Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks

Neil Band[*,†] University of Oxford
Tim G. J. Rudner[*,†] University of Oxford
Qixuan Feng University of Oxford
Angelos Filos University of Oxford
Zachary Nado Google Research
Michael W. Dusenberry Google Research
Ghassen Jerfel Google Research
Dustin Tran Google Research
Yarin Gal University of Oxford

Abstract

Bayesian deep learning seeks to equip deep neural networks with the ability to precisely quantify their predictive uncertainty, and has promised to make deep learning more reliable for safety-critical real-world applications. Yet, existing Bayesian deep learning methods fall short of this promise; new methods continue to be evaluated on unrealistic test beds that do not reflect the complexities of downstream real-world tasks that would benefit most from reliable uncertainty quantification. We propose a set of real-world tasks that accurately reflect such complexities and are designed to assess the reliability of predictive models in safety-critical scenarios. Specifically, we curate two publicly available datasets of high-resolution human retina images exhibiting varying degrees of diabetic retinopathy, a medical condition that can lead to blindness, and use them to design a suite of automated diagnosis tasks that require reliable predictive uncertainty quantification. We use these tasks to benchmark well-established and state-of-the-art Bayesian deep learning methods on task-specific evaluation metrics. We provide an easy-to-use codebase for fast and easy benchmarking following reproducibility and software design principles. We provide implementations of all methods included in the benchmark as well as results computed over 100 TPU days, 20 GPU days, 400 hyperparameter configurations, and evaluation on at least 6 random seeds each.

1 Introduction

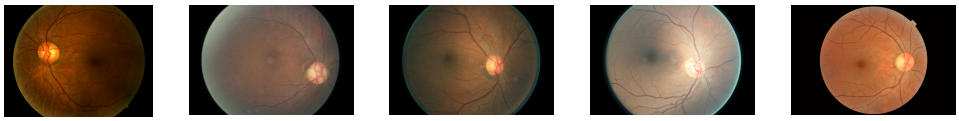
Bayesian deep learning has been applied successfully to a wide range of real-world prediction problems such as *medical diagnosis* [8, 28, 36, 65], *computer vision* [29, 30, 32], *scientific discovery* [37, 42], and *autonomous driving* [2, 17, 27, 30–32, 41].

Despite the demonstrated usefulness of Bayesian deep learning for such practical applications and a growing literature on inference methods [6, 14, 19, 21, 46, 47, 52, 61, 66], there exists no standardized benchmarking task that reflects the complexities and challenges of safety-critical real-world tasks while adequately accounting for the reliability of models' predictive uncertainty estimates.

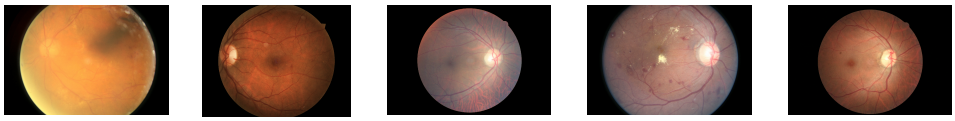
To make meaningful progress in the development and successful deployment of reliable Bayesian deep learning methods, we need easy-to-use benchmarking tasks that reflect the real world and hence serve as a legitimate litmus test for practitioners that aim to deploy their models in safety-critical

[*]Equal contribution.
[†]Corresponding authors: neil.band@cs.ox.ac.uk and tim.rudner@cs.ox.ac.uk.

Bayesian Deep Learning workshop, NeurIPS 2021.



(a) Retina images exhibiting non-sight-threatening diabetic retinopathy ($y = 0$).



(b) Retina images exhibiting sight-threatening diabetic retinopathy ($y = 1$).

Figure 1: Samples of retina scans from the EyePACS dataset showing varying degrees of diabetic retinopathy.

settings. Further, such tasks ought to be usable without the extensive domain expertise often necessary for appropriate experiment design and data preprocessing. Lastly, any such benchmarking task must include evaluation methods that test for predictive performance and assess different properties of models' predictive uncertainty estimates, while taking into account application-specific constraints.

In this paper, we propose a set of realistic safety-critical downstream tasks that respect these desiderata and use them to benchmark well-established and state-of-the-art Bayesian deep learning methods. To do so, we consider the problem of using machine learning to detect diabetic retinopathy, a medical condition considered the leading cause of vision impairment and blindness [54]. Unlike in prior works on diabetic retinopathy detection, the benchmarking tasks presented in this paper are specifically designed to assess the reliability of machine learning models and the quality of their predictive uncertainty estimates using both aleatoric and epistemic uncertainty estimates.

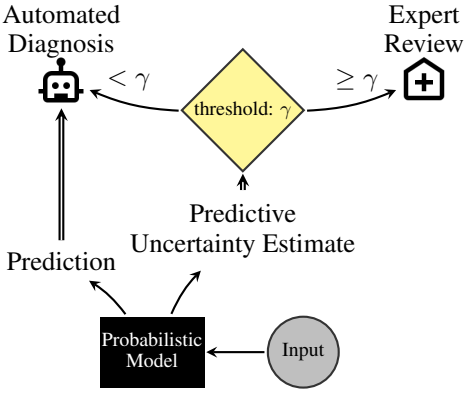


Figure 2: Automated diagnosis pipeline: For a given input, a model provides a prediction and a corresponding uncertainty estimate; if the uncertainty estimate is below a certain reference threshold γ (indicating a low degree of uncertainty) the diagnosis is processed without further review; otherwise, it is referred to a medical expert.

Medical diagnosis problems are particularly well-suited to assess reliability due to the severe harm caused by predictive models that make confident but poor predictions (for example, when a disease is not recognized). As a general desideratum, we want a model's predictive uncertainty to correlate with the correctness of its predictions. Good predictive uncertainty estimates can be a fail-safe against incorrect predictions. If a given data point might result in an incorrect prediction because it is meaningfully different from data in the training set—for example, because it shows signs of the disease not captured there, exhibits visual artifacts, or was obtained using different measurement devices—a good predictive model will express a high level of predictive uncertainty and flag the example for further review by a medical expert.

Contributions. We present an easy-to-use, expert-guided, open-source *suite of diabetic retinopathy detection benchmarking tasks* for Bayesian deep learning. In particular, we design safety-critical downstream tasks from publicly available datasets. On these downstream tasks, we evaluate well-established and state-of-the-art Bayesian and non-Bayesian methods on a set of task-specific reliability and performance metrics. Lastly, we provide a modular and extensible implementation of the benchmarking tasks and methods, as well as pre-trained models obtained from an extensive hyperparameter optimization over more than 400 total configurations and evaluation, using over 100 TPU days and 20 GPU days of compute. Code to reproduce our results and benchmark new methods is available at:

```
github.com/google/uncertainty-baselines/.../diabetic_retinopathy_detection.
```

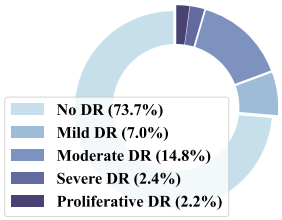
2 Downstream Benchmarking Tasks for Diabetic Retinopathy Detection

In this section, we present two real-world scenarios in diabetic retinopathy detection and describe how we merge two publicly available datasets to design corresponding prediction tasks.

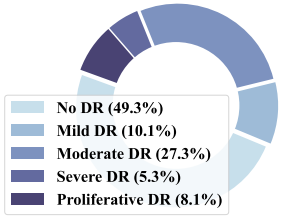
2.1 Data and Preprocessing

EyePACS Dataset. We construct training datasets for different tasks from the EyePACS dataset, previously used for the Kaggle Diabetic Retinopathy Detection Challenge [13]. It contains high-resolution labeled images of human retinas exhibiting varying degrees of diabetic retinopathy. The dataset consists of 35,126 training, 10,906 validation, and 42,670 test images, each an RGB image of a human retina graded by a medical expert on the following scale: 0 (no diabetic retinopathy), 1 (mild diabetic retinopathy), 2 (moderate diabetic retinopathy), 3 (severe diabetic retinopathy), and 4 (proliferative diabetic retinopathy).

APTOS Dataset. To construct tasks that assess model performance under distribution shift, we use the APTOS 2019 Blindness Detection dataset [3]. The dataset also contains labeled images of human retinas exhibiting varying degrees of diabetic retinopathy, but was collected in India, from a different patient population, using different medical equipment. We use 80% of the images (2,929 images) as a test set and the other 20% (733 images) as a secondary validation set. Moreover, the images are significantly noisier than the images in the EyePACS dataset, with distinct visual artifacts (cf. Figure 7, Appendix A.8). Each image was graded on the same 0-to-4 scale as the EyePACS dataset.



(a) EyePACS [13].



(b) APTOS [3].

Figure 3: Data class labels.

Prediction Targets. We follow Leibig et al. [36] and binarize all examples from both the EyePACS and APTOS datasets by dividing the classes up into sight-threatening diabetic retinopathy—defined as moderate diabetic retinopathy or worse (classes $\{2, 3, 4\}$)—and non-sight-threatening diabetic retinopathy—defined as no or mild diabetic retinopathy (classes $\{0, 1\}$). By international guidelines, this is the threshold at which a case should be referred to an ophthalmologist [64]. Example EyePACS retina images from the two classes are shown in Figure 1. Reflecting real-world challenges, the datasets are unbalanced—e.g., for EyePACS, only 19.6% of the training set and 19.2% of the test set have a positive label—and images have visual artifacts and noisy labels (some labels are incorrect).

Data Preprocessing. Data preprocessing on examples from both the EyePACS and APTOS datasets follows the winning entry of the Kaggle Challenge [13]: Images are rescaled such that retinas have a radius of 300 pixels, are smoothed using local Gaussian blur, and finally, are clipped to 90% size to remove boundary effects. Examples of original and corresponding processed images are provided in Figure 6 (Appendix A.8). We conduct an empirical study investigating how varying the strength of the Gaussian blur smoothing affects downstream performance and uncertainty quality in Appendix B.4.

2.2 Diabetic Retinopathy Detection under Severity Shift

Diabetes and diabetes-related illnesses such as diabetic retinopathy are becoming widespread. Yet cases of sight-threatening diabetic retinopathy are still relatively rare, and scans of retinas exhibiting signs of no or mild diabetic retinopathy are more easily obtainable. As a result, predictive models for detecting diabetic retinopathy may be trained on only a very small number of retina images showing signs of severe or proliferative retinopathy.

