilistic interpretation of deep learning models by learning a posterior distribution over the weights. As a result, the model will be robust to overfitting and is able to offer uncertainty estimates over the output probabilities.

In Bayesian statistics, network parameters are considered as one large random vector w, where the prior distribution of the weights is expressed as $p(w)$. If $X = \{x_1, ..., x_\beta\}$ denotes a set of training samples and $y = (y_1, ..., y_\beta)^T$ stands for the corresponding class labels, the posterior probability of the weights after observing the dataset is expressed as:

$$p(w|y, X) = \frac{p(y|w, X)p(w)}{\int p(y|w, X)p(w)dw}. \quad (1)$$

[1] EG-760 Colonoscope (Fujifilm® Corporation, Tokyo, Japan).
[2] EC38-i10F2 Colonoscope (PENTAX® Medical, Hoya Corp., Tokyo, Japan).

Classical assumptions on stochastic independence and modeling in deep learning, expresses the probability $p(y|w, X)$ as the product of the neural network outputs for all the training samples. Therefore, the integration over the very high-dimensional space of weights in the denominator of $p(y|w, X)$, makes the posterior generally intractable. Variational inference aims at approximating the posterior $p(w|y, X)$ by a distribution $(q_\Theta(w))$ that is most similar to the posterior distribution obtained by the model. This can be accomplished by Monte Carlo sampling of the posterior of model parameters and minimizing the Kullback-Leibler divergence *(KL-divergence)* between the variational distribution and the posterior: $D_{KL}(q_\Theta(w)||p(w|y, X))$.

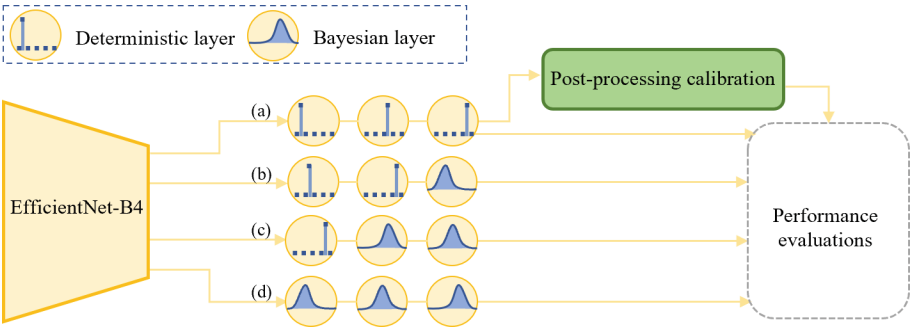

Fig. 1. Deterministic and hybrid Bayesian model architectures. On the left, the EfficientNet-B4 network is used as the base network and is followed by: (a) 3 fully-connected (FC) layers (addressed as the DNN); (b) 2 FC layers and 1 Bayesian layer; (c) 1 FC layer and 2 Bayesian layers; (d) 3 Bayesian layers (addressed as the BNN). The predictions made by the DNN are also passed to a post-processing calibration block. The outputs of all the networks are compared in the performance evaluation block.

2.3 Model Architecture

The block diagram of the employed framework is presented in Fig. 1. We use the EfficientNet-B4 architecture [17], pre-trained on ImageNet [18], as a base network and replace the classification layers with different sets of layers. As shown in Fig. 1 (a), the base architecture with 3 fully-connected (FC) layers serves as the Deterministic neural network (DNN).

The hybrid Bayesian models are achieved by gradually replacing each of the FC output layers with a Bayesian variational layer, which results in architectures (b) to (d) from Fig. 1. A Gaussian distribution is adopted to model the prior distribution of the weights and bias parameters in the Bayesian variational layers. During the training procedure, the aim is to minimize the KL-divergence by making multiple inference passes through the hybrid Bayesian networks. Inference passes are implemented using the Gradient Accumulation technique [19], to reduce the memory consumption. Flipout layers, as introduced by Wen *et al.* [20],

serve as our Bayesian linear layers, due to their ability to decorrelate the gradients within a mini-batch as a result of implicitly sampling pseudo-independent weight perturbations for each data point.

In a first experiment, the calibration performance of the DNN is compared to the hybrid Bayesian model with 3 variational layers (BNN), using the reliability diagrams [21,22] and confidence measures that will be introduced later. A temperature scaling [10] method is also used to calibrate output results of the DNN (Calibrated DNN) to provide a better comparison with the BNN results. In another experiment, we will gradually adapt the DNN architecture towards a BNN by replacing its FC layers with variational layers, and evaluate the generalization and robustness property of the different degree of hybridization.

