IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2022, accepted 11 July 2022, date of publication 18 July 2022, date of current version 26 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3192024

RESEARCH ARTICLE

Uncertainty-Aware Deep Learning Methods for Robust Diabetic Retinopathy Classification

JOEL JASKARI[1], JAAKKO SAHLSTEN[1], THEODOROS DAMOULAS[2,3], JEREMIAS KNOBLAUCH[4], SIMO SÄRKKÄ[5], (Senior Member, IEEE), LEO KÄRKKÄINEN[5], KUSTAA HIETALA[6], AND KIMMO K. KASKI[1,2]

[1]Department of Computer Science, Aalto University, 00076 Aalto, Finland
[2]The Alan Turing Institute, London NW1 2DB, U.K.
[3]Department of Computer Science and the Department of Statistics, The University of Warwick, Coventry CV4 7AL, U.K.
[4]Department of Statistical Science, University College London, London WC1E 6BT, U.K.
[5]Department of Electrical Engineering and Automation, Aalto University, 00076 Aalto, Finland
[6]Central Finland Health Care District, 40620 Jyväskylä, Finland

Corresponding author: Kimmo K. Kaski (kimmo.kaski@aalto.fi)

The work of Joel Jaskari, Jaakko Sahlsten, and Kimmo K. Kaski was supported in part by the Academy of Finland under Project 345449.

ABSTRACT Automatic classification of diabetic retinopathy from retinal images has been increasingly studied using deep neural networks with impressive results. However, there is clinical need for estimating uncertainty in the classifications, a shortcoming of modern neural networks. Recently, approximate Bayesian neural networks (BNNs) have been proposed for this task, but previous studies have only considered the binary referable/non-referable diabetic retinopathy classification applied to benchmark datasets. We present novel results for 9 BNNs by systematically investigating a clinical dataset and 5-class classification scheme, together with benchmark datasets and binary classification scheme. Moreover, we derive a connection between entropy-based uncertainty measure and classifier risk, from which we develop a novel uncertainty measure. We observe that the previously proposed entropy-based uncertainty measure improves performance on the clinical dataset for the binary classification scheme, but not to such an extent as on the benchmark datasets. It improves performance in the clinical 5-class classification scheme for the benchmark datasets, but not for the clinical dataset. Our novel uncertainty measure generalizes to the clinical dataset and to one benchmark dataset. Our findings suggest that BNNs can be utilized for uncertainty estimation in classifying diabetic retinopathy on clinical data, though proper uncertainty measures are needed to optimize the desired performance measure. In addition, methods developed for benchmark datasets might not generalize to clinical datasets.

INDEX TERMS Approximate Bayesian neural networks, deep learning, diabetic retinopathy, reject option classification, uncertainty estimation.

I. INTRODUCTION

Deep neural networks have achieved impressive results in a wide variety of problems, ranging from large scale image classification [1], to natural language understanding [2], and to medical image segmentation [3]. However, the standard methods have been found to produce over-confident predictions, meaning that they are poorly calibrated [4]. In classification tasks, a poorly calibrated network can place high probability on one of the classes, even when the predicted class is incorrect, whereas a well calibrated classifier would place less probability mass on uncertain classes. In the medical domain, the issue of uncertainty estimation is especially important for having trust to confident model predictions in screening automation and referring uncertain cases for intervention by medical experts. In this work we refer to the classifiers that can indicate their uncertainty as robust classifiers.

Over the past few years, the automatic classification of diabetic retinopathy by using deep neural networks has been

The associate editor coordinating the review of this manuscript and approving it for publication was Nagarajan Raghavan.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

under growing interest [5]–[8]. More recently, the focus of attention has turned on developing robust deep learning methods for the classification task, most commonly using the approximate Bayesian deep learning approach that approximates the Bayesian neural network (BNN) posterior distribution in a computationally scalable manner. The previous works have considered a variety of aspects from studying the benefits of uncertainty estimates [9] to algorithmic development of robust methods [10], [11]. Although the variety of different studied algorithms has been diverse [9], [12], the used datasets have so far been benchmark datasets. This leaves open the question of whether the algorithms generalize to real clinical data. In addition, these recent works have mainly focused on the classification of diabetic retinopathy using binary classification schemes, i.e. "referable vs. non-referable" (RDR) or "healthy vs. any" diabetic retinopathy. However, in clinically oriented approaches there has been a shift towards the 5-class proposed international diabetic retinopathy classification system (PIRC) [8], [13], [14]. In order for these approaches to have clinical use, it is of paramount importance that the algorithms generalize to clinical datasets and to clinical diabetic retinopathy grading systems, which is the novel scope we adopt in this study.

The common aspect among the works focusing on robust methods is the use of uncertainty information to simulate a referral process, introduced by [9]. Each prediction is associated with an uncertainty estimate, and the least certain predictions are referred to experts, while the more certain predictions are used for evaluation. This process mimics a situation in which the automated system asks human intervention for uncertain cases, i.e. refers them to an expert. In practice, the holdout test set predictions are ordered according to their uncertainty and several referral levels are defined corresponding to a percentile of referred examples, i.e. 10% referral level means that 10% of the most uncertain examples are left out of the evaluation.

In [9], a Monte Carlo (MC) dropout neural network was used for two binary classification tasks: classification of any diabetic retinopathy and RDR. They used the EyePACS dataset [15] for training and testing and evaluated the out-of-distribution performance with the Messidor dataset [16] and observed improved robustness in comparison to a baseline standard neural network. Reference [12] conducted a more methodologically extended study for the classification of RDR. The EyePACS dataset was used for training and testing, and the out-of-distribution performance was evaluated with the APTOS [17]. The work examined a number of approximate Bayesian deep learning methods, such as the MC dropout, Mean Field Variational Inference (MFVI), deep ensembles, and MC dropout ensemble. They observed that the approximate Bayesian methods outperformed the standard neural network in all the experiments where network uncertainty was utilized. Reference [10] proposed the Radial BNN method and benchmarked it against MFVI, MC dropout, and deep ensembles using the EyePACS dataset, as both the training and test sets, for the RDR classification task. A single Radial BNN was found to outperform the MC dropout and MFVI, but not the deep ensemble. Overall, an ensemble of Radial BNN's turned out to outperform all other methods at all referral levels. In [18] the RDR task was examined using the EyePACS dataset for training and testing, and the out-of-distribution performance was evaluated with the APTOS, similar to [12]. They also studied how the robust deep learning methods generalize when no images of severe and even severer proliferative diabetic retinopathy cases are used in the training. The considered approximate deep learning methods were MFVI, Radial BNN, Function-Space Variational Inference, MC dropout, and Rank-1 Parameterized BNN, and ensembles of them. Also the deep ensembles were used. It was found that the MC dropout ensemble performed the best for the within-distribution and the MFVI ensemble for the out-of-distribution experiments.

In this study, our objective is to analyze robust neural networks, i.e. networks that are inherently well calibrated, for the task of diabetic retinopathy classification using the RDR and PIRC classification schemes. For this reason we leave out the analysis of *post-hoc* calibration methods, such as the test-time augmentation introduced in [11] and [19], neural network softmax temperature scaling, and probability binning strategies [4]. We also omit methods based on standard neural networks, such as the DR|GRADUATE [20], that is trained to predict the PIRC label and an uncertainty score, using a specific ordinal regression setting.

Many measures of uncertainty have been used in the previous works. In [9], the standard deviation of the output of the BNN and the entropy of the posterior predictive distribution were considered as measures of uncertainty and were found to perform similarly. In [12], [19], and [18] the entropy was also selected as the measure of uncertainty. On the other hand, the mutual information between the parameters of the model and the output was considered as the measure of uncertainty in [10].

In the present work, we explore the benefits of uncertainty estimates for a clinical dataset of a Finnish hospital on both the binary RDR and the 5-class PIRC classification schemes. We investigate 9 different approximate Bayesian methods to extensively analyze recently proposed methods. In addition, we study the out-of-distribution performance using three benchmark datasets: the EyePACS [15], the Messidor-2 [16], [21], and the APTOS [17]. We observe that the entropy uncertainty estimates improve the area under the receiver operating characteristic curve (AUC) performance in the binary RDR system, but in the case of the 5-class PIRC system, the quadratic weighted Cohen's kappa (QWK) performance only improves across the benchmark datasets. For the clinical dataset and PIRC system, we observe less robust classifier performance using entropy-based uncertainty. To improve the quality of the uncertainty estimates, we additionally propose a novel classifier risk based uncertainty measure, which improves the within-distribution uncertainty performance on both the Finnish hospital dataset and the EyePACS dataset. As far as we know, clinical diabetic retinopathy severity schemes and clinical workflow datasets have not been studied before using robust neural networks.

II. METHODS

A. DATASETS

All our datasets consist of color images of the human retina and are graded using the following 5-class PIRC system for the severity scheme of diabetic retinopathy: *no diabetic retinopathy* (class 0), *mild diabetic retinopathy* (class 1), *moderate diabetic retinopathy* (class 2), *severe diabetic retinopathy* (class 3), and *proliferative diabetic retinopathy* (class 4). From the PIRC system, we derive the binary *referable/non-referable diabetic retinopathy* (RDR) classification, which is defined as the union of no diabetic retinopathy or mild diabetic retinopathy (≤ 1) and referable as retinopathy worse than or equal to moderate diabetic retinopathy (≥ 2). We chose to include the binary RDR system for comparison with the previous works.

We use the following four datasets for our experiments: the EyePACS [15], KSSHP [22], Messidor-2 [16], [21], and APTOS [17]. The EyePACS and APTOS datasets were introduced for two different Kaggle competitions of diabetic retinopathy detection. The Messidor-2 is a common benchmark dataset that was introduced for research in computer-assisted diagnosis of diabetic retinopathy. The EyePACS, APTOS, and Messidor-2 datasets have been widely used in literature for training and analyzing robust neural networks. In order to examine if the results generalize to clinical hospital datasets, we use the non-public KSSHP dataset. The KSSHP set was collected from clinical workflow data from the Central Finland Health Care District. Central Finland Health Care District approved the research permit regarding the KSSHP data in January 2021. The research was conducted as a register based research, and thus did not need ethical board evaluation according to EU General Data Protection Regulation (GDPR) and the Finnish law [23], [24]. All the used datasets originate from different countries: The EyePACS from USA, KSSHP from Finland, APTOS from India, and Messidor-2 from France, allowing for extensive analysis for the generalization of the models under distribution shift by country.