We design a prediction task that simulates this setting and allows us to assess the reliability of predictive models when they are evaluated on images that have been assigned a severity higher than any encountered in the training data. Specifically, we train models only on retina images showing signs of at most moderate diabetic retinopathy and evaluate them on retina images showing signs of severe or proliferative diabetic retinopathy. Given that many signs of moderate diabetic retinopathy are similar in appearance to signs of severe or proliferative diabetic retinopathy (just weaker), we would expect a good predictive model to be able to correctly classify the latter, but to exhibit increased predictive uncertainty. There are certain features of diabetic retinopathy progression that are unique to

more severe cases, such as vitreous hemorrhage, or bleeding into the vitreous humor [12]. However, we consider uncertainty-aware downstream tasks that tolerate such unfamiliar cases (cf. Section 2.4).

In this *Severity Shift* task, we partition the EyePACS dataset into a subset containing all retina images labeled as no, mild, or moderate diabetic retinopathy (original classes $\{0, 1, 2\}$) and a subset of retina images labeled as severe or proliferative diabetic retinopathy (original classes $\{3, 4\}$). Next, the samples in each subset are binarized (cf. Section 2.1): The subset of retina images showing signs of at most moderate diabetic retinopathy (subset "moderate") contains images of binarized classes $\{0, 1\}$; and the subset of retina images showing signs of severe or proliferative diabetic retinopathy (subset "severe") only contains the binarized class 1. This results in 33,545 images in the training set, and 40,727 and 3,524 images in the in-domain and distributionally shifted evaluation sets, respectively.

2.3 Diabetic Retinopathy Detection under Country Shift

Similar to the scarcity of scans of sight-threatening diabetic retinopathy, the availability of retina scans is limited in countries without widespread screening. Hence, a predictive model may be trained on images collected in the United States—where many scans are performed—and used to evaluate scans from another country, where scans are rarer and performed using different medical devices.

We design a prediction task that simulates this setting and allows us to evaluate the reliability of predictive models when the training and test data are not obtained from the same patient population nor collected with the same medical equipment. In this *Country Shift* task, we train models on retina images from the EyePACS dataset and evaluate them on retina images from the APTOS dataset. We use the entire training and test data provided in the EyePACS dataset and convert the task into binary classification as described in Section 2.1. This results in 35,126 images in the training set, and 42,670 and 2,929 images in the in-domain and distributionally shifted evaluation sets, respectively.

2.4 Downstream Task: Selective Prediction and Expert Referral

In real-world settings where the evaluation data may be sampled from a shifted distribution, incorrect predictions may become increasingly likely. To account for that possibility, predictive uncertainty estimates can be used to identify datapoints where the likelihood of an incorrect prediction is particularly high and refer them for further review as described in Figure 2. We consider a corresponding selective prediction task, where the predictive performance of a given model is evaluated for varying expert referral rates. That is, for a given referral rate of $\tau \in [0, 1]$, a model's predictive uncertainty is used to identify the τ proportion of images in the evaluation set for which the model's predictions are most uncertain. Those images are referred to a medical professional for further review, and the model is assessed on its predictions on the remaining $(1 - \tau)$ proportion of images. By repeating this process for all possible referral rates and assessing the model's predictive performance on the retained images, we estimate how reliable it would be in a safety-critical downstream task, where predictive uncertainty estimates are used in conjunction with human expertise to avoid harmful predictions.

Importantly, selective prediction tolerates out-of-distribution examples. For example, even if unfamiliar vitreous hemorrhages appear in certain *Severity Shift* images (cf. Section 2.2), a model with reliable uncertainty estimates will perform better in selective prediction by assigning these images high epistemic (and predictive) uncertainty, therefore referring them to an expert at a lower τ. Appendix A.6 discusses best- and worst-case uncertainty estimates for the selective prediction task.

To assess how well different models' predictive uncertainty estimates can be used to separate correct from incorrect diagnoses, we perform selective prediction on three different evaluation settings for the prediction problems described in Sections 2.2 and 2.3 to account for the possibility that the evaluation dataset may contain samples from the in-domain distribution, a shifted distribution, or both.

2.5 Model Diagnostic: Predictive Uncertainty Histograms

We may also investigate how a model's predictive uncertainty estimates vary with respect to the ground-truth clinical label (0-to-4). For each task (*Country* or *Severity Shift*) and each uncertainty quantification method (cf. Section 5), we bin examples by their ground-truth clinical label. Then, for each (task, method, clinical label) tuple, we plot the distribution of predictive uncertainty estimates for correctly and incorrectly predicted examples (in blue and red, respectively). See Appendix B.1 for further setup details and plots for both tasks. A model that produces reliable uncertainty estimates

should assign low predictive uncertainty to examples that it classifies correctly (the blue distribution should have most of its mass near $x = 0$) and high predictive uncertainty to examples that it classifies incorrectly (the red distribution should have its mass concentrated at a higher x-value).

3 Related Work

This benchmark builds on prior works that demonstrated the usefulness of predictive uncertainty estimates in diabetic retinopathy detection and related downstream tasks [36]. We significantly extend the empirical evaluation in Leibig et al. [36] by designing new prediction problems and corresponding safety-critical downstream tasks for diabetic retinopathy detection, benchmarking a wide array of Bayesian deep learning methods, and providing a modular, extensible, and easy-to-use codebase. We also significantly extend Filos et al. [16] (of which this paper is a direct extension; with contributions from some of the authors), which does not consider severity shifts, only compares two variational inference methods, uses an outdated neural network architecture (with only $\approx$10% of the parameters of the ResNet-50 architecture used in this work), and considers only a small subset of the evaluation procedures included in this benchmark (cf. Appendix B.4 for the full set of results).

Previous works have evaluated methods by predictive performance and quality of their predictive uncertainty estimates on curated datasets such as CIFAR-10 and FashionMNIST [47, 48, 24, 52]. Some prior works provide datasets and benchmarks for robustness and uncertainty quantification in real-world settings but have significant shortcomings. Le et al. [35] considers object detection using a real-world dataset [20] but benchmarks only two methods, neither of which can quantify epistemic uncertainty (cf. Section 4), and does not consider distribution shifts. Other works [5, 15] use methods which quantify both epistemic and aleatoric uncertainty, and consider distribution shifts, but use performance metrics which do not assess quality of uncertainty estimates, such as average precision and log-likelihood (cf. Section 6.3). Finally, Koh et al. [33] considers real-world datasets in domain adaptation problems, but restrictively assumes that the training data is composed of multiple training distributions with domain labels, and does not take into account models' predictive uncertainty.

In contrast, our benchmark **(i)** considers real-world safety-critical tasks and accompanying uncertainty-aware metrics in an important application domain, **(ii)** is composed of large amounts of high-dimensional data (>80 GB), **(iii)** compares a larger set of methods than prior works and incorporates both aleatoric and epistemic uncertainty, and is implemented in adherence to the *Uncertainty Baselines* repository[3] practices for easy future use and extension, making it easier to benchmark other Bayesian deep learning methods not only on the tasks presented but also on a range of other datasets.

4 Uncertainty Estimation

Predictive models' total uncertainty can be decomposed into aleatoric and epistemic uncertainty. A model's aleatoric uncertainty is an estimate of the uncertainty inherent in the data (e.g., due to noisy inputs or targets), whereas a model's epistemic uncertainty is an estimate of the uncertainty due to constraints on the model (e.g., due to model misspecification) or the training process (e.g., due to convergence to bad local optima) [10]. Optimal uncertainty estimates would be perfectly correlated with the model error. Hence, because both aleatoric and epistemic uncertainty may contribute to an incorrect prediction, total uncertainty is our uncertainty measure of choice. For a model with stochastic parameters $\boldsymbol{\Theta}$, pre-likelihood outputs $f(\mathbf{X}; \boldsymbol{\Theta})$, and a likelihood function $p(\mathbf{y}_* \mid \mathbf{x}_*; \boldsymbol{\theta})$, the model's predictive uncertainty can be decomposed as

$$\underbrace{\mathcal{H}(\mathbb{E}[p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}))])}_{\text{Total Uncertainty}} = \underbrace{\mathbb{E}[\mathcal{H}(p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta})))]}_{\text{Aleatoric Uncertainty}} + \underbrace{\mathcal{I}(\mathbf{y}_*; \boldsymbol{\Theta})}_{\text{Epistemic Uncertainty}}, \tag{1}$$

where the expectation is taken with respect to the distribution over the model parameters, $\mathcal{H}(\cdot)$ is the entropy functional, and $\mathcal{I}(\mathbf{y}_*; \boldsymbol{\Theta})$ is the mutual information between the model parameters and its predictions [9, 55].

In binary classification settings with classes $\{0, 1\}$, the total predictive uncertainty is given by

$$\mathcal{H}(\mathbb{E}[p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}))]) = -\sum_{c \in \{0,1\}} \mathbb{E}[p(\mathbf{y}_* = c \mid f(\mathbf{x}_*; \boldsymbol{\theta}))] \log \mathbb{E}[p(\mathbf{y}_* = c \mid f(\mathbf{x}_*; \boldsymbol{\theta}))], \tag{2}$$

[3] See `https://github.com/google/uncertainty-baselines`.

where $f(\mathbf{x}_*; \boldsymbol{\theta})$ are logits and $p(\mathbf{y}_* = c \mid f(\mathbf{x}_*; \boldsymbol{\theta}))$ is a binary cross-entropy likelihood function. The total predictive uncertainty is high when either the aleatoric uncertainty is high (e.g., because the input is noisy), or when the epistemic uncertainty is high (e.g., because the model has many possible explanations for the input). In practice, the total predictive uncertainty $\mathcal{H}(\mathbb{E}[p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}))])$ is computed with a Monte Carlo estimator $\mathbb{E}[p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}))] \approx \frac{1}{S} \sum_i^S p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}^{(i)}))$, where $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^S$ are sampled from some distribution over the network parameters, and $p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta}^{(i)}))$ is a deterministic forward pass given parameter realization $\boldsymbol{\theta}^{(i)}$.

5 Methods

Estimating a model's predictive uncertainty in terms of both aleatoric and epistemic uncertainty requires a *distribution over model predictions*. Such a distribution over model predictions can be obtained by treating the parameters of a neural network as random variables and inferring a posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$—a distribution over the network parameters conditioned on a set of training data $\mathcal{D} = (\mathbf{X}_\mathcal{D}, \mathbf{y}_\mathcal{D})$—according to the rules of Bayesian inference. Neural networks with such distributions over the network parameters—referred to as a Bayesian neural networks (BNN)—define distribution over predictions and are able to capture both aleatoric and epistemic uncertainty [18, 38, 45]. Unfortunately, computing a posterior distribution over the parameters of a neural network according to the rule of Bayesian inference is analytically intractable and requires the use of approximate inference methods [18, 21, 25, 45, 50]. Below, we describe the baseline and state-of-the-art methods for which we implement standardized and optimized runscripts in the benchmark, which are readily extensible for experimentation and deployment in application settings.