2.4 Evaluation Metrics

The performance of both approaches is measured and compared by computing various types of calibration error metrics. These metrics are calculated from the reliability diagram [21,22]. A reliability diagram shows accuracy as a function of the predicted confidence of samples, by grouping predictions into bins, based on their predicted confidence. To calculate the metrics predictions are grouped into M bins of size $\frac{1}{M}$, and the accuracy of each bin is computed. Let B_m be the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The accuracy and average confidence within bin B_m is defined as:

$$\mathrm{acc}(\mathrm{B}_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i). \tag{2}$$

$$\mathrm{conf}(\mathrm{B}_m) = \frac{1}{|B_m|} \sum_{i \in B_m} (\hat{p}_i). \tag{3}$$

In the above equation, $\hat{y}_i$ and y_i are the predicted and true class labels, respectively, and $\hat{p}_i$ is the confidence for sample i. One important error metric is Expected Calibration Error (ECE), that is the weighted average of the calibration error across all bins. Moreover, the Maximum Calibration Error (MCE) determines the largest error across the bins. In line with the MCE, we also use the Average Calibration Error (ACE), that determines the average error across the non-empty bins (M^+). Finally, the Over-Confidence Error (OE) is specified as the weighted average of the errors across bins where confidence exceeds accuracy. These errors are covered by the following equations:

$$ECE = \sum_{m=1}^{M} \frac{B_m}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|. \tag{4}$$

$$MCE = \max_{m \in 1,\ldots,M} \mathrm{acc}(B_m) - \mathrm{conf}(B_m)|. \tag{5}$$

$$ACE = \frac{1}{M^+} \sum_{m=1}^{M} \mathrm{acc}(B_m) - \mathrm{conf}(B_m)|. \tag{6}$$

$$OE = \sum_{m=1}^{M} \frac{B_m}{n} [\mathrm{conf}(B_m). \max(\mathrm{conf}(B_m) - \mathrm{acc}(B_m), 0)]. \tag{7}$$

3 Results

3.1 Experimental Setting

We evaluate our Bayesian network (BNN) in multiple experiments using the introduced CRP dataset and compare the obtained results with the uncalibrated and calibrated deterministic versions of the network (DNN). Input images for both networks are resized to 256×256 pixels, while compatibility of the dataset with ImageNet pre-trained networks is ensured by channel-wisely subtracting the mean and dividing by the standard deviation of ImageNet data. For an improved generalization, data augmentation is applied using the following transformations: horizontal and vertical flipping, rotation, Gaussian blurring, contrast/saturation/brightness enhancements, random affine, and perspective transforms. For the optimization, we use the Adam optimizer with a learning rate of 10^{-5}, (β_1,β_2)=(0.9,0.999). Due to the class imbalance, an independent batch generator is used to ensure that each of the classes is represented during training.

A mini-batch size of 16 (7 benign/9 pre-malignant) images is used and a the data is shuffled after each epoch. The training iteration ends when all benign images are seen once by the network. The experiments are implemented using the PyTorch framework and executed on a GeForce RTX 2080 Ti. To train the different hybrid BNN model variations, we perform multiple stochastic forward passes on the final (1–3) FC variational layers with Monte Carlo sampling on the weight posterior distributions. In our experiments, for a better generalization of the model, 10 forward passes provide reliable estimates. Subsequently, the predictive mean is obtained by averaging the confidence estimates from inference passes.

3.2 Calibration-performance Assessment

In order to verify the ability of the BNN to provide reliable confidence measures, we visualize and compare the calibration performance of the BNN with the DNN by using reliability diagrams. In these diagrams, the degree of miscalibration can be assessed by the gap between the plotted accuracy and the ideal diagonal. A high calibration performance can be achieved when the bin accuracy aligns closely with the ideal diagonal (expected accuracy). Figure 2 shows the reliability diagrams of the BNN, DNN, and the calibrated DNN with temperature scaling. The Green/Red bars indicate the Under/Over-Confidence, respectively. It can be observed that the BNN is better calibrated, as the achieved accuracies of the bins better approximate the expected accuracies (i.e. the bars align closer along the ideal diagonal). For the DNN and the calibrated DNN, the reliability diagrams show larger gaps between the achieved and the expected accuracies, especially for the higher confidence values.