The division of data to training, validation, and test sets were performed for EyePACS and KSSHP image datasets. For the EyePACS dataset, the official "Train" set was used for training, the "Public" set was used for validation, and the "Private" set was used as the test set. These training, validation, and test splits have also been used in [18] and in the Tensorflow Datasets Python package. For the KSSHP dataset, we used 70%, 10%, and 20% of images for the training, validation, and test sets, respectively. For the splits, we used stratified pseudo random sampling to preserve PIRC class distribution of the original set, but taking into account that images from the same patient cannot reside in more than one set. Due to the low amount of images in the APTOS and Messidor-2 datasets, 3662 and 1744 respectively, they were used only as out-of-distribution test sets. Complete description of the training, validation, and test sets are described in Table 1 and the test set class distributions in Table 2. The images were resized into standard size 512 x 512 and, to reduce known variability, preprocessed with steps described in the Supplementary Section 1.

B. APPROXIMATE BAYESIAN DEEP LEARNING

The approximate Bayesian deep learning models take the uncertainty in account by computing the following posterior predictive distribution:

$$p(y \mid \boldsymbol{x}, D) = \int_{\boldsymbol{\theta} \in \Theta} p(y \mid \boldsymbol{x}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}. \tag{1}$$

Here $\boldsymbol{x}$ and y denote the image and target, respectively, D the training data, and $\boldsymbol{\theta}$ the model parameters. The predictions are weighted averages using the posterior distribution $p(\boldsymbol{\theta} \mid D)$ [25], [26]. For the standard neural networks, the maximum likelihood (ML) or maximum a posteriori (MAP) estimates are typically used, which completely ignore the uncertainty in the parameters.

The exact solution for Equation (1) is intractable for deep neural networks, and Markov Chain Monte Carlo is prohibitively expensive for real world scenarios. Thus approximations need to be utilized. We selected the deep ensemble, MC dropout, MFVI, Generalized Variational Inference (GVI), and Radial BNN as our approximate Bayesian methods. The posterior predictive distribution can be inexpensively approximated with these methods using an approximate posterior distribution $q(\boldsymbol{\theta})$ and Monte Carlo integation with N samples:

$$p(y \mid \boldsymbol{x}, D) \approx \frac{1}{N} \sum_{i=1}^{N} p(y \mid \boldsymbol{x}, \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i \sim q(\boldsymbol{\theta}). \tag{2}$$

C. DEEP ENSEMBLES

The deep ensemble [27] is a collection of multiple ML or MAP neural network models. The models are trained using different initializations in order to produce a diverse set of models. The predictions of the models are averaged to produce the prediction of the ensemble model, corresponding to heuristically setting the posterior as uniform distribution over the set if models in Equation (2).

D. MONTE CARLO DROPOUT - MC DROPOUT

The Monte Carlo dropout method introduced in [28] allows approximate Bayesian inference during test-time for networks that have been trained using the dropout regularization method. The dropout works by sampling a binary mask $\boldsymbol{r} = [r_1, \ldots, r_d]^\top$ from a Bernoulli distribution $r_i \sim \mathrm{Bern}(p)$ and masks the activations h of a layer by computing the Hadamard product $r \odot h$ [29], which is equivalent to masking rows of the weight matrix and elements of the bias vector [28]. The posterior predictive distribution is then computed using the dropout distribution in Equation (2).

TABLE 1. Number of images in each subset for each dataset.

Subset	EyePACS [15]	KSSHP [22]	APTOS [17]	Messidor-2 [21]
Train	35125 (39.6%)	39482 (70.0%)	-	-
Validation	10906 (12.3%)	5652 (10.0%)	-	-
Test	42669 (48.1%)	11285 (20.0%)	3662 (100%)	1744 (100%)
Total	88700	56419	3662	1744

TABLE 2. Class distribution for PIRC and RDR classification schemes of the test sets.

Class	EyePACS [15]	KSSHP [22]	APTOS [17]	Messidor-2 [21]
PIRC 0	31403 (73.6%)	7723 (68.4%)	1805 (49.3%)	1017 (58.3%)
PIRC 1	3042 (7.1%)	2431 (21.5%)	370 (10.1%)	270 (15.5%)
PIRC 2	6281 (14.7%)	930 (8.2%)	999 (27.3%)	347 (19.9%)
PIRC 3	977 (2.3%)	177 (1.6%)	193 (5.3%)	75 (4.3%)
PIRC 4	966 (2.3%)	24 (0.2%)	295 (8.1%)	35 (2%)
RDR 0	34445 (80.7%)	10154 (90.0%)	2175 (59.4%)	1279 (73.3%)
RDR 1	8224 (19.3%)	1131 (10.0%)	1487 (40.6%)	465 (26.7%)

E. MEAN FIELD VARIATIONAL INFERENCE - MFVI

The Mean Field variational approximations assume that the neural network parameters are independent and typically that the approximate posterior and prior are Gaussians [10], [12], [30]. The evidence lower-bound (ELBO) can then be maximized, which provides a lower bound for the posterior probability, and for the diagonal multivariate Gaussian case, the equations become simple [30]:

$$\begin{aligned}\mathcal{L}_{ELBO}(\mathcal{D},\boldsymbol{\theta}) &= \mathbb{E}_{q(\boldsymbol{\theta})}[\log(p(\mathcal{D}\mid\boldsymbol{\theta}))] \\ &\quad - D_{KL}[q(\boldsymbol{\theta})\mid\mid p(\boldsymbol{\theta})], \\ &= \mathbb{E}_{q(\boldsymbol{\theta})}[\log(p(\mathcal{D}\mid\boldsymbol{\theta}))] \\ &\quad - \sum_{j=1}^{J}\log\frac{s_j}{\sigma_j} \\ &\quad + \frac{1}{2s_j^2}[(\mu_j-m_j)^2+\sigma_j^2-s_j^2]. \qquad (3)\end{aligned}$$

Here the $KL[\cdot\mid\mid\cdot]$ denotes the Kullback-Leibler (KL) divergence, the prior is $\mathcal{N}(m_j, s_j^2)$, and the variational posterior is $\mathcal{N}(\mu_j,\sigma_j^2)$ for a parameter θ_j. The remaining expected log-likelihood term is computed using Monte Carlo integration and the reparametrization trick [31]. Similar to the MC dropout, the posterior predictive distribution is computed using samples from the fitted approximate posterior distribution. We used multivariate standard normal distribution as the prior in all experiments.

F. GENERALIZED VARIATIONAL INFERENCE - GVI

In [32], a novel optimization-centric view to posterior inference is proposed. The problem of finding posterior distributions is viewed in terms of the so-called "Rule-of-Three", which separates the loss, divergence, and the space of feasible solutions as different aspects of the optimization procedure. Different configurations result in different posterior inference methods, for example the standard Bayesian posterior inference and variational inference. This view allows for principled use of divergences, which are robust to the mis-specification of the prior. Since the diagonal standard normal is typically chosen for computational convenience (rather than to incorporate any prior knowledge about good values of the neural network weights) [32], this approach is promising. We select the robust divergence as Rényi's α-Divergence ($D_{AR}^{\alpha}[\,\cdot\,\mid\mid\,\cdot\,]$), with $\alpha = 0.5$. As the loss function, we use the negative log-likelihood, and as the space of feasible solutions the mean field normal posteriors and priors, where the prior was selected as the diagonal standard normal. The total minimization objective is then:

$$\begin{aligned}\mathcal{L}_{GVI} &= -\mathbb{E}_{q(\boldsymbol{\theta})}[\log(p(\mathcal{D}\mid\boldsymbol{\theta}))] \\ &\quad + D_{AR}^{\alpha}[q(\boldsymbol{\theta})\mid\mid p(\boldsymbol{\theta})], \qquad (4)\end{aligned}$$

$$D_{AR}^{\alpha}[q(\boldsymbol{\theta})\mid\mid p(\boldsymbol{\theta})] = \frac{1}{\alpha(1-\alpha)}\log\int q(\boldsymbol{\theta})^{\alpha}p(\boldsymbol{\theta})^{(1-\alpha)}d\boldsymbol{\theta}. \qquad (5)$$

G. RADIAL BAYESIAN NEURAL NETWORKS - RADIAL BNN

The multivariate normal distribution has the so-called "soap bubble" pathology in high dimensions, meaning that most of the probability mass is concentrated on thin shell far from the mean, resulting in samples with a high norm. In [10], the high norm is proposed to be the problem in training deep neural networks using the MFVI approach with Gaussian posteriors. To address this issue, they propose a novel "Radial BNN" posterior distribution. The proposed distribution is constructed such that the samples have the same expected norm as univariate standard normal distribution, regardless of the dimension. The sampling process is defined as follows for a single weight vector:

$$\boldsymbol{w} = \boldsymbol{\mu} + \boldsymbol{\sigma}\odot\hat{\boldsymbol{\epsilon}}\cdot r, \qquad (6)$$

$$\hat{\boldsymbol{\epsilon}} = \frac{\boldsymbol{\epsilon}}{\mid\mid\boldsymbol{\epsilon}\mid\mid_2}, \qquad (7)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I), \qquad (8)$$

$$r \sim \mathcal{N}(0, 1). \qquad (9)$$

The resulting random variable $\boldsymbol{w}$ avoids the soap bubble pathology, however, it has no closed form probability density function or KL-divergence. The authors observe that a stochastic estimate of the KL-divergence can be computed similar to [26], allowing for optimizing the ELBO up to a constant. Multivariate standard normal distribution was selected as the prior, similar to MFVI.

H. NEURAL NETWORK ARCHITECTURE AND IMPLEMENTATION

For all our experiments, we used a VGG16 type network as the base architecture, similar to most of the previous studies [9], [10], [12]. The VGG16 architecture was chosen to be identical to the one used by [10], as their study demonstrated high performance for approximate Bayesian neural networks on the RDR task. This network consists of five ''blocks'' with the first two blocks containing two convolutional layers each, and all the later blocks have three convolutional layers in each block. All the convolutional layers have a kernel size of 3×3, and each of them is followed by a Leaky ReLU activation with a negative slope of $\alpha = 0.2$. A max pooling operation with kernel size and stride of 2×2 is applied as the last operation of each block, with the exception of the final block. The output of the final block is globally average pooled and max pooled over the spatial dimensions, and their output vector representations are concatenated. The final fully-connected layer has one neuron with a sigmoid activation for the binary RDR classification task, and five neurons with a softmax activation for the 5-class PIRC task. The MAP and MC dropout networks had 64 channels for the first convolution and the channel count was doubled after each block. The MFVI, GVI, and Radial networks had 46 channels that were doubled after each block to keep the total number of trainable parameters between models comparable, similarly to [10]. All the models had approximately 15 million parameters. The ensemble models consisted of three individual models that were trained with different random seeds. The architecture is visually illustrated in Table 3.