5.1 Maximum A Posteriori Estimation in Bayesian Neural Networks

As an alternative to inferring a posterior distribution over neural network parameters, maximum a posteriori (MAP) estimation yields network parameter values equal to the mode of the exact posterior distribution. For a prior distribution over network parameters with zero mean and precision λ, the maximum a posteriori estimate is equal to the solution of the ℓ_2-regularized optimization problem $\arg\min_{\boldsymbol{\theta}}\{-\log p(\mathbf{y}_\mathcal{D} \mid f(\mathbf{X}; \boldsymbol{\theta})) + \lambda ||\boldsymbol{\theta}||_2^2\}$, and as such is equivalent to parameter values obtained by training a neural network with weight decay. Since MAP estimation yields a point estimate of the MAP parameters, the MAP solution defines a deterministic neural network and is thus unable to capture any epistemic uncertainty. In classification tasks, they represent aleatoric uncertainty estimates via the predicted class probabilities [31]. We use neural networks with MAP estimation as a baseline for the benchmark.

5.2 Variational Inference in Bayesian Neural Networks

Variational inference is an approximate inference method that seeks to sidestep the intractability of exact posterior inference over the network parameters by framing posterior inference as a variational optimization problem. In particular, variational inference in neural networks seeks to find an approximation to the posterior distribution over parameters by solving the optimization problem

$$\text{argmax}_{q \in \mathcal{Q}}\{\mathbb{E}_q[\log p(\mathbf{y}_\mathcal{D} \mid f(\mathbf{X}_\mathcal{D}; \boldsymbol{\theta}))] - \mathbb{D}_{\text{KL}}(q \,\|\, p)\}, \tag{3}$$

where $\mathcal{Q}$ is a variational family of distributions and p is a prior distribution.

Gaussian Mean-Field Variational Inference. If $p \doteq p_{\boldsymbol{\Theta}}$ and $q \doteq q_{\boldsymbol{\Theta}}$ are distributions over parameters, $\mathcal{Q}$ is the family of mean-field (i.e., fully-factorized) Gaussian distributions, and the prior distribution over parameters $p_{\boldsymbol{\Theta}}$ is also a diagonal Gaussian, the resulting variational objective is amenable to stochastic variational inference and can be optimized using stochastic gradient methods [6, 21, 25, 26, 59]. Henceforth, we refer to BNN inference methods that make these variational assumptions as mean-field variational inference. To optimize this objective, the expectation is estimated using Monte Carlo sampling and the network parameters are reparameterized as $\boldsymbol{\Theta} \doteq \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Throughout, we use the flipout estimator [61] to reduce the variance of the gradient estimates, and temper the Kullback-Leibler divergence term in the variational objective [62].

Radial-Gaussian Mean-Field Variational Inference. Radial-Gaussian mean-field variational inference [14] uses the same variational objective, prior, and variational distribution as standard Gaussian mean-field variational inference, but uses an alternative gradient estimator to obtain an

improved signal-to-noise ratio in the gradient estimates. Specifically, the network parameters are reparameterized as $\boldsymbol{\Theta} \doteq \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \frac{\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||_2} \cdot |r|$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $r \sim \mathcal{N}(0, 1)$.

Function-Space Variational Inference. Rudner et al. [52] proposed a tractable function-space variational objective for Bayesian neural networks. If $p \doteq p_{f([\mathbf{X}_\mathcal{D}, \mathbf{X}_\mathcal{I}]; \boldsymbol{\Theta})}$ and $q \doteq q_{f([\mathbf{X}_\mathcal{D}, \mathbf{X}_\mathcal{I}]; \boldsymbol{\Theta})}$ are distributions over functions evaluated at the training inputs $\mathbf{X}_\mathcal{D}$ and at a set of inducing inputs $\mathbf{X}_\mathcal{I}$, $\mathcal{Q}$ is the family of distributions over functions induced by some distribution over network parameters, and the Kullback-Leibler divergence between distributions over functions evaluated at $[\mathbf{X}_\mathcal{D}, \mathbf{X}_\mathcal{I}]$ is approximated by a linearization of the neural network mapping, then the resulting variational objective is amenable to stochastic variational inference [52, 53]. In our benchmark, we define a Gaussian mean-field distribution over the final layer of the neural network and reparameterize the parameters as $\boldsymbol{\Theta} \doteq \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Monte Carlo Dropout. Gal and Ghahramani [19] showed that training a deterministic neural network with ℓ_2- and dropout regularization [57], that is, solving the optimization problem $\arg\min_{\boldsymbol{\theta}} \{-\mathbb{E}_q[\log p(\mathbf{y}_\mathcal{D} \mid f(\mathbf{X}; \boldsymbol{\theta}))] + \lambda ||\boldsymbol{\theta}||_2^2\}$, where $q_{\boldsymbol{\Theta}}$ is the distribution over parameters obtained by applying dropout with a given dropout rate, approximately corresponds to variational inference in a Bayesian neural network. To sample from the approximate posterior predictive distribution, dropout is applied to the deterministic network parameters. To optimize the objective above, the expectation is estimated using a single Monte Carlo sample (i.e., by applying dropout).

Rank-1 Parameterization. Dusenberry et al. [11] propose a rank-1 parameterization of Bayesian neural networks, where each weight matrix involves only a distribution on a rank-1 subspace, that is, each stochastic weight matrix is defined as $\mathbf{W}_k' = \mathbf{W}_k \odot \mathbf{r}_k \mathbf{s}_k^\top$, where $\mathbf{W}_k$ is a deterministic set of weights, and $\mathbf{r}_k$ and $\mathbf{s}_k$ are random vectors of parameters. Variational distributions over $\mathbf{r}_k$ and $\mathbf{s}_k$ and a Dirac delta distribution over $\mathbf{W}_k$ for all layers k are obtained by optimizing a variational objective.

5.3 Model Ensembling

Deep Ensembles. A deep ensemble [34] is a mixture of multiple independently-trained deterministic neural networks. As such, unlike BNNs, deep ensembles do not explicitly infer a distribution over the parameters of a single neural network. Instead, they marginalize over multiple deterministic models to obtain a predictive distribution that captures both aleatoric and epistemic uncertainty. We construct deep ensembles from multiple MAP neural networks trained with different random seeds.

Ensembles of Bayesian Neural Networks. Ensemble of Bayesian neural networks [16, 52, 56] are mixtures of multiple independently-trained Bayesian neural networks. They can account for the possibility that any individual approximate posterior distribution obtained via variational inference may be a poor approximation to the exact posterior distribution and may hence yield a poor predictive distribution. A common issue in the Bayesian deep learning literature is that ensembles are frequently compared to single models, often due to computational constraints. In our benchmark, we provide a unified comparison and construct ensembles for all predictive models, including BNNs.

6 Benchmark

6.1 Evaluation Protocol

Network Architecture. We use a ResNet-50 architecture for all experiments [22]. A sigmoid transformation is applied to the final linear layer of all networks to obtain class probabilities corresponding to the outcomes of the binary classification problems described in Sections 2.2 and 2.3.

Validation Data and Hyperparameter Tuning. Reliable uncertainty estimation on data points from shifted distributions is the central challenge for Bayesian deep learning methods. In training and evaluating such methods, practitioners must decide how they should choose validation data: specifically, in which settings they would benefit from using "out-of-distribution" data points for hyperparameter tuning. We consider two real-world settings: **(i)** No distributionally shifted data is available during hyperparameter tuning. This setting reflects scenarios in which practitioners do not know what data or distributional shift they might encounter during deployment and hence cannot make assumptions about it at training time. **(ii)** Shifted validation data *is* available for hyperparameter tuning. This setting reflects scenarios in which practitioners may intend to train a model on data collected from one subpopulation and deploy it on data collected from another subpopulation, but are able to acquire a small number of examples from the deployment subpopulation for use in

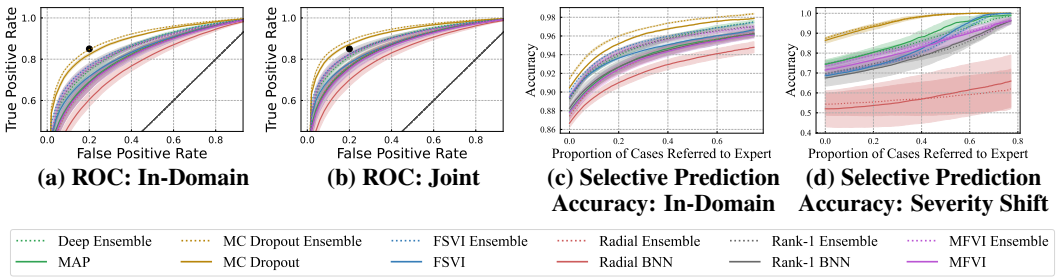


Figure 4: Severity Shift. We jointly assess model predictive performance and uncertainty quantification on the in-domain test dataset composed only of cases with either no, mild, or moderate diabetic retinopathy, and the *Severity Shift* evaluation set composed only of severe and proliferative cases. **Left:** The receiver operating characteristic curve (ROC) for in-domain diagnosis (**a**) and for a joint dataset composed of examples from both the in-domain and *Severity Shift* evaluation sets (**b**). The dot in **black** denotes the NHS-recommended 85% sensitivity and 80% specificity ratios [63]. **Right:** Selective prediction on accuracy in the in-domain (**c**) and *Severity Shift* (**d**) settings. Shading denotes standard error computed over six random seeds. See Section 6.2.

tuning to improve generalization. Prior works on out-of-distribution detection [23] and uncertainty quantification [39] have considered setting **(ii)**, but have not provided a comparative analysis, which would inform practitioners on when they ought to collect shifted validation data for tuning. We rigorously investigate the two settings across downstream tasks in Appendix B.4. In the main paper, we report results for models tuned under setting **(i)**.