Using the introduced calibration measures, a more quantitative comparison of the calibration performance of the three networks is achieved. Table 1 demonstrates a lower MCE of 0.2539 for the BNN compared to 0.2654 for the DNN,

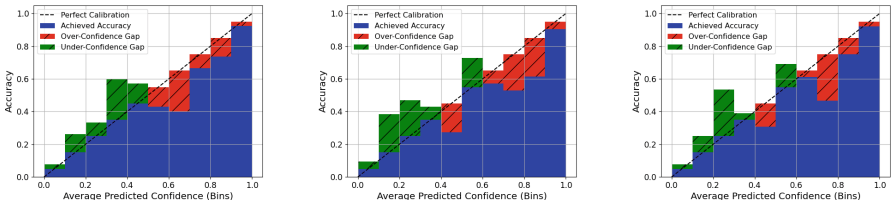

Fig. 2. Reliability diagrams for (left) the BNN, (center) the DNN, and (right) the calibrated DNN with temperature scaling.

Table 1. Calibration performance comparison for various experimented networks.

Network	ECE	MCE	ACE	OE
Bayesian network (BNN)	0.1699	0.2539	0.1296	0.0753
Deterministic network (DNN)	0.2255	0.2910	0.1721	0.0978
Calibrated deterministic network (Cal. DNN)	0.1870	0.2654	0.1388	0.0758

and 0.2910 for the calibrated DNN. In addition, the BNN network is able to achieve lower error rates using other calibration measures compared to the DNN as well as the calibrated DNN with the temperature scaling technique.

3.3 Model Performance Comparison

We evaluate our Bayesian and Deterministic models on the CRP dataset. In Table 2, a comparison of the obtained classification accuracy, Sensitivity, Specificity, area under curve (AUC), negative predictive value (NPV), and positive predictive value (PPV) on the test dataset are presented for each of the networks. The results show a very similar overall performance between the two networks, with most of the employed metrics exhibiting only a negligible difference.

3.4 Generalization and Robustness to Over-Fitting Assessment

In another experiment, we investigate the effect of increasing the Bayesian level of the deterministic network, by gradually replacing each of the FC layers of the Deterministic network by a Flipout variational layer, and obtain a network with increasingly more Bayesian layers (see Fig. 1 (b)–(d)), and finally obtain a network with 3 Bayesian FC layers. We have compared various versions of this

Table 2. Bayesian vs. Deterministic neural network performance assessment.

Network	Phase	Acc.	Sens.	Spec.	AUC	NPV	PPV
Bayesian network (BNN)	Test	84.10	90.55	61.40	0.89	64.81	89.22
Deterministic network (DNN)	Test	84.88	90.05	66.67	0.89	65.52	90.50

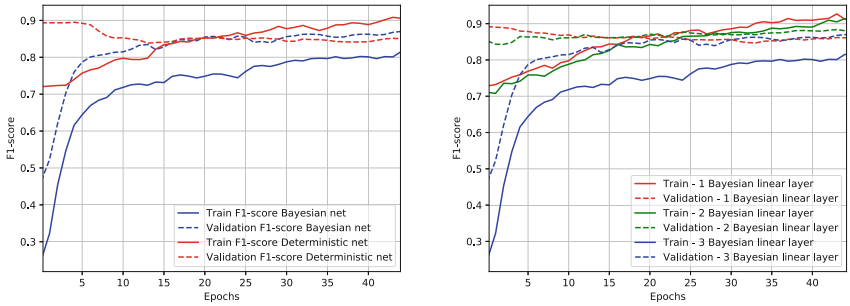

Fig. 3. Comparison of the F1-score during training and validation phase for (left) the Bayesian (BNN) and the Deterministic (DNN) networks, and (right) the different levels of hybridization for the Bayesian network.

hybrid BNN, in terms of F1-score of the training and validation phase and the result is available in Fig. 3 at the right. As demonstrated by the plots, the DNN shows a drop in validation F1-score as the training process advances, which shows that the model is over-fitting on the training data. On the other hand, the hybrid network with 3 Bayesian layers obtains a higher validation F1-score regarding its training F1-score and, therefore, offers a better generalized performance. Another important observation is that both networks with 1 and 2 Bayesian layers have a similar performance as the DNN, while experiencing the over-fitting problem to a lower degree. This possibly indicates the insufficiency of the Bayesian effect of the networks. It can be noticed that the network with 3 Bayesian layers (BNN) is capable of achieving comparable validation F1-score and expresses a better robustness towards over-fitting.