TABLE 3. Architecture of the VGG network. *k* denotes the kernel size and *s* the stride used for the convolutional layer. *C* = 64 for MAP, MC dropout, deep ensemble, and MC dropout ensemble, and *C* = 46 for MFVI, GVI, Radial, MFVI ensemble, GVI ensemble, and radial ensemble.

Block	Layers	Output Size
Input	-	$512 \times 512 \times 3$
1	Convolution, $k = 3$, $s = 2$ Leaky ReLU	$256 \times 256 \times C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$256 \times 256 \times C$
	Max Pool	$128 \times 128 \times C$
2	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$128 \times 128 \times 2C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$128 \times 128 \times 2C$
	Max Pool	$64 \times 64 \times 2C$
3	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$64 \times 64 \times 4C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$64 \times 64 \times 4C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$64 \times 64 \times 4C$
	Max Pool	$32 \times 32 \times 4C$
4	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$32 \times 32 \times 8C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$32 \times 32 \times 8C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$32 \times 32 \times 8C$
	Max Pool	$16 \times 16 \times 8C$
5	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$16 \times 16 \times 8C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$16 \times 16 \times 8C$
	Convolution, $k = 3$, $s = 1$ Leaky ReLU	$16 \times 16 \times 8C$
	Concat $\begin{pmatrix}\text{Global Avg. Pool}\\ \text{Global Max Pool}\end{pmatrix}$	$16C$
Classifier	Fully-connected Sigmoid if RDR Softmax if PIRC	1 if RDR 5 if PIRC

Our baseline approach is a standard neural network trained with dropout and L2 weight regularization. The L2 weight regularization is equivalent to the MAP estimation if a fully factorized normal prior is used on the network parameters. The MAP network, MC dropout, deep ensemble, and MC dropout ensemble models were trained using the negative log-likelihood of a Bernoulli distribution for RDR and the negative log-likelihood of a categorical distribution for PIRC. In addition, these models were regularized using the L2 weight regularization. The MFVI, Radial, MFVI ensemble, and Radial ensemble models were trained to maximize the ELBO in Equation (3), and the GVI and GVI ensemble model to minimize the loss in Equation (4). The $\log p(\mathcal{D} \mid \boldsymbol{\theta})$ in Equations (3) and (4) was also the log-likelihood of a Bernoulli distribution for RDR and the log-likelihood of a categorical distribution for PIRC. We did not employ over-sampling of the minority classes or weighting of the log-likelihood term, as previous studies suggest that the standard negative log-likelihood loss can reach equal or better performance, e.g. when comparing the results by [10] and [12]. Mini-batch sizes, optimizer settings, and the used data augmentation scheme are detailed in the Supplementary Section 1.

I. UNCERTAINITY ESTIMATION

Many uncertainity estimates have been used in previous works. The entropy of the posterior predictive distribution has been a typical choice, and has been observed to work well for the binary RDR classification. However, for the 5-class PIRC classification, we observe that the entropy does not systematically improve the referral process QWK on the clinical dataset. We seek to identify alternative measures of uncertainty, which would give more informed rejection rules.

We measure the performance of our 5-class PIRC classifiers using the quadratic weighted Cohen's kappa [33]:

$$\kappa_{QW}(C) = 1 - \frac{\sum_{i=1}^{M} \sum_{j=1}^{M} (i-j)^2 C_{i,j}}{\sum_{i=1}^{M} \sum_{j=1}^{M} (i-j)^2 E_{i,j}}, \tag{10}$$

$$E_{i,j} = \frac{1}{N} \sum_{a=1}^{M} C_{i,a} \sum_{b=1}^{M} C_{b,j}. \tag{11}$$

The C is the confusion matrix where element $C_{i,j}$ is the number of cases where the predicted label is i and the true label is j, and E is the expected agreement matrix. For a perfect classifier the confusion matrix is diagonal and thus the numerator term is zero, which results in a QWK value of 1. The QWK weights the misclassifications with squared distance of the numerical class labels, as well as with the expected agreement.

The referral of uncertain examples is essentially a case of *reject option classification* [34]. In reject option classification, the risk of a classification decision is the measure of uncertainty, and when the risk exceeds a certain threshold, the prediction is discarded [35]. When the classifier risk is the minimum risk over the decisions, and the error is defined using the 0/1 loss, a classic result is that the prediction is rejected if $\max_i p(y = i \mid x) < \tau$ for some threshold τ [34], [36].

Instead of the minimum risk estimator, we choose to use an average risk, similar to the expected risk of classifier in [37]. The average risk view reveals an interesting connection between the classifier risk analysis and the entropy referral process, and thus helps in the analysis of constructing alternative uncertainty measures. The expected risk of a classifier that estimates the likelihood of discrete labels y given $\boldsymbol{x}$ is defined as [37]

$$\mathcal{I} = \int_{\boldsymbol{x}} \sum_{y} \mathcal{L}(p(y \mid \boldsymbol{x}), y) p(\boldsymbol{x}, y) d\boldsymbol{x}. \tag{12}$$

The $\mathcal{L}(\cdot, \cdot)$ is the loss function associated with a certain prediction and label combination. The risk associated with using the classifier for a certain input $\boldsymbol{x}$ can be derived by leaving the marginalization over $\boldsymbol{x}$ out, which is the expected (over y) conditional (on $\boldsymbol{x}$) risk:

$$\mathcal{R}(\boldsymbol{x}) = \sum_{y} \mathcal{L}(p(y \mid \boldsymbol{x}), y) p(y \mid \boldsymbol{x}). \tag{13}$$

For a given classifier, the risk is now completely defined by the choice of the loss function. Indeed, the expected conditional risk can be used for any performance measure we want to optimize by choosing a loss function that corresponds to the performance measure.

The negative log-likelihood of a target label c given a categorical distribution $p(y \mid \boldsymbol{x})$ with M classes is:

$$\begin{aligned} \mathcal{L}_{NLL}(p(y \mid \boldsymbol{x}), c) &= -\sum_{j=1}^{M} [c = j] \log(p(y = j \mid \boldsymbol{x})), \\ &= -\log(p(y = c \mid \boldsymbol{x})). \end{aligned} \tag{14}$$

When we plug this to the expected conditional risk, we obtain the entropy of the posterior predictive distribution:

$$\begin{aligned} \mathcal{R}(\boldsymbol{x}) &= \sum_{i=1}^{M} \mathcal{L}_{NLL}(p(y \mid \boldsymbol{x}), i) p(y = i \mid \boldsymbol{x}), \\ &= \sum_{i=1}^{M} -\log(p(y = i \mid \boldsymbol{x})) p(y = i \mid \boldsymbol{x}). \end{aligned} \tag{15}$$

Thus the negative log-likelihood induces a risk measure that is the entropy of the posterior predictive distribution.

In order to apply the methodology to the QWK, we need a loss function that directly reflects it. For a single prediction, the numerator and denominator terms in Equation (10) will be the same and thus the κ_{QWK} will always be zero. Thus we need the confusion matrix to be a non single entry matrix. For this purpose, we utilize an initial estimate of the confusion matrix C, computed on the validation set of the corresponding training set the model was trained on, and define the confusion matrix as a sum of the initial confusion matrix and a single entry matrix $S_{j,i}$ with 1 on index j, i. The entry j, i denotes a combination of a prediction-target pair where the predicted label is j and the target label is i. We propose the loss to then be the negative expected QWK:

$$\mathcal{L}_{QWK}(p(y \mid \boldsymbol{x}), i) = -\sum_{j=1}^{M} p(y = j \mid \boldsymbol{x}) \kappa_{QW}(C + S_{j,i}). \tag{16}$$

The final QWK-Risk uncertainty estimate is then:

$$\begin{aligned} \mathcal{R}_{QWK}(\boldsymbol{x}) = &-\sum_{i=1}^{M} p(y = i \mid \boldsymbol{x}) \\ &\times \sum_{j=1}^{M} p(y = j \mid \boldsymbol{x}) \kappa_{QW}(C + S_{j,i}). \end{aligned} \tag{17}$$

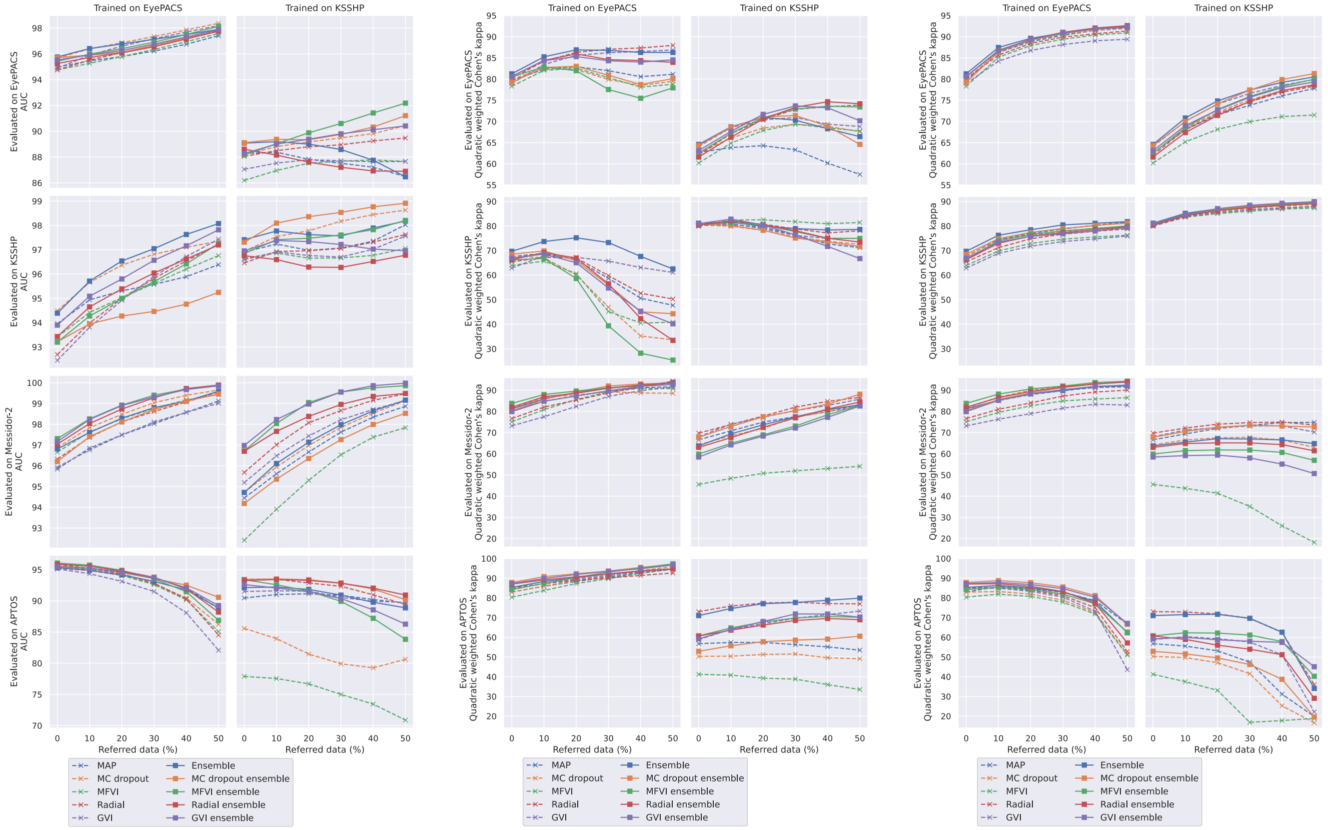
The QWK-Risk can be interpreted as the expected negative QWK value for an input example $\boldsymbol{x}$, similar to entropy being the expected negative log-likelihood.