The aim of this benchmark is to adequately represent the challenges of real-world distributional shift, and rigorously assess the reliability of (Bayesian) uncertainty quantification in deep learning. Our selective prediction downstream tasks demonstrate two real-world use cases:

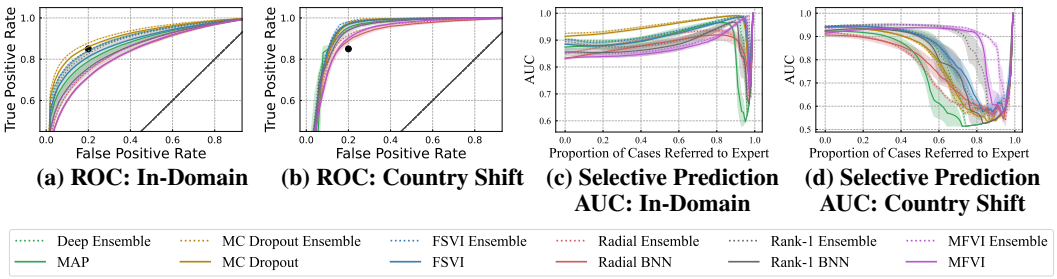
- **Tuning Referral Thresholds.** On the *Severity Shift* task, models demonstrate reasonable uncertainty estimates: predictive performance increases monotonically with an increasing referral rate τ. Therefore, practitioners can infer which referral rate will lead to a desired predictive performance, or infer the performance for a predetermined referral rate (i.e., respecting a budget of expert time).
- **Detecting Low-Quality Predictive Uncertainty.** On the *Country Shift* task, most methods fail: predictive performance on the shifted dataset *declines* as τ increases, indicating that the quality of uncertainty estimates is no better than random referral. Importantly, this failure is *not reflected* in the standard performance measure for retinopathy diagnosis, the receiver operating characteristic (ROC) curve [36]—the area under the ROC curve (AUC) is *higher* on the shifted evaluation dataset than the in-domain dataset—meaning that a practitioner using only AUC might wrongly conclude that these models would perform well as part of an automated diagnosis pipeline (cf. Figure 2) on distributionally shifted data.

For each method, we assess both AUC and accuracy as a function of the referral rate τ, evaluating the models' predictions for the $(1 - \tau)$ proportion of cases on which they are most certain, as indicated by their predictive uncertainty estimates. We additionally examine predictive uncertainty histograms for each task, method, and ground-truth clinical label (cf. Section 2.5, Appendix B.1) to determine if methods have particularly good or bad uncertainty estimates at particular severity levels. We also investigate other metrics to assess the reliability of models' uncertainty estimates, including *expected calibration error* and *out-of-distribution detection AUC*, in Appendix B.4.

6.2 Severity Shift

On the *Severity Shift* task (Figure 4, Table 2), models are trained on EyePACS images that show signs of at most moderate diabetic retinopathy. We assess their ability to generalize to images showing signs of severe or proliferative retinopathy. Surprisingly, we find that models generalize well from cases with no worse than moderate diabetic retinopathy (in-domain) (Figure 4(**a**)) to severe cases (Figure 4(**b**)), improving their AUC under the distribution shift.

Methods Generalize Reasonably Well Under Severity Shift. Reliable predictive uncertainty estimates correlate with predictive error, and therefore we would expect a model's performance (e.g., measured in terms of accuracy or AUC) to increase as more examples on which the model exhibits high uncertainty are referred to an expert. On both the in-domain and *Severity Shift* evaluation sets



(a) ROC: In-Domain **(b) ROC: Country Shift** **(c) Selective Prediction AUC: In-Domain** **(d) Selective Prediction AUC: Country Shift**

Figure 5: Country Shift. We jointly assess model predictive performance and uncertainty quantification on both in-domain and distributionally shifted data. **Left:** The *receiver operating characteristic curve (ROC)* for in-population diagnosis on the EyePACS [13] test set (**a**) and for changing medical equipment and patient populations on the APTOS [3] test set (**b**). The dot in **black** denotes the NHS-recommended 85% sensitivity and 80% specificity ratios [63]. **Right:** *selective prediction* on AUC in the EyePACS [13] (**c**) and the APTOS [3] (**d**) settings. Shading denotes standard error computed over six random seeds. See Section 6.3.

(Figures 4(**c**) and (**d**)), models demonstrate reasonable uncertainty in that accuracy monotonically increases as τ increases. This highlights two ways that practitioners may use selective prediction to prepare models for a real-world deployment in the presence of potential distribution shifts. First, given a performance target (e.g., $\geq 95\%$ accuracy) the referral curve can be used to determine the minimum τ achieving this target, estimating a medical experts' workload. Second, for a maximum acceptable referral rate (e.g., a clinic has medical experts to handle referral of $\tau \leq 20\%$ of patients) the referral curve can be used to determine the optimal τ value and the corresponding performance. For monotonically increasing referral curves, the optimal τ is uniquely the maximum acceptable referral rate.

Taking into Account Epistemic Uncertainty Can Improve Reliability. On the *Severity Shift* task (Figure 4(**d**)) many models achieve near-perfect accuracy well before all examples have been referred. For example, MC DROPOUT, which incorporates both epistemic and aleatoric uncertainty (cf. Section 4), achieves 100% predictive accuracy near the 50% referral rate—nearly 20% lower than the referral rate at which a deterministic neural network (MAP), which only represents aleatoric uncertainty, achieves this level of accuracy. Other variational inference methods underperform MAP, underscoring the importance of continued work on approximate inference in BNNs.

Predictive Uncertainty Histograms Identify Harmful Uncertainty Quantification. In Figure 8 (Appendix B.1), we find that MAP, RANK-1, and MFVI generate worse uncertainty estimates than other methods on the shifted data (labels 3 and 4); many of their incorrect predictions are assigned low predictive uncertainty (i.e., the red distribution is concentrated near 0). These false negatives with low uncertainty are particularly dangerous in automated diagnosis settings (cf. Figure 2), as a medical expert would not be able to catch the model's failure to recognize the condition.

6.3 Country Shift

In the *Country Shift* task (Figure 5, Table 1), we consider the performance of models trained on the US EyePACS [13] dataset and evaluated under distributional shift, on the Indian APTOS dataset [3]. The left two plots of Figure 5 present the ROC curves of methods evaluated on the in-domain (**a**) and *Country Shift* (**b**) evaluation datasets. The black dot in Figures 5(**a**) and (**b**) denotes the minimum sensitivity–specificity threshold for the deployment of automated diabetic retinopathy diagnosis systems set by the British National Health Service (NHS) [63]. On the in-domain test dataset, only the MC DROPOUT variants meet the NHS standard; on the APTOS dataset, essentially all methods surpass the standard.[4] Hence, practitioners using only the ROC curve and its AUC (cf. Table 1) might conclude that their model generalizes under the distribution shift although the ROC curve provides no information on the application of uncertainty estimates to real-world scenarios (cf. Figure 2).

Selective Prediction Can Indicate Failures in Uncertainty Estimation. Unlike the ROC curve, the selective prediction metric conveys how a model would perform in an automated diagnosis pipeline in which the reliability of models' uncertainty estimates directly impacts performance

[4]We investigate this in Appendix B.4 and find that class proportions do not account for the improved predictive performance on APTOS, implying other contributing factors such as demographics or camera type.

Table 1: Country Shift. Prediction and uncertainty quality of baseline methods in terms of the area under the receiver operating characteristic curve (AUC) and classification accuracy, as a function of the proportion of data referred to a medical expert. All methods are tuned on in-domain validation AUC, and ensembles have $K = 3$ constituent models (true for all subsequent tables unless specified otherwise). On in-domain data, MC DROPOUT performs best across all thresholds. On distributionally shifted data, no method consistently performs best.

	No Referral		50% Data Referred		70% Data Referred	
Method	**AUC (%) ↑**	**Accuracy (%) ↑**	**AUC (%) ↑**	**Accuracy (%) ↑**	**AUC (%) ↑**	**Accuracy ↑**
EyePACS Dataset (In-Domain)						
MAP (Deterministic)	87.4±1.0	88.6±0.6	91.1±1.4	95.9±0.3	94.9±0.8	96.5±0.2
MFVI	83.3±0.1	85.7±0.1	85.5±0.5	94.5±0.1	88.2±0.5	95.9±0.1
RADIAL-MFVI	83.2±0.4	74.2±3.8	88.9±0.7	81.8±4.7	91.2±1.0	83.8±4.3
FSVI	88.5±0.1	89.8±0.0	91.0±0.3	96.4±0.0	94.3±0.2	97.2±0.0
MC DROPOUT	91.4±0.1	90.9±0.0	95.3±0.2	97.4±0.0	97.4±0.1	98.1±0.0
RANK-1	85.6±1.1	87.7±0.6	87.1±1.8	95.3±0.4	90.9±1.5	96.4±0.3
DEEP ENSEMBLE	90.3±0.1	90.3±0.2	91.7±0.5	97.2±0.0	95.0±0.4	97.9±0.0
MFVI ENSEMBLE	85.4±0.0	87.8±0.0	86.3±0.3	95.4±0.0	89.2±0.3	96.7±0.0
RADIAL-MFVI ENSEMBLE	85.1±0.0	77.8±1.7	91.6±0.3	87.9±1.7	94.0±0.3	90.5±1.6
FSVI ENSEMBLE	90.3±0.1	90.6±0.0	92.1±0.2	97.1±0.0	95.2±0.1	97.8±0.0
MC DROPOUT ENSEMBLE	**92.5±0.0**	**91.6±0.0**	**95.8±0.1**	**97.8±0.0**	**97.7±0.1**	**98.4±0.0**
RANK-1 ENSEMBLE	89.3±0.7	89.3±0.4	88.5±1.1	96.9±0.3	91.6±1.1	97.6±0.2
APTOS 2019 Dataset (Population Shift)						
MAP (Deterministic)	92.2±0.2	86.2±0.4	80.1±2.8	87.6±1.1	55.4±3.3	85.4±0.9
MFVI	91.4±0.2	84.1±0.3	93.8±0.3	92.1±0.4	93.0±0.5	92.7±0.4
RADIAL-MFVI	90.7±0.5	71.8±3.5	82.0±2.0	81.5±2.1	66.4±1.7	85.9±0.7
FSVI	94.1±0.1	87.6±0.4	90.6±0.7	90.7±0.6	77.2±3.6	89.8±0.2
MC DROPOUT	94.0±0.2	86.8±0.2	87.4±0.3	88.1±0.2	65.3±1.3	88.2±0.3
RANK-1	92.5±0.2	86.2±0.4	90.1±1.9	91.4±0.8	75.1±6.0	89.5±1.2
DEEP ENSEMBLE	94.2±0.2	87.5±0.1	91.2±1.4	92.4±0.7	67.4±5.6	90.1±0.9
MFVI ENSEMBLE	93.2±0.1	87.0±0.2	**94.9±0.3**	**93.7±0.3**	**94.2±0.2**	**94.0±0.3**
RADIAL-MFVI ENSEMBLE	91.8±0.2	71.7±1.9	81.8±1.5	82.7±1.3	65.9±3.1	87.6±0.5
FSVI ENSEMBLE	**94.6±0.0**	**88.9±0.1**	90.7±0.4	91.1±0.5	74.1±2.6	89.8±0.2
MC DROPOUT ENSEMBLE	94.1±0.1	87.6±0.1	86.8±0.2	88.0±0.1	62.3±0.3	87.7±0.2
RANK-1 ENSEMBLE	94.1±0.1	88.2±0.1	**94.8±0.3**	93.3±0.2	92.1±1.2	93.7±0.3

(cf. Figure 2). Recall that if a model generates reliable predictive uncertainty estimates, the AUC should increase as more patients with uncertain predictions are referred for expert review. This mechanism is illustrated well by the application of MFVI to the *Country Shift* task (Figure 5**(d)** and Table 1), since the AUC improves from an initial 91.4% up to 93.8% when referring 50% of the patients, but then deteriorates as the model is forced to refer patients on which it is both certain and correct. In contrast, other models' AUCs trend downwards; using uncertainty to refer patients actively hurts model performance on this shifted dataset.