4 Discussion and Conclusion

Both an optimal clinical workflow integration and the physician-AI collaboration necessitate a reliable CADx system that is capable of capturing an accurate and well-calibrated classification confidence. In this regard, we incorporate Bayesian variational inference and investigate the performance of a hybrid Bayesian neural network architecture for the characterization of CRPs. The presented quantitative and qualitative results demonstrate that the BNN is capable of expressing reliable uncertainty measures and better calibrated classification confidence compared to a peer Deterministic network. Furthermore, the hybrid BNN approach is able to outperform a temperature-scaling calibrated DNN and provides lower calibration errors. Moreover, it alleviates the need for an additional calibration data set. A further hybridization experiment, based on replacing output layers with Bayesian variational layers, shows that the best performance is obtained by using 3 Bayesian layers. The better generalization property and being less prone to over-fitting, makes BNNs a suitable choice for small datasets. However, dealing with imbalanced classes can be an important challenge that should be

further investigated. Bayesian networks are generally slower and have high memory consumption during training due to the required sampling for inference, and are heavily reliable on the prior distribution initialization for achieving a good predictive accuracy. This opens an interesting direction for future work, investigating whether assigning class-specific prior distributions can be beneficial for classes with less data availability.

References

1. Sung, H.,et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clinic. **71**(3), 209–249 (2021)
2. Shin, Y., Qadir, H.A., Aabakken, L., Bergsland, J., Balasingham, I.: Automatic colon polyp detection using region based deep CNN and post learning approaches. IEEE Access **6**, 40950–40962 (2018)
3. Meng, J., et al.: Automatic detection and segmentation of adenomatous colorectal polyps during colonoscopy using Mask R-CNN. Open Life Sci. **15**(1), 588–596 (2020)
4. Zhang, R., et al.: Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. IEEE J. Biomed. Health Inf. **21**(1), 41–47 (2016)
5. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. **60**, 101619 (2020)
6. Alam, S., Tomar, N.K., Thakur, A., Jha, D., Rauniyar, A.: Automatic polyp segmentation using u-net-resnet50. arXiv preprint arXiv:2012.15247 (2020)
7. Weigt, J., et al.: Performance of a new integrated computer-assisted system (CADe/CADx) for detection and characterization of colorectal neoplasia. Endoscopy. **54**(02), 180–184 (2022)
8. Usami, H., et al.: Colorectal polyp classification based on latent sharing features domain from multiple endoscopy images. Proc. Comput. Sci. **176**, 2507–2514 (2020)
9. Fonollà, R., et al.: A CNN CADx system for multimodal classification of colorectal polyps combining WL, BLI, and LCI modalities. Appl. Sci. **10**(15), 5040 (2020)
10. Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q.: On calibration of modern neural networks. In: PLMR, International Conference on Machine Learning, pp. 1321–1330 (2017)
11. Kusters, K.C., et al.: Colorectal polyp classification using confidence-calibrated convolutional neural networks. In: SPIE, Medical Imaging 2022: Computer-Aided Diagnosis, vol. 12033, pp. 442–454(2022)
12. Carneiro, G., Pu, L.Z.C.T., Singh, R., Burt, A.: Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med. Image Anal. **62**, 101653 (2020)
13. Krishnan, R., Subedar, M. and Tickoo, O.: BAR: Bayesian activity recognition using variational inference. arXiv preprint arXiv:1811.03305 (2018)
14. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)

15. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)
16. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning, PMLR (2015)
17. Tan, M. and Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
18. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Computer Vision and Pattern Recognition Conference, CVPR (2009)
19. Nazarovs, J., Mehta, R.R., Lokhande, V.S., Singh, V.: Graph reparameterizations for enabling 1000+ Monte Carlo iterations in Bayesian deep neural networks. In: Uncertainty in Artificial Intelligence, pp. 118–128. PMLR (2021)
20. Wen, Y., Vicol, P., Ba, J., Tran, D., Grosse, R.: Flipout: efficient pseudo-independent weight perturbations on mini-batches. arXiv preprint arXiv:1803.04386 (2018)
21. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. J. R. Statist. Soc. Ser. D (The Statist.) **32**(1–2), 12–22 (1983)
22. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: 22nd International Conference on Machine Learning (2005)