The utility of the uncertainty information is evaluated in a similar manner as in previous works. We compute the uncertainty as the entropy or as the QWK-Risk of the posterior predictive distribution, that is computed using 100 Monte Carlo samples for MC dropout, MFVI, GVI, and Radial BNN, and with three ensemble members for the ensemble experiments. The uncertainty is then used to reject some proportion of the most uncertain cases, which is called the referral level. The performance is evaluated for the binary RDR task using the area under the receiver operating characteristic curve (AUC), and for the 5-class PIRC task we use the quadratic weighted Cohen's kappa (QWK), similar to [13] and [14]. Both the AUC and QWK are evaluated at 0% (no referral), 30%, and 50% referral levels, and are presented in a scale with maximum of 100 to improve readability.

III. RESULTS

A. RDR CLASSIFICATION

Visual illustration of results for performance after referral for the models trained on RDR classification is presented in Fig. 1A. Detailed RDR results are presented in Table 4. On the full test set, i.e. 0% referral level, our EyePACS data trained models had better results on all of the benchmark sets in comparison to previous works, thus setting a higher standard as our baseline performance. Indeed, our worst models are better than the best models in previous works: on EyePACS [9] had 92.7 AUC, [12] 82.5 AUC, [10] 94.5 AUC, and [18] 92.5 AUC, in comparison to our, MFVI and Radial, with



(a)

(b)

(c)

FIGURE 1. Performance after referral using (a) posterior predictive entropy to refer on the RDR task, (b) posterior predictive entropy to refer on the PIRC task, and (c) QWK-Risk to refer on the PIRC task. Each column shows which dataset was used for training and each row which test dataset was used for the results. The y-axis values are scaled to have a theoretical maximum of 100 and the limits chosen for better readability. The y-axis limits are shared in the PIRC task between entropy and QWK-Risk referral plots for each test set, in order to enable direct visual comparison of the methods.

94.8 AUC. In addition, on APTOS dataset [12] had 77.2 AUC and [18] 94.6 AUC, whereas our worst models, MC dropout and GVI, had 95.2 AUC. The best Messidor result in [9] was 95.5 AUC, whereas our worst model, deterministic MAP network, had 95.9 AUC, however, our results are not directly comparable since they use the Messidor set, which is a subset of the more recent Messidor-2 set.

Also, all of the KSSHP data trained models had a high AUC on 0% referral level for the KSSHP test set. Indeed, every model outperforms the corresponding model trained and tested on the EyePACS dataset. However, the generalization of the KSSHP trained models to Messidor-2 and APTOS sets was worse in every comparison than the generalization of the EyePACS trained models.

The uncertainty enabled referral process results for the EyePACS data trained models are similar to [9], as we observe the referral process to systematically increase the AUC on the EyePACS test set and on the Messidor-2 set, and as a novel result, we observe that the referral process also increases the clinical KSSHP dataset AUC. We observe different behaviour on the APTOS dataset than in [12] but similar to some models in [18], as the AUC of all models systematically decreases when utilizing the uncertainty. Surprisingly, we observe strong uncertainty estimates for the deterministic MAP network, unlike in [9] and [12], that observed generally worse uncertainty estimates for the point estimate approach. This difference is likely due to the strong regularization we used in training our MAP network.

It turns out that the models trained with the KSSHP data had different utility from the uncertainty referral than those trained using the EyePACS set. Specifically, on the KSSHP test set, from the 0% to 30% referral level, the MFVI and Radial ensemble models exhibit slightly decreased performance. In contrast, the GVI model performance stayed constant across this range, but decreased from 30% to 50%. In addition, the uncertainty estimates do not generalize to the EyePACS test set to such an extent as they did using the EyePACS set trained models on the KSSHP set. We see that on the EyePACS set, the uncertainty estimates given by the MAP network, deep ensemble, and the Radial ensemble cannot be used to improve the performance by referral, and the MFVI and GVI models do not improve upon the 30% referral rate. However, all the models perform well to the Messidor-2 dataset on all referral levels with the GVI ensemble even attaining 100.0 AUC for the 50% referral level. In the case of APTOS dataset we observe similar performance as with the EyePACS trained model, as the performance decreases along the referral rate, with the exception of baseline network improving from 0% to 30%.

In terms of overall performance, we observe that best results for the EyePACS, KSSHP, and the Messidor-2 datasets were obtained by referring data. For EyePACS and KSSHP the best results were for within-distribution trained models. The best performance on the EyePACS was obtained with MC dropout model that reaches 98.4 AUC with 50% referral level and on the KSSHP with MC dropout ensemble that reached 98.9 AUC with 50% referral level. The best out-of-distribution performance on the Messidor-2 set was obtained with the KSSHP trained GVI ensemble with 50% referral level that reached 100.0 AUC. For the APTOS set the referral did not improve performance, as the EyePACS trained MFVI ensemble with 0% referral level reached the highest AUC of 96.1.

B. PIRC CLASSIFICATION

As no previous works have utilized robust neural networks for the 5-class clinical PIRC scheme, we cannot directly compare the absolute performance. However, standard deep learning methods have been utilized, with similar data to our benchmark datasets. In [13], a deterministic Inception-v4 model was trained using over 1.6 million images from EyePACS affiliated clinics, 3 eye hospitals in India, one of them the same as the origin of the APTOS dataset, and the Messidor-2 dataset. The test set consisted of EyePACS originating images, on which the model achieved QWK of 84.0, which we consider the high performance baseline.

Visual illustration of PIRC results are presented in Fig. 1B, full results in 5. When no images are referred, the best performance is achieved with different ensembles: the deep ensemble achieves 81.3 QWK on the EyePACS dataset and 81.1 QWK on the KSSHP dataset. Similarly, the MFVI ensemble achieves 83.9 QWK on the Messidor-2 dataset, and the MC dropout ensemble achieves 87.9 QWK on the APTOS dataset. EyePACS trained models had better out-of-distribution performance than the KSSHP trained models.

While using entropy as the measure of uncertainty, the only EyePACS trained models that improved within distribution performance for all the referral levels were GVI, Radial, and GVI ensemble. These models also surpassed the 84.0 QWK when referring $\geq$ 30% of images. Also, deep ensemble achieved 86.8 QWK and Radial ensemble 84.6 QWK for the 30% referral level. All the models, except the MC dropout, consistently improve on the Messidor-2 dataset and reach over 84.0 QWK when referring $\geq$ 30% of images. On the APTOS dataset, all models consistently improve for all referral levels and they surpass the 84.0 QWK when referring $\geq$ 30% of images. However, we can see that on the KSSHP dataset, the models degrade consistently, apart from GVI and deep ensemble for 0% to 30% referral level, and no model reaches competitive performance on the KSSHP for any referral level.

When using the KSSHP data for training, we get significantly worse results compared to the EyePACS trained models, as no model consistently improves on the within distribution set, and overall, no model reaches the 84.0 QWK. Indeed, only the MFVI improves from 0% to 30% referral level. In terms of EyePACS generalization, all models improve from 0% to 30%, and Radial, MFVI ensemble, and Radial ensemble improve also from 30% to 50% referral level. However, no model reaches over 80.0 QWK. All models consistently improve for the Messidor-2 dataset and GVI, Radial, MC dropout ensemble, and Radial ensemble reach over 84.0 QWK for the 50% referral level. However, the MFVI has generally poor performance, as the QWK only increases from 45.7 to 54.5. On the APTOS dataset, MC dropout, GVI, Radial, and all ensemble models improve from 0% to 30% referral level. On the 30% to 50% referral level, GVI and all ensembles except the GVI ensemble improve. Even though some models improve in the referral process, no model reaches over 80.0 QWK, the deep ensemble being closest with 79.9 QWK on the 50% referral level.

The overall best performance for the EyePACS set is obtained using EyePACS trained Radial when referring 50% of images, QWK of 87.9, and for the KSSHP set using KSSHP trained MFVI and referring 30% of images, QWK of 81.6. For the out-of-distribution sets, the best Messidor-2 results are obtained using EyePACS trained Radial when referring 50% of images, QWK of 94.3, and for the APTOS set using EyePACS trained MFVI ensemble when referring 50% of images, QWK of 97.3. Since the uncertainty estimation is motivated from the point of view of clinical interest, it is extremely concerning that the methods appear to not work in a similar manner when trained on a clinical dataset in comparison to the benchmark datasets, especially on the clinical set itself.