Different Prediction Tasks Yield Different Method Rankings. In Figure 5(**c**), variational inference methods, including MC DROPOUT, FSVI, and DEEP ENSEMBLE, outperform MAP inference. This highlights that rankings are task-dependent, and underscores the importance of generic evaluation frameworks to enable rapid benchmarking on many tasks.

7 Conclusions

The deployment of modern machine learning models in safety-critical real-world settings necessitates trust in the reliability of the models' predictions.

To encourage the development of Bayesian deep learning methods that are capable of generating reliable uncertainty estimates about their predictions, we introduced a set of safety-critical real-world clinical prediction tasks, which highlight various shortcomings of existing uncertainty quantification methods. We demonstrate that by taking into account the quality of predictive uncertainty estimates, selective prediction can help identify whether methods might fail when deployed as part of an automated diagnosis pipeline (cf. Figure 2), whereas standard metrics such as ROC curves cannot.

While no single set of benchmarking tasks is a panacea, we hope that the tasks and evaluation methods presented in this paper will significantly lower the barrier for assessing the reliability of Bayesian deep learning methods on safety-critical real-world prediction tasks.

Acknowledgments and Disclosure of Funding

We thank Google Research for providing computational and storage resources. We thank Intel Labs for their computational support. We thank Ranganath Krishnan for his contributions to the RANK-1 implementation by porting a CIFAR-10 training script, and for sharing his expertise on variational inference in Bayesian neural networks. We thank Sebastian Farquhar for his contributions to the RADIAL-MFVI implementation, including a TensorFlow Probability distribution, code review, and feedback on tied means, ℓ_2-regularization, variance reduction, and hyperparameters. We thank Jorge Cuadros, OD, PhD (CEO of EyePACS) and Jan Brauner, MD (University of Oxford) for lending their domain expertise in task design. We thank all other contributors to the *Uncertainty Baselines* project [43] (into which this benchmark is integrated): Mark Collier, Josip Djolonga, Marton Havasi, Rodolphe Jenatton, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, and Jasper Snoek. NB and TGJR acknowledge funding from the Rhodes Trust. TGJR also acknowledges funding from Qualcomm and the Engineering and Physical Sciences Research Council (EPSRC).

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016.

[3] APTOS. APTOS 2019 Blindness Detection Dataset, 2019.

[4] Lukas Biewald. Experiment Tracking with Weights and Biases, 2020.

[5] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A Benchmark for Safe Semantic Segmentation in Autonomous Driving. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2403–2412, 2019.

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

[7] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. *https://arxiv.org/abs/1511.06349*, 2015.

[8] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *bioRxiv*, page 142760, 2018.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[10] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-Sensitive Learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[11] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2782–2792. PMLR, 13–18 Jul 2020.

[12] Jaafar El Annan and Petros E. Carvounis. Current Management of Vitreous Hemorrhage due to Proliferative Diabetic Retinopathy. *Int Ophthalmol Clin.*, 2014.

[13] EyePACS. Diabetic Retinopathy Detection Dataset, 2015.

[14] Sebastian Farquhar, Michael A. Osborne, and Yarin Gal. Radial Bayesian Neural Networks: Beyond Discrete Support In Large-Scale Bayesian Deep Learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1352–1362. Pmlr, 26–28 Aug 2020.

[15] Di Feng, Ali Harakeh, Steven L. Waslander, and Klaus Dietmayer. A Review and Comparative Study on Probabilistic Object Detection in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, page 1–20, 2021. ISSN 1558-0016.

[16] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks, 2019.

[17] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can Autonomous Vehicles Identify, Recover From, and Adapt to Distribution Shifts? In *International Conference on Machine Learning*, pages 3145–3153. Pmlr, 2020.

[18] Yarin Gal. Uncertainty in Deep Learning. *University of Cambridge*, 2016.

[19] Yarin Gal and Zoubin Ghahramani. Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, Icml 2016, pages 1050–1059, 2016.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[21] Alex Graves. Practical Variational Inference for Neural Networks. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, page 2348–2356, Red Hook, NY, USA, 2011. Curran Associates Inc.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[23] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure, 2019.

[24] Maximilian Henne, Adrian Schwaiger, Karsten Roscher, and Gereon Weiss. Benchmarking Uncertainty Estimation Methods for Deep Learning With Safety-Related Metrics, 2020.

[25] Geoffrey E. Hinton and Drew van Camp. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, Colt '93, page 5–13, New York, NY, USA, 1993. Citeseer, Association for Computing Machinery.

[26] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013. ISSN 1532-4435.

[27] Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-Aware Reinforcement Learning for Collision Avoidance. *arXiv preprint arXiv:1702.01182*, 2017.

[28] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Medical Image Analysis*, 36: 61–78, 2017.

[29] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*, pages 680–688. Ieee, 2016.

[30] Alex Kendall and Roberto Cipolla. Modelling Uncertainty in Deep Learning for Camera Relocalization. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4762–4769. IEEE, 2016.

[31] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems 30*, 2017.

[32] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[33] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of In-the-Wild Distribution Shifts, 2021.

[34] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413, 2017.

[35] Michael Truong Le, Frederik Diehl, Thomas Brunner, and Alois Knoll. Uncertainty Estimation for Deep Neural Object Detectors in Safety-Critical Applications, 2018.

[36] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging Uncertainty Information From Deep Neural Networks for Disease Detection. *Nature Scientific Reports*, 7(1):17816, 2017.

[37] Laurence Perreault Levasseur, Yashar D Hezaveh, and Risa H Wechsler. Uncertainties in Parameters Estimated with Neural Networks: Application to Strong Gravitational Lensing. *The Astrophysical Journal Letters*, 850(1):L7, 2017.

[38] David JC MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 1992.

[39] Andrey Malinin and Mark Gales. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[40] Andrey Malinin and Mark JF Gales. Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness. 2019.

[41] Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[42] Robert T McGibbon, Andrew G Taube, Alexander G Donchev, Karthik Siva, Felipe Hernández, Cory Hargus, Ka-Hei Law, John L Klepeis, and David E Shaw. Improving the Accuracy of Møller-Plesset Perturbation Theory with Neural Networks. *The Journal of Chemical Physics*, 147(16):161725, 2017.

[43] Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael Dusenberry, Sebastian Farquhar, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim G. J. Rudner, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty Baselines: Benchmarks for Uncertainty & Robustness in Deep Learning. *arXiv preprint arXiv:2106.04015*, 2021.

[44] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation. *Medical Image Analysis*, 59:101557, 2020.

[45] Radford M Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 1995.

[46] Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variance Networks: When Expectation Does Not Meet Your Expectations. *arXiv preprint arXiv:1803.03764*, 2018.

[47] Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical Deep Learning with Bayesian Principles. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 4287–4299. Curran Associates, Inc., 2019.

[48] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Advances in Neural Information Processing Systems 32*. 2019.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[50] Carsten Peterson. A Mean Field Theory Learning Algorithm for Neural Networks. *Complex systems*, 1:995–1019, 1987.

[51] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Inherent Brain Segmentation Quality Control from Fully Convnet Monte Carlo Sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2018.

[52] Tim G. J. Rudner, Zonghao Chen, and Yarin Gal. Rethinking Function-Space Variational Inference in Bayesian Neural Networks. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

[53] Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual Learning via Function-Space Variational Inference. In *ICML Workshop on Theory and Foundations of Continual Learning*. 2021.

[54] Charumathi Sabanayagam, Riswana Banu, Miao Li Chee, Ryan Lee, Ya Xing Wang, Gavin Tan, Jost B Jonas, Ecosse L Lamoureux, Ching-Yu Cheng, Barbara E K Klein, Paul Mitchell, Ronald Klein, C M Gemmy Cheung, and Tien Y Wong. Incidence and Progression of Diabetic Retinopathy: A Systematic Review. *The Lancet Diabetes & Endocrinology*, 7(2):140–149, 2021/08/27 2019.

[55] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana and Chicago, 1949.

[56] Lewis Smith and Yarin Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569. AUAI Press, 2018.

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[58] TFDS Team. TensorFlow Datasets, A Collection of Ready-to-Use Datasets. `https://www.tensorflow.org/datasets`.

[59] Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

[60] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks. *Neurocomputing*, 338:34–45, 2019. ISSN 0925-2312.

[61] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. *arXiv preprint arXiv:1803.04386*, 2018.

[62] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020.

[63] D. T. S. Widdowson. The Management of Grading Quality: Good Practice in the Quality Assurance of Grading. *Tech. Rep.*, 2016.

[64] Tien Y Wong, Jennifer Sun, Ryo Kawasaki, Paisan Ruamviboonsuk, Neeru Gupta, Van Charles Lansingh, Mauricio Maia, Wanjiku Mathenge, Sunil Moreker, Mahi M.K. Muqit, Serge Resnikoff, Juan Verdaguer, Peiquan Zhao, Frederick Ferris, Lloyd P. Aiello, and Hugh R. Taylor. Guidelines on Diabetic Eye Care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. *Opthalmology*, 2018.

[65] Daniel E Worrall, Clare M Wilson, and Gabriel J Brostow. Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 68–76. Springer, 2016.

[66] Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E Turner, José Miguel Hernández-Lobato, and Alexander L Gaunt. Deterministic Variational Inference for Robust Bayesian Neural Networks. *International Conference on Learning Representations*, 2019.

Supplementary Material

Table of Contents

Appendix A Implementation, Training, and Evaluation Details

A.1 Benchmark Software Design Principles

Reproducibility in machine learning is often hampered by the wide variety of experimental artifacts made available in papers. Perhaps the most common approach is a GitHub dump of experimental code lacking documentation and testing. This common practice fails to enforce a rigorous standard across works: for example, experiment protocol on cross-validation, access to distributionally shifted validation data, and various tweaks in optimization such as learning rate annealing.

The Diabetic Retinopathy Detection Benchmark is implemented in the open-sourced *Uncertainty Baselines* [43] repository. All models implemented in this repository conform to explicit design principles intended to facilitate easy extension and reproduction of dataset loading utilities, metrics, and evaluation.

Extensibility. Each model baseline (e.g., MAP, MC DROPOUT, FSVI) is implemented in its own self-contained experiment pipeline. This minimizes external dependencies, and therefore provides researchers and practitioners an immediate starting point for experimenting with a particular model. Datasets are implemented as lightweight wrappers around TensorFlow Datasets [58]. Users that wish to extend our benchmark with new datasets (e.g., clinical practitioners that wish to apply our methods on their own diabetic retinopathy tasks) can follow our custom implementation of the APTOS [3] data loader, which constructs the dataset from raw images and a CSV containing metadata, and applies the preprocessing used by the winner of the EyePACS Kaggle competition [13]. Dataset implementation can be found here.[5]

Framework Agnosticity. The Diabetic Retinopathy Detection Benchmark is framework-agnostic. For example, FSVI is implemented in JAX, a variant of MC Dropout is in PyTorch [49] (though we use

[5] https://github.com/google/uncertainty-baselines/tree/main/uncertainty_baselines/datasets

in this work a TensorFlow variant to simplify TPU tuning), and other models in raw TensorFlow [1]. This interoperability means that users can easily incorporate our datasets and evaluation utilities, including an arrangement of robustness and uncertainty metrics such as *selective prediction*, *out-of-distribution detection*, and *expected calibration error*.

Reproducibility. All models include testing, and all results are reported over multiple seeds. For each method (e.g., MC DROPOUT or MFVI), downstream task (*Country* and *Severity Shift*), and tuning assumption (whether or not distributionally shifted validation data is available for tuning), we test over at least 32 hyperparameter configurations. Instead of using a domain-specific and limiting tuning framework for this, we simply provide hyperparameters through Python flags, and implement for convenience of the user the ability to specify automatic logging to TensorBoard and Weights & Biases, an increasingly popular deep learning experiment management service [4].

A.2 Class Imbalance Adjustment

We compensate for the class imbalance discussed in Section 2 by reweighing the cross-entropy portion of each objective function, placing more weight on the minority class based on the relative class frequencies in each mini-batch of M samples, $p(k)_{\text{mini-batch}}$ [36]:

$$\mathcal{L} = -\frac{1}{KM}\sum_{i=1}^{M}\frac{\mathcal{L}_{\text{cross-entropy}}(i)}{p(k)_{\text{mini-batch}}}, \tag{A.1}$$

where k is the class of sample i. We also tried using constant class weights, but found that this resulted in lower overall performance.

A.3 Mean-Field Variational Inference Implementation

We employ a set of standard optimizations to improve training stability for the MFVI and RADIAL-MFVI methods. We fix the mean of the prior to that of the variational posterior, which causes the KL term to only penalize the standard deviation of the weight posterior, and not its mean. We use flipout for lower-variance gradients in convolutional layers and the final dense layer [61], and KL annealing using a cyclical schedule, following [7]. Finally, for RADIAL-MFVI, the prior's standard deviation is by default set to the He initializer standard deviation $\sqrt{2/\text{fan\_in}}$ [45].

A.4 Uncertainty Estimation and Related Work

The Monte Carlo estimator used to computed the total uncertainty is biased but consistent and commonly used in practice [6, 10, 19]. A model's aleatoric uncertainty, $\mathbb{E}[\mathcal{H}(p(\mathbf{y}_* \mid f(\mathbf{x}_*; \boldsymbol{\theta})))]$ is estimated analogously, and the epistemic uncertainty can then be computed as the difference between the total and the aleatoric predictive uncertainty estimates.

Some other works consider uncertainty estimation in medical imaging. [60] uses test-time augmentation for uncertainty estimation, but captures only aleatoric uncertainty. [44, 51] considers uncertainty estimation with a Monte Carlo dropout model but does not isolate how their various measures of uncertainty correspond to epistemic or aleatoric uncertainty. None of the above works contribute and open-source tasks designed to emulate real-world distribution shifts, nor do they implement and benchmark a significant number of baseline uncertainty quantification models considering both aleatoric and epistemic uncertainty.

A.5 Receiver Operating Characteristic Curves

The ROC curve (e.g., see Figure 5(**a**) and (**b**)) illustrates the diagnostic ability of a binary classification system as a function of the discrimination threshold. The curve is created by plotting the true positive rate (that is, the sensitivity) against the false positive rate (that is, $1 -$ specificity). The quality of the ROC curve can be summarized by the area under the curve, which ranges from 0.5 (chance level) to 1.0 (perfect classification).

A.6 Selective Prediction Curves

For the purposes of selective prediction, a model with optimal uncertainty estimates on a given dataset would have uncertainty perfectly correlate rank-wise with the model error. For example, the image on which the model has the highest error should be assigned the highest uncertainty, the image with the second highest error should be assigned the second highest uncertainty, and so on. On the other hand, the worst possible uncertainty estimates are random, which would be uninformative to referral.

Finally, we explain in more detail the dip observed at the right side of selective prediction curves using AUC as the base metric (e.g., Figure 5**(c)** and **(d)**). At relatively high threshold values τ, models begin to refer examples on which they are both confident and correct. This results in the selective prediction curve decreasing. At the highest τ values (the last few examples), for many models, nearly all remaining predictions are correct with high certainty, and the AUC increases.

A.7 Hyperparameter Tuning

We provide full tuning details so that users of the benchmark will be able to reproduce our results.

All tuning scripts across all methods, tasks (*Country* and *Severity Shift*), and tuning procedures (on in-domain validation AUC and area under the selective prediction accuracy curve using the joint validation dataset, described in Appendix B.3) are documented in the Uncertainty Baselines repository.[6]

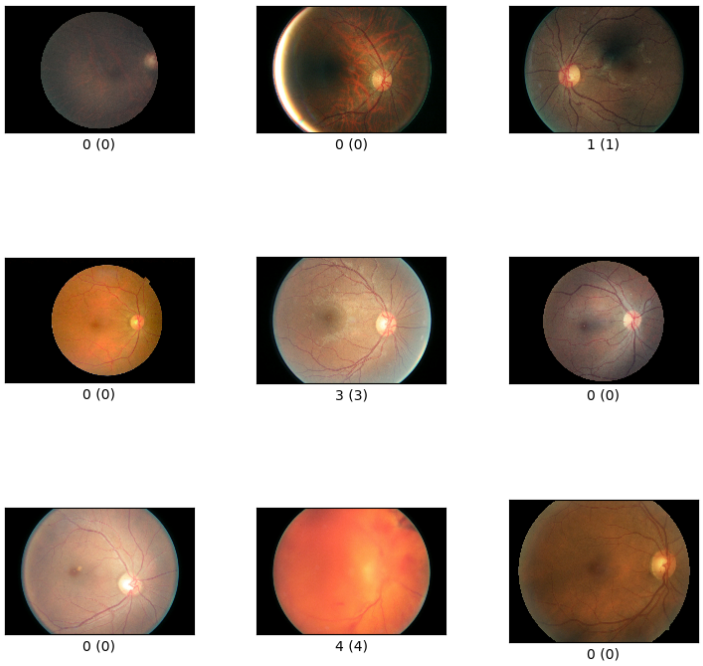
We tuned each model with a quasi-random search on several hyperparameters including learning rate, momentum, ℓ_2 regularization, and method-specific variables including dropout rate and variational posterior initializations. We used a minimum of 32 trials per model. Because of the large size of the input data and significant expense of multiple Monte Carlo samples at training time for some of the variational methods (in particular, MFVI, RANK-1, and RADIAL-MFVI), we were unable to achieve a large batch size with multiple variational samples at training time. With a single variational sample at training time, we were able to fit more reasonable batch sizes (≥ 64) and found this to significantly improve convergence and performance on validation metrics. We attribute this to the batch size increase and the usage of variance reduction techniques such as flipout layers [61], which mitigate the impact of only using a single variational sample at training time.

We considered model selection for each of the models on each of the two tasks (*Country* and *Severity Shift*) using two different validation metrics: in-domain validation AUC, and area under the accuracy referral curve constructed using both in-domain and distributionally shifted validation data. We describe the reasoning behind the latter metric in Appendix B.3. We used this validation performance to select the best hyperparameter setting and retrained a configuration for each combination of model, task, and validation tuning metric for 6 random seeds. We evaluated single models by averaging performance over those seeds, and evaluated ensembles by randomly sampling ensembles of size 3 without replacement from the 6 available models, and averaging over 6 such ensemble constructions.

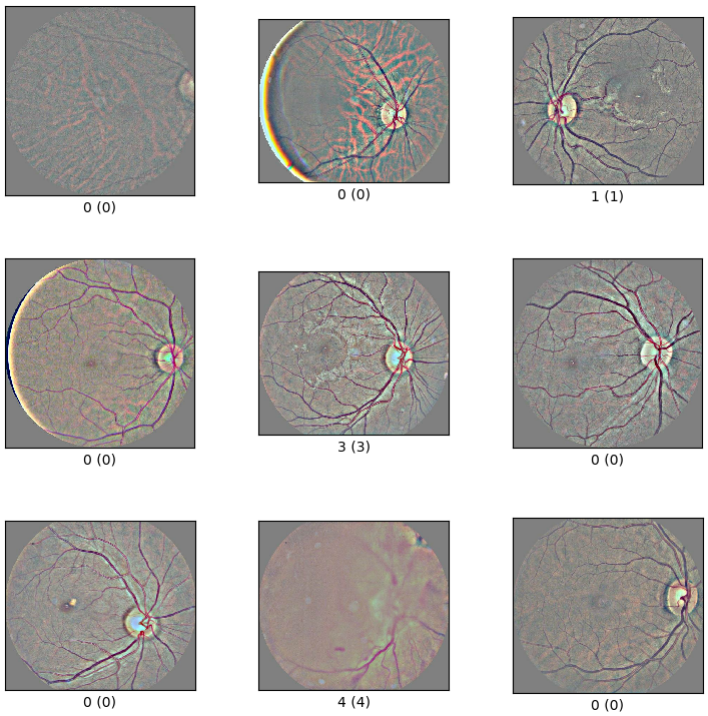
Compute Resources. The majority of methods were tuned on TPU v2-8 nodes. MFVI had particularly high memory requirements which required the use of TPU v3-8 nodes to achieve a reasonable batch size and stable training. Evaluation was performed on NVIDIA A100 GPUs with 40 GB memory, though GPUs with standard sizes (e.g., >6 GB) will be sufficient to run evaluation and inference with the models in the benchmark, e.g., using the model checkpoints. Approximately 100 TPU days and 20 GPU days were used collectively across the initial hyperparameter tuning, fine-tuning with selected configurations, and evaluation across the various tasks. Though a significant cost, we hope that our open-sourcing of all code along with hyperparameter sweep details and checkpoints will significantly decrease future consumption of researchers interested in designing deep models for diabetic retinopathy, along with Bayesian deep learning researchers using our configurations to inform their hyperparameter tuning, or our generally applicable evaluation utilities.