C. QWK-RISK AS AN ALTERNATIVE UNCERTAINTY MEASURE

Our proposed QWK-Risk uncertainty measure results are presented in Fig. 1C and Table 6. We can see that the QWK-Risk uncertainty based method systematically improves both the within distribution test results and cross out-of-distribution results for the two train sets. Within distribution for the EyePACS set, all models reach over 84.0 QWK when referring $\geq$ 30% of images. No EyePACS trained model reaches the 84.0 QWK on the KSSHP set. However, QWK-Risk enables deep ensemble to reach $\geq$ 80.0 QWK for all referral levels. In addition, deterministic MAP network, MC dropout ensemble, and MFVI ensemble reach $\geq$ 80.0 QWK for 50% referral level. On the Messidor-2 set, the performance of some models decreases and some improve in comparison to the entropy based uncertainty, most notably the GVI no longer reaches over 84.0 QWK but all other models reach over 84.0 QWK for referral levels $\geq$ 30%. Systematic decrease in the performance is seen for the APTOS dataset,

TABLE 4. RDR results. Mean and standard deviation computed for 100 bootstrap resamples of the test data. Relative improvement to previous referral level is denoted with a green up-pointing triangle, orange equals sign, or red down-pointing triangle for increasing, equal, or worse performance, respectively.

Test Dataset	Method	EyePACS trained			KSSHP trained		
		AUC Ref. 50%	AUC Ref. 30%	AUC Ref. 0%	AUC Ref. 50%	AUC Ref. 30%	AUC Ref. 0%
EyePACS	MAP	▲ 97.4 ± 0.2	▲ 96.2 ± 0.2	95.0 ± 0.1	▼ 86.5 ± 0.5	▼ 87.5 ± 0.4	**88.3 ± 0.2**
	MC dropout	▲ **98.4 ± 0.1**	▲ **97.3 ± 0.2**	**95.7 ± 0.1**	▲ **90.4 ± 0.4**	▲ **89.5 ± 0.3**	88.2 ± 0.2
	MFVI	▲ 97.6 ± 0.2	▲ 96.3 ± 0.2	94.8 ± 0.1	= 87.7 ± 0.5	▲ 87.7 ± 0.3	86.2 ± 0.2
	GVI	▲ 98.2 ± 0.1	▲ 97.2 ± 0.1	94.9 ± 0.1	= 87.7 ± 0.5	▲ 87.7 ± 0.4	87.0 ± 0.2
	Radial	▲ 97.8 ± 0.2	▲ 96.7 ± 0.2	94.8 ± 0.1	▲ 89.5 ± 0.4	▲ 88.9 ± 0.3	88.0 ± 0.2
KSSHP	MAP	▲ 96.4 ± 0.7	▲ 95.5 ± 0.7	93.9 ± 0.5	▲ 98.0 ± 0.8	▲ 97.0 ± 0.7	**96.9 ± 0.3**
	MC dropout	▲ 97.3 ± 0.6	▲ **96.7 ± 0.5**	**94.4 ± 0.4**	▲ **98.6 ± 0.5**	▲ **98.1 ± 0.5**	**96.9 ± 0.3**
	MFVI	▲ 96.8 ± 0.6	▲ 95.6 ± 0.6	93.4 ± 0.5	▲ 96.9 ± 0.8	▼ 96.5 ± 0.7	96.6 ± 0.3
	GVI	▲ **97.4 ± 0.4**	▲ 95.9 ± 0.5	92.4 ± 0.5	▲ 97.4 ± 0.7	= 96.5 ± 0.7	96.5 ± 0.3
	Radial	▲ 97.2 ± 0.5	▲ 95.8 ± 0.5	92.7 ± 0.5	▲ 97.5 ± 0.6	▲ 96.9 ± 0.6	96.4 ± 0.3
Messidor-2	MAP	▲ 99.1 ± 0.3	▲ 98.1 ± 0.4	95.9 ± 0.4	▲ 98.9 ± 0.4	▲ 97.7 ± 0.5	94.5 ± 0.5
	MC dropout	▲ **99.7 ± 0.1**	▲ **99.1 ± 0.2**	**96.9 ± 0.4**	▲ 99.2 ± 0.3	▲ 97.9 ± 0.4	94.8 ± 0.6
	MFVI	▲ 99.4 ± 0.3	▲ 98.8 ± 0.3	96.7 ± 0.4	▲ 97.9 ± 0.4	▲ 96.5 ± 0.5	92.4 ± 0.6
	GVI	▲ 99.0 ± 0.3	▲ 98.2 ± 0.4	96.0 ± 0.5	▲ 99.1 ± 0.4	▲ 98.2 ± 0.4	95.3 ± 0.5
	Radial	▲ 99.6 ± 0.2	▲ 98.6 ± 0.4	96.4 ± 0.5	▲ **99.5 ± 0.3**	▲ **98.7 ± 0.3**	**95.8 ± 0.4**
APTOS	MAP	▼ **88.8 ± 1.0**	▼ **93.6 ± 0.5**	95.5 ± 0.3	▼ 89.7 ± 0.8	▲ 91.0 ± 0.6	90.6 ± 0.5
	MC dropout	▼ 86.2 ± 1.3	▼ 92.7 ± 0.6	95.2 ± 0.4	▲ 80.7 ± 1.0	▼ 79.9 ± 0.9	85.6 ± 0.7
	MFVI	▼ 85.2 ± 1.3	▼ 92.7 ± 0.6	95.6 ± 0.3	▼ 71.0 ± 1.2	▼ 75.2 ± 1.0	78.0 ± 0.8
	GVI	▼ 82.3 ± 1.6	▼ 91.6 ± 0.6	95.2 ± 0.4	▼ **89.8 ± 0.8**	▼ 90.6 ± 0.6	91.5 ± 0.5
	Radial	▼ 84.5 ± 1.5	▼ 92.9 ± 0.6	**95.9 ± 0.3**	▼ 89.6 ± 0.8	▼ **92.4 ± 0.6**	**93.5 ± 0.4**
EyePACS	Deep ensemble	▲ 97.9 ± 0.2	▲ **97.1 ± 0.2**	**95.8 ± 0.1**	▼ 86.5 ± 0.6	▼ 88.6 ± 0.3	**89.1 ± 0.2**
	MC dropout ensemble	▲ 97.7 ± 0.2	▲ 96.7 ± 0.2	95.7 ± 0.1	▲ 91.2 ± 0.4	▲ 89.7 ± 0.3	**89.1 ± 0.2**
	MFVI ensemble	▲ **98.1 ± 0.2**	▲ 96.9 ± 0.2	95.4 ± 0.1	▲ **92.2 ± 0.3**	▲ **90.6 ± 0.3**	88.2 ± 0.2
	GVI ensemble	▲ 97.9 ± 0.2	▲ 96.8 ± 0.2	95.5 ± 0.1	▲ 90.4 ± 0.4	▲ 89.8 ± 0.3	88.3 ± 0.2
	Radial ensemble	▲ 97.8 ± 0.2	▲ 96.6 ± 0.2	95.3 ± 0.1	▼ 86.9 ± 0.5	▼ 87.2 ± 0.4	88.6 ± 0.2
KSSHP	Deep ensemble	▲ **98.0 ± 0.5**	▲ **97.0 ± 0.4**	**94.4 ± 0.4**	▲ 98.1 ± 1.0	▲ 97.5 ± 0.7	**97.4 ± 0.2**
	MC dropout ensemble	▲ 95.1 ± 0.9	▲ 94.3 ± 0.8	93.2 ± 0.5	▲ **98.9 ± 0.5**	▲ **98.5 ± 0.5**	97.3 ± 0.2
	MFVI ensemble	▲ 97.1 ± 0.6	▲ 95.6 ± 0.6	93.1 ± 0.5	▲ 98.1 ± 0.6	▲ 97.5 ± 0.5	96.8 ± 0.3
	GVI ensemble	▲ 97.7 ± 0.6	▲ 96.4 ± 0.6	93.8 ± 0.4	▼ 96.8 ± 0.9	▲ 97.1 ± 0.7	96.9 ± 0.3
	Radial ensemble	▲ 97.2 ± 0.6	▲ 96.0 ± 0.6	93.4 ± 0.4	▲ 96.6 ± 0.8	▼ 96.1 ± 0.7	96.7 ± 0.3
Messidor-2	Deep ensemble	▲ 99.6 ± 0.1	▲ 98.8 ± 0.3	96.8 ± 0.4	▲ 99.3 ± 0.4	▲ 98.1 ± 0.4	94.9 ± 0.6
	MC dropout ensemble	▲ 99.4 ± 0.2	▲ 98.7 ± 0.3	96.3 ± 0.4	▲ 98.6 ± 0.6	▲ 97.3 ± 0.6	94.3 ± 0.6
	MFVI ensemble	▲ **99.9 ± 0.1**	▲ **99.5 ± 0.2**	**97.4 ± 0.3**	▲ 99.9 ± 0.1	▲ **99.6 ± 0.2**	96.8 ± 0.4
	GVI ensemble	▲ **99.9 ± 0.1**	▲ 99.3 ± 0.2	97.2 ± 0.3	▲ **100.0 ± 0.0**	▲ 99.5 ± 0.2	**97.0 ± 0.3**
	Radial ensemble	▲ **99.9 ± 0.1**	▲ 99.3 ± 0.2	97.1 ± 0.4	▲ 99.5 ± 0.3	▲ 99.0 ± 0.3	96.8 ± 0.4
APTOS	Deep ensemble	▼ 88.9 ± 1.1	▼ 93.1 ± 0.5	95.3 ± 0.4	▼ 89.1 ± 0.8	▼ 91.0 ± 0.6	92.2 ± 0.5
	MC dropout ensemble	▼ **90.6 ± 0.9**	▼ 93.6 ± 0.5	95.6 ± 0.3	▼ 90.3 ± 0.8	▼ **92.9 ± 0.5**	**93.5 ± 0.4**
	MFVI ensemble	▼ 86.9 ± 1.2	▼ 93.5 ± 0.6	**96.1 ± 0.3**	▼ 84.1 ± 1.2	▼ 90.0 ± 0.7	93.4 ± 0.4
	GVI ensemble	▼ 89.3 ± 0.9	▼ **93.8 ± 0.5**	95.5 ± 0.3	▼ 86.5 ± 1.0	▼ 90.5 ± 0.7	92.7 ± 0.4
	Radial ensemble	▼ 88.3 ± 1.1	▼ **93.8 ± 0.5**	95.9 ± 0.3	▼ **91.1 ± 0.8**	▼ **92.9 ± 0.6**	93.3 ± 0.4

and now only MC dropout ensemble and GVI ensemble reach over 84.0 QWK when referring 30% of images.