[6] https://github.com/google/uncertainty-baselines/tree/main/baselines/diabetic_retinopathy_detection

A.8 EyePACS and APTOS Input Data Examples



(a) Original samples from the EyePACS Diabetic Retinopathy dataset [13].



(b) Processed and augmented samples from the EyePACS Diabetic Retinopathy dataset, following the procedure of the Kaggle competition winner [13].

Figure 6: Illustrative examples of retina images in the original EyePACS dataset (top) and after preprocessing (bottom).

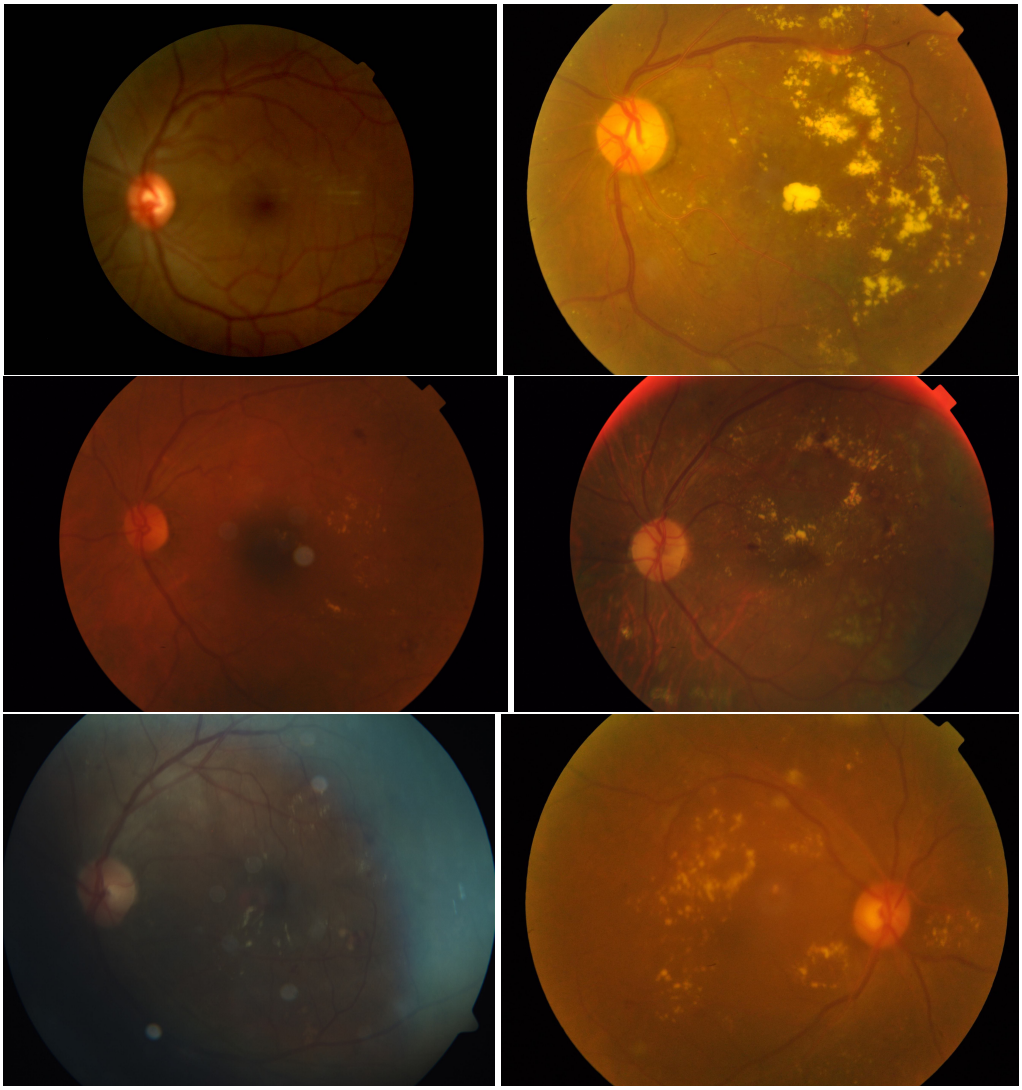


Figure 7: Illustrative examples of retina images in the APTOS dataset. The images are collected using different measurement devices than the EyePACS dataset. Note the artifacts present in the images including blur, low background lighting, and effects around the edges of the retina.

Appendix B Further Empirical Results

B.1 Predictive Uncertainty Histograms

In the figures below, predictive uncertainty (cf. Section 4) is displayed as a normalized density for correct (blue) and incorrect (red) predictions. All histograms are normalized and are displayed with the same range on the x- and y-axis. Some bars of the histograms are cut off because the plots are zoomed-in along the y-axis to improve legibility. See Section 2.5 for a description of predictive uncertainty histograms as a model diagnostic tool, including a discussion of the expected behavior of reliable models. See Section 6 for a discussion of the results for single models on the shifted datasets.

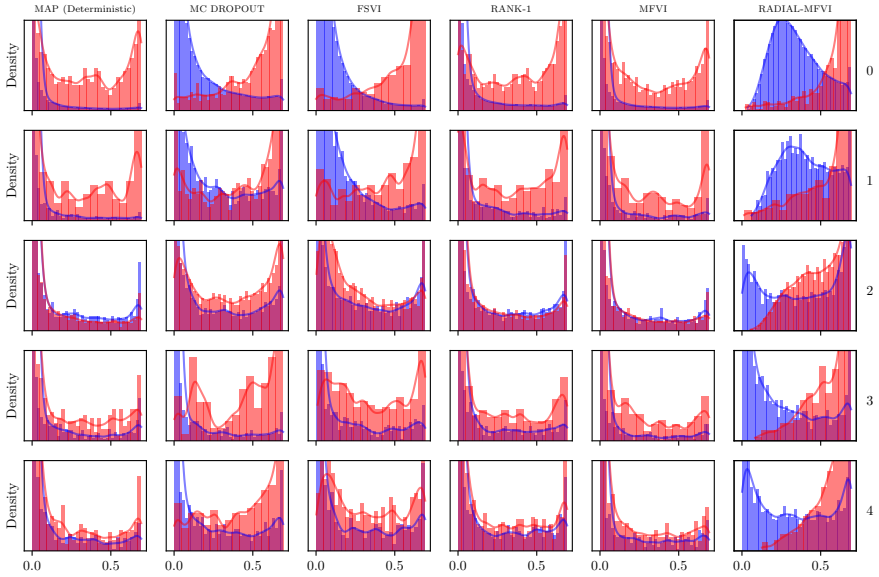


Figure 8: Clinical Label Binning – Severity Shift, Single Models. We analyze predictive uncertainty for each underlying clinical severity label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider both the in-domain and distributionally shifted Severity Shift evaluation datasets, and single models ($K = 1$). Predictive uncertainty, as measured by total uncertainty (cf. Section 4), is displayed as a normalized density for correct (blue) and incorrect (red) predictions.

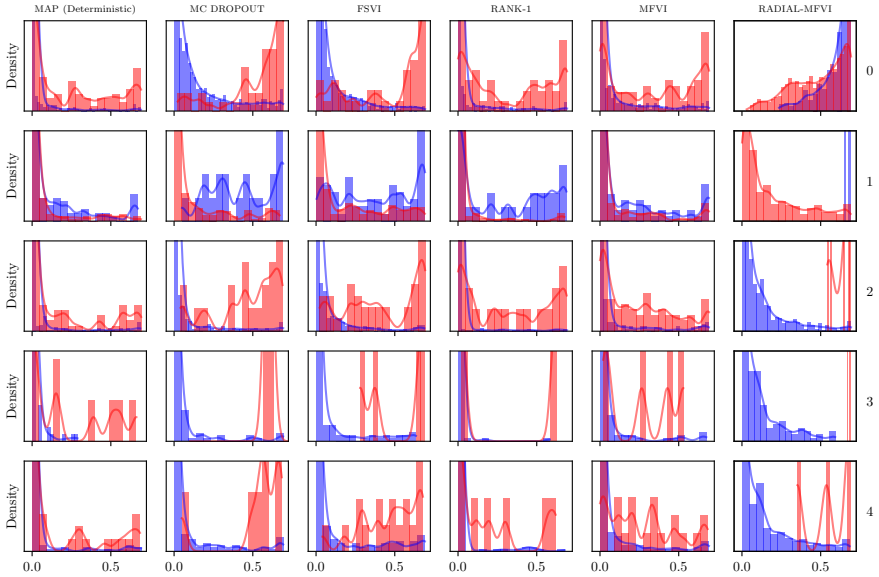


Figure 9: Clinical Label Binning – Country Shift (Shifted), Single Models. We analyze predictive uncertainty for each underlying clinical severity label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider the distributionally shifted Country Shift evaluation dataset (APTOS), and single models ($K = 1$). Predictive uncertainty, as measured by total uncertainty (cf. Section 4), is displayed as a normalized density for correct (blue) and incorrect (red) predictions.