From the clinical perspective, an important finding is that using the QWK-Risk, all our models reach $\geq$ 84.0 QWK for referral levels $\geq$ 30% when trained and tested on the KSSHP set. Indeed, the GVI ensemble even reaches 89.9 QWK for the within KSSHP distribution test when 50% of examples are referred. The generalization of the uncertainty estimates to the EyePACS set also increases and now the deep ensemble, MC dropout ensemble, and MFVI ensemble reach $\geq$ 80.0 QWK, however, no model can reach the 84.0 QWK. The QWK-Risk decreases the performance for the Messidor-2 and APTOS datasets in comparison to entropy.

Using the QWK-Risk as the uncertainty measure, the best performance for the EyePACS set is obtained using EyePACS trained MC dropout ensemble and Radial ensemble when referring 50% of images, QWK of 92.6, for the KSSHP set using KSSHP trained GVI ensemble and referring 50% of images, QWK of 89.9, for the Messidor-2 using EyePACS trained MC dropout ensemble and MFVI ensemble when referring 50% of images, QWK of 94.3, and for the APTOS set using EyePACS trained MC dropout ensemble when referring 0% of images, QWK of 87.9.

IV. DISCUSSION

We have replicated the usefulness of entropy as an uncertainty estimate in RDR classification task for most robust neural networks when training with the EyePACS benchmark dataset, and demonstrated to a somewhat lesser extend the same finding for the KSSHP clinical hospital dataset. We also observe that the quality of the uncertainty estimates on out-of-distribution tests decreases when the models are trained with the KSSHP set. In addition, we show that entropy is less suitable as a measure of uncertainty in the clinical 5-class PIRC classification task for the EyePACS and KSSHP datasets. We have proposed and demonstrated that

TABLE 5. PIRC results using posterior predictive entropy as the uncertainty measure. Mean and standard deviation computed for 100 bootstrap resamples of the test data. Relative improvement to previous referral level is denoted with a green up-pointing triangle, orange equals sign, or red down-pointing triangle for increasing, equal, or worse performance, respectively.

Test Dataset	Method	EyePACS trained			KSSHP trained		
		QWK Ref. 50%	QWK Ref. 30%	QWK Ref. 0%	QWK Ref. 50%	QWK Ref. 30%	QWK Ref. 0%
EyePACS	MAP	▼81.3 ± 1.0	▲82.0 ± 0.6	79.7 ± 0.4	▼57.4 ± 1.0	▲63.3 ± 0.7	62.7 ± 0.4
	MC dropout	▼79.5 ± 1.2	▲80.0 ± 0.7	79.9 ± 0.4	▼67.6 ± 0.8	▲69.3 ± 0.5	62.5 ± 0.4
	MFVI	▼79.0 ± 1.1	▲80.5 ± 0.7	78.4 ± 0.4	▼67.7 ± 0.7	▲69.3 ± 0.5	60.2 ± 0.5
	GVI	▲86.9 ± 0.6	▲86.2 ± 0.4	79.0 ± 0.3	▼68.8 ± 0.9	▲70.9 ± 0.6	**64.1 ± 0.5**
	Radial	▲**87.9 ± 0.5**	▲**87.0 ± 0.4**	**80.2 ± 0.3**	▲**73.8 ± 0.8**	▲**72.8 ± 0.5**	62.0 ± 0.4
KSSHP	MAP	▼47.6 ± 3.1	▼58.8 ± 2.0	67.5 ± 0.9	▼70.9 ± 1.6	▼76.0 ± 1.0	**81.0 ± 0.6**
	MC dropout	▼32.6 ± 4.8	▼47.2 ± 2.7	**67.7 ± 0.8**	▼72.1 ± 1.4	▼75.3 ± 1.1	80.3 ± 0.6
	MFVI	▼40.5 ± 3.2	▼45.3 ± 2.4	63.8 ± 0.9	▼**81.4 ± 1.1**	▲**81.6 ± 0.8**	79.9 ± 0.6
	GVI	▼**61.1 ± 1.8**	▲**65.6 ± 1.2**	62.8 ± 0.8	▼74.8 ± 1.7	▼76.1 ± 1.1	80.0 ± 0.6
	Radial	▼49.9 ± 2.9	▼59.6 ± 2.1	65.4 ± 0.9	▼77.6 ± 1.6	▼78.5 ± 1.3	79.9 ± 0.6
Messidor-2	MAP	▲91.7 ± 1.3	▲**90.0 ± 1.0**	80.7 ± 1.2	▲83.2 ± 1.2	▲77.6 ± 1.3	66.7 ± 1.4
	MC dropout	▼88.5 ± 1.8	▲89.5 ± 1.2	**81.8 ± 1.4**	▲82.5 ± 1.3	▲77.2 ± 1.3	64.3 ± 1.4
	MFVI	▲90.7 ± 1.3	▲89.3 ± 1.2	75.3 ± 1.6	▲54.5 ± 2.3	▲51.8 ± 1.7	45.7 ± 1.7
	GVI	▲91.8 ± 1.1	▲87.2 ± 1.1	73.3 ± 1.4	▲85.8 ± 1.2	▲80.2 ± 1.1	68.1 ± 1.4
	Radial	▲**94.3 ± 0.9**	▲88.8 ± 1.1	76.8 ± 1.3	▲**86.6 ± 0.8**	▲**81.9 ± 1.0**	**69.8 ± 1.3**
APTOS	MAP	▲94.8 ± 0.6	▲90.5 ± 0.6	83.1 ± 0.6	▼53.6 ± 1.4	▼56.3 ± 1.0	56.8 ± 0.9
	MC dropout	▲**95.9 ± 0.5**	▲91.2 ± 0.5	83.0 ± 0.6	▼49.1 ± 1.3	▲51.5 ± 1.0	50.4 ± 0.9
	MFVI	▲94.5 ± 0.7	▲89.8 ± 0.6	80.4 ± 0.6	▼33.7 ± 1.3	▼38.8 ± 1.2	41.3 ± 1.0
	GVI	▲94.6 ± 0.5	▲**91.4 ± 0.4**	84.6 ± 0.5	▲73.4 ± 1.1	▲69.6 ± 1.0	59.0 ± 1.0
	Radial	▲92.7 ± 0.4	▲90.1 ± 0.5	**85.2 ± 0.5**	▼**77.1 ± 1.0**	▲**77.7 ± 0.8**	**73.0 ± 0.7**
EyePACS	Deep ensemble	▼**86.4 ± 0.6**	▲**86.8 ± 0.4**	**81.3 ± 0.4**	▼66.4 ± 0.9	▲70.3 ± 0.6	**64.5 ± 0.4**
	MC dropout ensemble	▼80.1 ± 1.1	▲80.9 ± 0.7	79.2 ± 0.4	▼64.6 ± 1.0	▲71.3 ± 0.5	64.3 ± 0.4
	MFVI ensemble	▲77.9 ± 1.6	▼77.5 ± 1.0	80.6 ± 0.3	▲73.4 ± 0.6	▲72.9 ± 0.4	62.5 ± 0.4
	GVI ensemble	▲84.6 ± 0.8	▲84.3 ± 0.6	80.7 ± 0.4	▼70.1 ± 0.8	▲**73.7 ± 0.5**	63.1 ± 0.4
	Radial ensemble	▼83.9 ± 0.9	▲84.6 ± 0.5	80.7 ± 0.3	▲**74.2 ± 0.6**	▲73.3 ± 0.5	61.6 ± 0.4
KSSHP	Deep ensemble	▼**62.3 ± 2.0**	▲**73.1 ± 1.1**	**69.6 ± 0.8**	▼**78.4 ± 1.3**	▼**78.8 ± 1.0**	**81.1 ± 0.5**
	MC dropout ensemble	▼43.7 ± 3.9	▼56.0 ± 2.0	68.4 ± 0.9	▼71.3 ± 1.8	▼75.0 ± 1.0	80.6 ± 0.6
	MFVI ensemble	▼24.4 ± 5.3	▼39.4 ± 2.9	66.1 ± 0.9	▼74.7 ± 1.5	▼77.7 ± 1.1	80.5 ± 0.6
	GVI ensemble	▼40.0 ± 3.4	▼55.0 ± 1.9	66.8 ± 0.9	▼66.6 ± 2.0	▼76.4 ± 1.1	80.9 ± 0.5
	Radial ensemble	▼33.4 ± 3.5	▼56.2 ± 2.2	66.2 ± 1.0	▼73.2 ± 1.6	▼77.8 ± 1.1	80.2 ± 0.6
Messidor-2	Deep ensemble	▲**93.8 ± 1.0**	▲91.1 ± 0.9	81.4 ± 1.2	▲83.2 ± 1.2	▲77.6 ± 1.2	63.7 ± 1.5
	MC dropout ensemble	▲93.5 ± 1.1	▲**92.1 ± 0.8**	80.3 ± 1.3	▲**88.1 ± 1.1**	▲**80.4 ± 1.3**	**68.0 ± 1.5**
	MFVI ensemble	▲93.2 ± 1.3	▲91.3 ± 1.0	**83.9 ± 1.2**	▲83.2 ± 1.4	▲73.2 ± 1.6	59.9 ± 1.5
	GVI ensemble	▲92.6 ± 1.2	▲89.9 ± 1.0	80.0 ± 1.4	▲82.6 ± 1.5	▲72.3 ± 1.5	58.4 ± 1.4
	Radial ensemble	▲93.7 ± 1.0	▲91.0 ± 1.0	82.2 ± 1.3	▲84.4 ± 1.2	▲77.2 ± 1.2	63.1 ± 1.5
APTOS	Deep ensemble	▲94.8 ± 0.5	▲92.2 ± 0.5	85.4 ± 0.5	▲**79.9 ± 0.8**	▲**77.6 ± 0.8**	**71.0 ± 0.7**
	MC dropout ensemble	▲97.0 ± 0.4	▲**93.7 ± 0.4**	**87.9 ± 0.4**	▲60.8 ± 1.3	▲58.6 ± 1.1	53.0 ± 1.0
	MFVI ensemble	▲**97.3 ± 0.4**	▲93.1 ± 0.4	84.3 ± 0.6	▲70.2 ± 1.0	▲69.9 ± 0.8	60.7 ± 0.9
	GVI ensemble	▲96.7 ± 0.4	▲93.5 ± 0.4	87.3 ± 0.5	▼70.3 ± 1.1	▲71.8 ± 0.9	59.2 ± 0.8
	Radial ensemble	▲94.8 ± 0.4	▲92.2 ± 0.4	87.1 ± 0.4	▲68.9 ± 1.2	▲68.6 ± 1.0	60.6 ± 0.9

an uncertainty measure based on the classifier risk using the quadratic weighted Cohen's kappa provides a useful measure for the within-distribution uncertainty quantification that improves uncertainty based retained performance in comparison to entropy. However, the QWK-Risk decreases the out-of-distribution quality of the uncertainty estimates for the Messidor-2 and APTOS datasets.

From the clinical perspective, the classifiers should be able to indicate uncertainty, such that the cases for which the classifier is confident can be automatically classified, while the difficult cases can be manually verified, and possibly corrected, by medical experts. Since most of the benchmark datasets only permit research use, for real world use-cases the classifiers should be able to be trained using ''in-house'' hospital datasets. When training the classifier on a hospital dataset, we expect that it can be utilized for future cases within the same hospital. In our studies we did not observe competitive performance for the KSSHP set on the clinical PIRC system, when entropy was utilized to quantify uncertainty. This highlights the concern that methods developed using benchmark datasets might not generalize to the clinical setting. However, the proposed QWK-Risk uncertainty measure enabled the models to surpass the state-of-the-art when 30% of the most uncertain cases were referred, which demonstrates that the robust neural networks with a more appropriate uncertainty function may be utilized for clinical datasets and clinical use-cases.

The degradation and variability in out-of-distribution performance may partly be caused by different grading conventions and grading variability may prevent correct evaluation. There are several mechanisms causing variable grading. Firstly, the frequency of photographic screening may affect the distribution of retinopathy severity with advanced grades of retinopathy being rarer when screening is more frequent. The patients are referred and treated before the advanced stages and followed by clinical visits instead of fundus photography. Secondly, there are no grading schemes available for the classification of treatment outcomes after the

TABLE 6. PIRC results using QWK-Risk as the uncertainty measure. Mean and standard deviation computed for 100 bootstrap resamples of the test data. Relative improvement to previous referral level is denoted with a green up-pointing triangle, orange equals sign, or red down-pointing triangle for increasing, equal, or worse performance, respectively.

Test Dataset	Method	EyePACS trained			KSSHP trained		
		QWK Ref. 50%	QWK Ref. 30%	QWK Ref. 0%	QWK Ref. 50%	QWK Ref. 30%	QWK Ref. 0%
EyePACS	MAP	▲ **92.1 ± 0.2**	▲ 90.3 ± 0.2	79.7 ± 0.4	▲ 77.8 ± 0.4	▲ 73.7 ± 0.4	62.7 ± 0.4
	MC dropout	▲ 92.0 ± 0.2	▲ **90.6 ± 0.2**	79.9 ± 0.4	▲ 78.5 ± 0.3	▲ 75.1 ± 0.3	62.5 ± 0.4
	MFVI	▲ 91.0 ± 0.2	▲ 89.5 ± 0.2	78.4 ± 0.4	▲ 71.5 ± 0.4	▲ 69.9 ± 0.4	60.2 ± 0.5
	GVI	▲ 89.4 ± 0.2	▲ 88.2 ± 0.3	79.0 ± 0.3	▲ **79.9 ± 0.4**	▲ **76.6 ± 0.4**	**64.1 ± 0.5**
	Radial	▲ 91.3 ± 0.2	▲ 89.9 ± 0.2	**80.2 ± 0.3**	▲ 78.2 ± 0.4	▲ 74.6 ± 0.4	62.0 ± 0.4
KSSHP	MAP	▲ **81.4 ± 0.9**	▲ **78.6 ± 0.9**	67.5 ± 0.9	▲ **89.4 ± 0.5**	▲ **87.9 ± 0.5**	**81.0 ± 0.6**
	MC dropout	▲ 79.6 ± 1.0	▲ 77.5 ± 0.9	**67.7 ± 0.8**	▲ 88.6 ± 0.5	▲ 87.1 ± 0.5	80.3 ± 0.6
	MFVI	▲ 76.1 ± 1.0	▲ 74.5 ± 0.9	63.8 ± 0.9	▲ 87.3 ± 0.5	▲ 85.9 ± 0.5	79.9 ± 0.6
	GVI	▲ 75.8 ± 0.8	▲ 73.6 ± 0.8	62.8 ± 0.8	▲ 87.9 ± 0.4	▲ 86.5 ± 0.4	80.0 ± 0.6
	Radial	▲ 78.7 ± 0.9	▲ 76.6 ± 0.8	65.4 ± 0.9	▲ 87.8 ± 0.5	▲ 86.5 ± 0.5	79.9 ± 0.6
Messidor-2	MAP	▲ 91.9 ± 0.8	▲ 89.9 ± 0.8	80.7 ± 1.2	▲ **74.8 ± 1.4**	▲ 73.4 ± 1.3	66.7 ± 1.4
	MC dropout	▲ **94.0 ± 0.6**	▲ **91.7 ± 0.7**	**81.8 ± 1.4**	▼ 63.1 ± 2.1	▲ 67.9 ± 1.5	64.3 ± 1.4
	MFVI	▲ 86.3 ± 1.6	▲ 84.8 ± 1.4	75.3 ± 1.6	▼ 18.3 ± 1.3	▼ 35.4 ± 1.8	45.7 ± 1.7
	GVI	▲ 83.0 ± 1.5	▲ 81.7 ± 1.3	73.3 ± 1.4	▼ 70.5 ± 1.8	▲ 73.4 ± 1.3	68.1 ± 1.4
	Radial	▲ 90.0 ± 1.0	▲ 87.4 ± 1.0	76.8 ± 1.3	▼ 73.5 ± 1.4	▲ **74.6 ± 1.1**	**69.8 ± 1.3**
APTOS	MAP	▼ **66.5 ± 2.4**	▼ **82.8 ± 0.7**	83.1 ± 0.6	▼ 19.7 ± 1.3	▼ 47.3 ± 1.5	56.8 ± 0.9
	MC dropout	▼ 51.4 ± 4.7	▼ 79.0 ± 0.9	83.0 ± 0.6	▼ 16.6 ± 1.1	▼ 41.6 ± 1.3	50.4 ± 0.9
	MFVI	▼ 50.8 ± 4.1	▼ 77.9 ± 0.8	80.4 ± 0.6	▲ 18.7 ± 1.2	▼ 17.2 ± 1.1	41.3 ± 1.0
	GVI	▼ 45.1 ± 4.4	▼ 80.4 ± 0.8	84.6 ± 0.5	▼ 22.0 ± 5.4	▼ 57.9 ± 1.4	59.0 ± 1.0
	Radial	▼ 52.5 ± 5.0	▼ 81.3 ± 0.8	**85.2 ± 0.5**	▼ **35.8 ± 5.9**	▼ **69.5 ± 1.1**	**73.0 ± 0.7**
EyePACS	Deep ensemble	▲ 92.2 ± 0.2	▲ 90.9 ± 0.2	**81.3 ± 0.4**	▲ 80.5 ± 0.3	▲ **77.4 ± 0.4**	**64.5 ± 0.4**
	MC dropout ensemble	▲ **92.6 ± 0.2**	▲ 90.7 ± 0.2	79.2 ± 0.4	▲ **81.3 ± 0.3**	▲ **77.4 ± 0.3**	64.3 ± 0.4
	MFVI ensemble	▲ 92.5 ± 0.2	▲ 90.8 ± 0.2	80.6 ± 0.3	▲ 80.0 ± 0.3	▲ 75.9 ± 0.3	62.5 ± 0.4
	GVI ensemble	▲ 92.3 ± 0.2	▲ 90.9 ± 0.2	80.7 ± 0.4	▲ 79.3 ± 0.3	▲ 75.7 ± 0.4	63.1 ± 0.4
	Radial ensemble	▲ **92.6 ± 0.2**	▲ **91.0 ± 0.2**	80.7 ± 0.3	▲ 78.7 ± 0.3	▲ 74.7 ± 0.4	61.6 ± 0.4
KSSHP	Deep ensemble	▲ **81.6 ± 0.8**	▲ **80.2 ± 0.8**	**69.6 ± 0.8**	▲ 89.5 ± 0.4	▲ 88.1 ± 0.4	**81.1 ± 0.5**
	MC dropout ensemble	▲ 81.1 ± 0.9	▲ 78.9 ± 0.9	68.4 ± 0.9	▲ 89.4 ± 0.4	▲ 87.9 ± 0.5	80.6 ± 0.6
	MFVI ensemble	▲ 80.0 ± 0.9	▲ 77.9 ± 0.9	66.1 ± 0.9	▲ 89.7 ± 0.4	▲ 87.7 ± 0.4	80.5 ± 0.6
	GVI ensemble	▲ 79.3 ± 0.9	▲ 76.7 ± 0.9	66.8 ± 0.9	▲ **89.9 ± 0.4**	▲ **88.4 ± 0.4**	80.9 ± 0.5
	Radial ensemble	▲ 79.4 ± 1.0	▲ 77.1 ± 0.9	66.2 ± 1.0	▲ 89.1 ± 0.5	▲ 87.4 ± 0.5	80.2 ± 0.6
Messidor-2	Deep ensemble	▲ 92.0 ± 0.8	▲ 90.0 ± 0.9	81.4 ± 1.2	▼ 65.0 ± 1.8	▲ 67.1 ± 1.4	63.7 ± 1.5
	MC dropout ensemble	▲ **94.3 ± 0.7**	▲ 91.4 ± 0.8	80.3 ± 1.3	▼ **72.9 ± 1.9**	▲ **73.6 ± 1.4**	**68.0 ± 1.5**
	MFVI ensemble	▲ **94.3 ± 0.8**	▲ **92.1 ± 0.8**	**83.9 ± 1.2**	▼ 57.3 ± 2.2	▲ 61.8 ± 1.7	59.9 ± 1.5
	GVI ensemble	▲ 92.4 ± 0.9	▲ 90.3 ± 0.8	80.0 ± 1.4	▼ 50.9 ± 2.5	▼ 58.1 ± 1.9	58.4 ± 1.4
	Radial ensemble	▲ 93.9 ± 0.8	▲ 91.7 ± 0.8	82.2 ± 1.3	▼ 61.9 ± 2.3	▲ 65.4 ± 1.6	63.1 ± 1.5
APTOS	Deep ensemble	▼ 62.3 ± 3.0	▼ 83.2 ± 0.6	85.4 ± 0.5	▼ 33.7 ± 6.4	▼ **69.6 ± 1.1**	**71.0 ± 0.7**
	MC dropout ensemble	▼ 66.4 ± 3.3	▼ **85.6 ± 0.6**	**87.9 ± 0.4**	▼ 19.5 ± 1.1	▼ 46.1 ± 1.2	53.0 ± 1.0
	MFVI ensemble	▼ 62.6 ± 3.1	▼ 82.1 ± 0.7	84.3 ± 0.6	▼ 39.6 ± 4.6	▲ 61.2 ± 1.1	60.7 ± 0.9
	GVI ensemble	▼ **66.9 ± 2.8**	▼ 84.7 ± 0.6	87.3 ± 0.5	▼ **45.2 ± 4.1**	▼ 58.0 ± 1.0	59.2 ± 0.8
	Radial ensemble	▼ 56.9 ± 4.4	▼ 83.3 ± 0.7	87.1 ± 0.4	▼ 29.3 ± 4.8	▼ 54.1 ± 1.2	60.6 ± 0.9

treatment. As a result, severe retinopathy may be arbitrarily classified into many different grades. The third reason for variable grading is poor quality of the retinal images. The grading of poor quality photographs may reflect some preconceived notions of patient status, based on such factors as age and type of diabetes, availability of treatments and also previous treatments.

Our future work includes a more fine-grained analysis of the out-of-distribution performance of robust neural networks. We will be evaluating a larger dataset of diabetic retinopathy images from the Helsinki region in Finland. In addition to the country distribution shift, our aim is to examine the within country hospital region distribution shift using this data, in addition to the KSSHP dataset. Additionally, our future work includes utilization of some of the computationally more intensive solutions for robust deep learning. Moreover, we will be examining the impact of joint training from the different regions on performance and robustness. Lastly, we aim to examine multi modal inputs with the fusion of multi-view retinal images with the task of even finer grain diabetic retinopathy grading.

V. CONCLUSION

Uncertainty estimates, given by approximate Bayesian deep neural networks, can be used to refer uncertain diabetic retinopathy classifications, with high performance on the certain classifications, using the binary referable/non-referable as well as clinical 5-class proposed international diabetic retinopathy classification scheme. For the 5-class scheme, uncertainty can be estimated using expected conditional risk based QWK-Risk uncertainty measure, to improve the performance on clinical data. Properly regularized standard neural networks exhibit also well calibrated performance. The results suggest that methods developed for benchmark datasets tend to generalize better to them than to the clinical dataset. Uncertainty estimates could improve safety and trustworthiness of the deep learning systems, as most uncertain classifications can be manually corrected, if needed, while expecting high performance on the more certain classifications.

REFERENCES

[1] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Jun. 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[2] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, in Lecture Notes in Computer Science, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 70, D. Precup and Y. W. Teh, Eds. Sydney, NSW, Australia, Aug. 2017, pp. 1321–1330. [Online]. Available: http://proceedings.mlr.press/v70/guo17a.html

[5] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Opthalmol. Vis. Sci.*, vol. 57, no. 13, p. 5200, Oct. 2016.

[6] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, J. Cuadros, and R. Kim, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.

[7] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017, doi: 10.1001/jama.2017.18152.

[8] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala, and K. Kaski, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Sci. Rep.*, vol. 9, no. 1, p. 10750, Jul. 2019, doi: 10.1038/S41598-019-47181-W.

[9] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, p. 17816, Dec. 2017, doi: 10.1038/S41598-017-17876-Z.

[10] S. Farquhar, M. A. Osborne, and Y. Gal, "Radial Bayesian neural networks: Beyond discrete support in large-scale Bayesian deep learning," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, in Proceedings of Machine Learning Research, vol. 108, S. Chiappa and R. Calandra, Eds., Aug. 2020, pp. 1352–1362. [Online]. Available: https://proceedings.mlr.press/v108/farquhar20a.html

[11] M. S. Ayhan and P. Berens, "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2018, pp. 1–9.

[12] A. Filos, S. Farquhar, A. N. Gomez, T. G. J. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal, "A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks," 2019, *arXiv:1912.10481*.

[13] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng, and D. R. Webster, "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, Aug. 2018, doi: 10.1016/j.ophtha.2018.01.034.

[14] P. Ruamviboonsuk *et al.*, "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program," *npj Digit. Med.*, vol. 2, no. 1, p. 25, Apr. 2019, doi: 10.1038/S41746-019-0099-8.

[15] J. Cuadros and G. Bresnick, "EyePACS: An adaptable telemedicine system for diabetic retinopathy screening," *J. Diabetes Sci. Technol.*, vol. 3, pp. 509–516, May 2009.

[16] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.

[17] Aravind Eye Hospital & PG Institute of Ophthalmology (Managed and Run by the Govel Trust) in Support of the Asia Pacific Tele-Ophthalmology Society (APTOS) Symposium. (2019). *APTOS 2019 Blindness Detection*. Accessed: Jan. 1, 2020. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection/overview/aptos-2019

[18] N. Band, T. G. Rudner, Q. Feng, A. Filos, Z. Nado, M. W. Dusenberry, G. Jerfel, D. Tran, and Y. Gal, "Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks," in *Proc. 35th Conf. Neural Inf. Process. Syst. Workshop Distrib. Shifts, Connecting Methods Appl.*, 2021, pp. 1–40.

[19] M. S. Ayhan, L. Kühlewein, G. Aliyeva, W. Inhoffen, F. Ziemssen, and P. Berens, "Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101724.

[20] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, and A. Campilho, "DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101715, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520300797

[21] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, and P. Massin, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, 2013.

[22] J. Sahlsten, J. Jaskari, K. Kaski, and K. Hietala, "Diabeettisen retinopatian ja makulaturvotuksen luokittelu syväoppivan tekoälyjärjestelmän avulla," *Aikakauskirja Duodecim*, vol. 17, pp. 1971–1978, Sep. 2020. [Online]. Available: https://www.duodecimlehti.fi/duo15766

[23] Ministry of Social Affairs and Health, Finland. *Medical Research Act No. 488/1999*. Accessed: Jan. 15, 2022. [Online]. Available: https://www.finlex.fi/fi/

[24] Ministry of Social Affairs and Health, Finland. *Secondary Use of Health and Social Data Act No. 552/2019*. Accessed: Jan. 15, 2022. [Online]. Available: https://www.finlex.fi

[25] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: CRC Press, 2013.

[26] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. 32nd Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 37, F. Bach and D. Blei, Eds. Lille, France, Jul. 2015, pp. 1613–1622. [Online]. Available: http://proceedings.mlr.press/v37/blundell15.html

[27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6402–6413.

[28] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1–10.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[30] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems*, vol. 24, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2011. [Online]. Available: https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf

[31] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. Banff, AB, Canada, 2014, pp. 1–13.

[32] J. Knoblauch, J. Jewson, and T. Damoulas, "Generalized variational inference: Three arguments for deriving new posteriors," 2019, *arXiv:1904.02063*.

[33] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968, doi: 10.1037/H0026256.

[34] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.

[35] V. Franc and D. Prusa, "On discriminative learning of prediction uncertainty," in *Proc. 36th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Jun. 2019, pp. 1963–1971. [Online]. Available: https://proceedings.mlr.press/v97/franc19a.html

[36] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970.

[37] L. Rosasco, E. Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?" *Neural Comput.*, vol. 16, no. 5, pp. 1063–1076, May 2004, doi: 10.1162/089976604773135104.

JOEL JASKARI received the B.Sc. (Tech.) and M.Sc. (Tech.) degrees from Aalto University, Finland, in 2013 and 2017, respectively, where he is currently pursuing the D.Sc. degree with the Department of Computer Science, School of Science. His research interests include machine learning, especially probabilistic deep learning and Bayesian methods and the application of machine learning in healthcare.

JAAKKO SAHLSTEN received the B.Sc. (Tech.) and M.Sc. (Tech.) degrees from Aalto University, Finland, in 2016 and 2018, respectively, where he is currently pursuing the D.Sc. degree with the Department of Computer Science, School of Science. His research interests include machine learning, especially robust deep neural networks and their applications for issues in the healthcare domain.

THEODOROS DAMOULAS is currently a Professor of machine learning with the Department of Computer Science and the Department of Statistics, The University of Warwick. His research interests include probabilistic machine learning and Bayesian statistics. In 2021, he was awarded the prestigious five-year UKRI Turing AI Fellowship to lead research that sets the ML Foundations of Digital Twins.

JEREMIAS KNOBLAUCH received the Ph.D. degree as part of the Oxford-Warwick Statistics Program, in 2022. He is currently an Assistant Professor/Lecturer with University College London. He is a fellow of EPSRC.

SIMO SÄRKKÄ (Senior Member, IEEE) is currently an Associate Professor with Aalto University. He has authored or coauthored over 100 peer-reviewed scientific articles and three books. His research interests include multi-sensor data processing systems and machine learning methods with applications in medical and health technology, target tracking, inverse problems, and location sensing. He is a member of the IEEE Machine Learning for Signal Processing Technical Committee. He has been serving as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.

LEO KÄRKKÄINEN received the Ph.D. degree in theoretical physics on high temperature quantum chromodynamics from the University of Helsinki, Finland, in 1990. He has several academic and industrial research positions. He is currently a Professor of practice position with the Department of Electrical Engineering, Aalto University, Finland, on leave of absence. His research interests include QCD, acoustics, nanotechnologies, medical technologies, AI, and AI ethics.

KUSTAA HIETALA received the M.D. degree from the University of Eastern Finland, in 1995, and the Specialist degree in ophthalmology and the D.M.Sc. degree from the University of Helsinki, in 2005 and 2013, respectively. His research interests include diabetic retinopathy and machine learning.

KIMMO K. KASKI received the M.Sc. (Tech.) and Lic.Tech. degrees in electrical engineering from the Helsinki University of Technology, Finland, in 1973 and 1977, respectively, and the D.Phil. degree in theoretical physics from the University of Oxford, U.K., in 1981. He is currently a Professor of computational science with the School of Science, Aalto University, Finland; a Supernumerary Fellow with the Wolfson College, University of Oxford, U.K.; an External Faculty with the Complexity Science Hub, Vienna, Austria; and a Visiting Fellow with The Alan Turing Institute, London, U.K. His research interests include computational science, statistical physics, complex system science, data science and artificial intelligence to their applications in various issues of social networks, and studies of digital health. He is a fellow of the American Physical Society, a fellow and a Chartered Physicist of the Institute of Physics, U.K., a fellow of the Academia Europaea and Finnish Academy of Science and Letters, a fellow of the Finnish Academy of Technical Sciences, and a fellow of the Mexican Academy of Sciences.