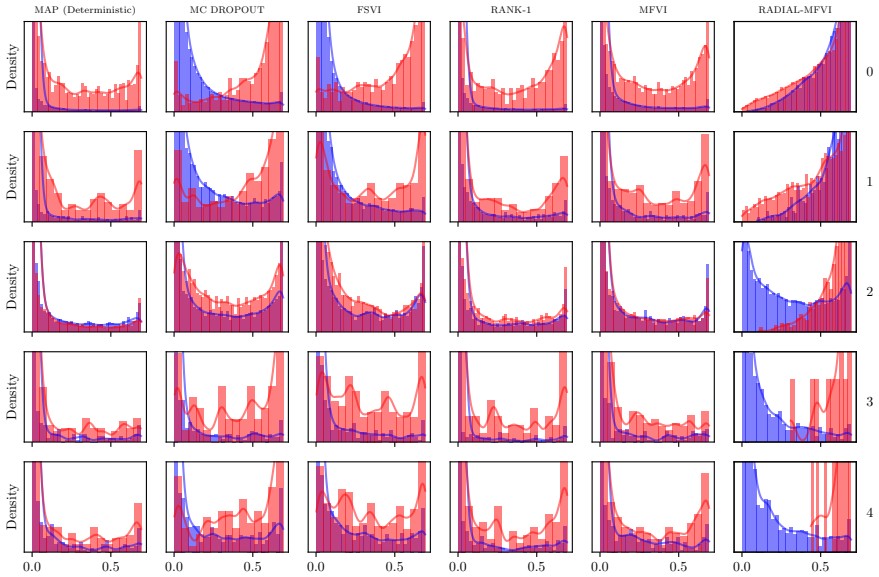


Figure 10: Clinical Label Binning – Country Shift (In-Domain), Single Models. We analyze predictive uncertainty for each ground-truth clinical label (rows) and each uncertainty quantification method (columns). Here, we consider the in-domain Country Shift evaluation dataset, and single models ($K = 1$). Predictive uncertainty, as measured by total uncertainty (cf. Section 4), is displayed as a normalized density for correct (blue) and incorrect (red) predictions.

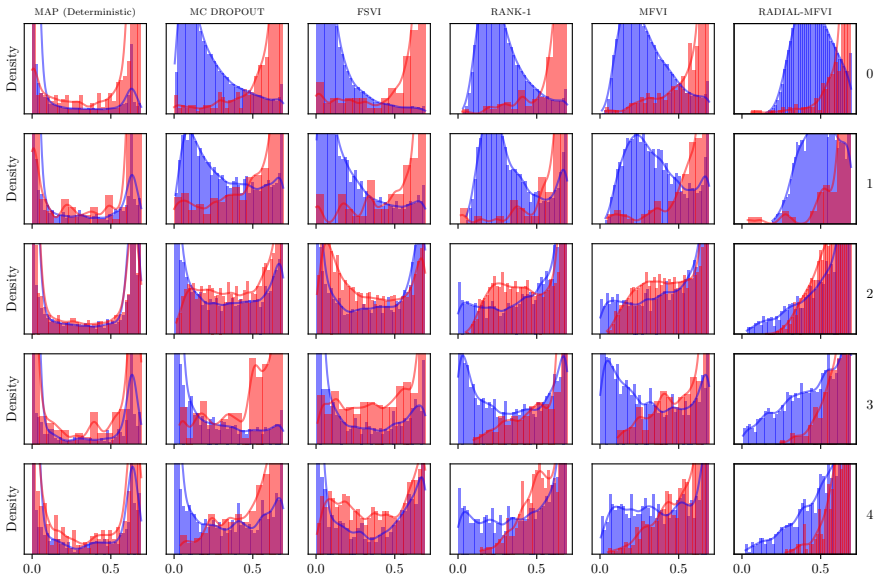


Figure 11: Clinical Label Binning – Severity Shift, Ensembles. We analyze predictive uncertainty for each ground-truth clinical label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider both the in-domain and distributionally shifted Severity Shift evaluation datasets, and ensembles ($K = 3$). Predictive uncertainty, as measured by total uncertainty (cf. Section 4), is displayed as a normalized density for correct (blue) and incorrect (red) predictions.

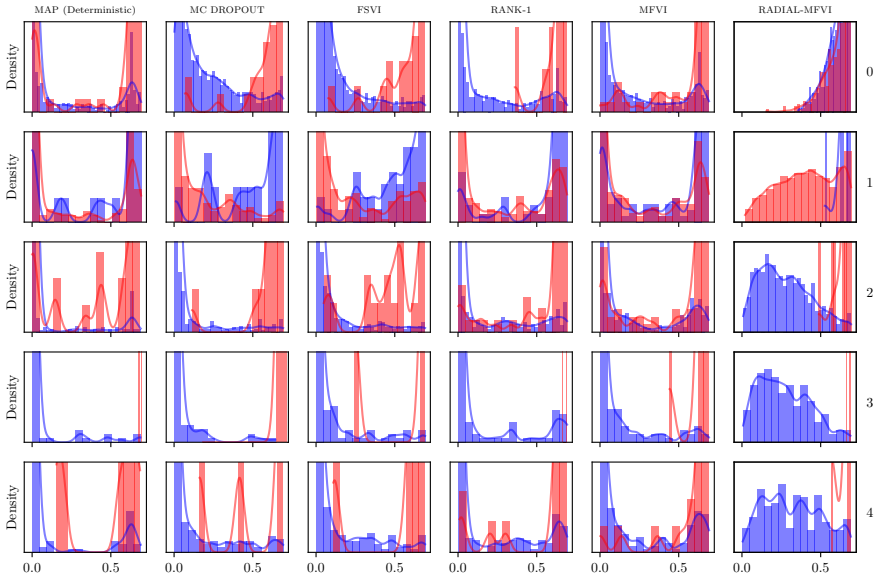


Figure 12: Clinical Label Binning – Country Shift (Shifted), Ensembles. We analyze predictive uncertainty for each ground-truth clinical label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider the distributionally shifted Country Shift evaluation dataset (APTOS), and ensembles ($K = 3$). Predictive uncertainty, as measured by total uncertainty (cf. Section 4), is displayed as a normalized density for correct (blue) and incorrect (red) predictions.

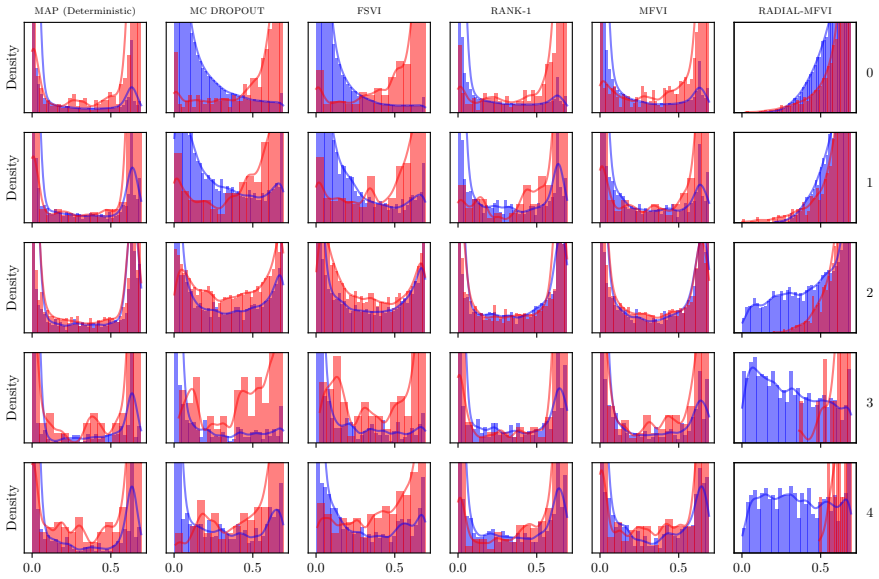
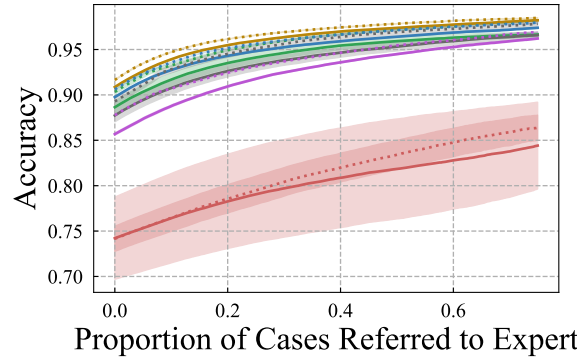


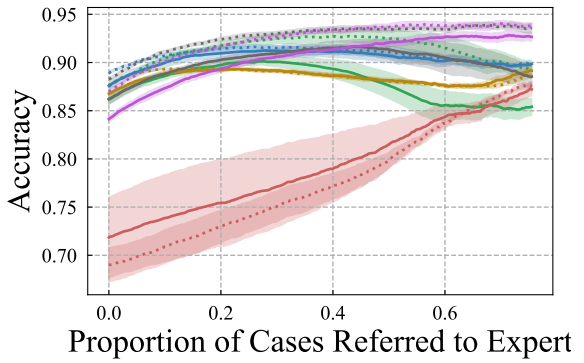
Figure 13: Clinical Label Binning – Country Shift (In-Domain), Ensembles. We analyze predictive uncertainty for each ground-truth clinical label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider the in-domain Country Shift evaluation dataset (APTOS), and ensembles ($K = 3$). Predictive uncertainty, as measured by total uncertainty (cf. Section 4), is displayed as a normalized density for correct (blue) and incorrect (red) predictions.

B.2 Tuning without Distributionally Shifted Data: Country Shift Accuracy.

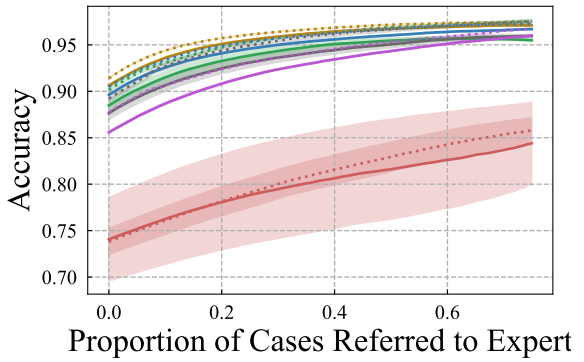
We provide referral curves on accuracy for *Country Shift* with in-domain validation tuning in Figure 14.



(a) Selective Prediction Accuracy: In-Domain



(b) Selective Prediction Accuracy: Country Shift



(c) Selective Prediction Accuracy: Joint

Figure 14: Selective Prediction: Country Shift (Accuracy). We use the binary accuracy for in-domain diagnosis on the EyePACS [13] test set (**a**), for changing medical equipment and patient populations on the shifted APTOS [3] evaluation set (**b**), and on a joint dataset composed of both the in-domain and APTOS datasets (**c**). Shading denotes one standard error.

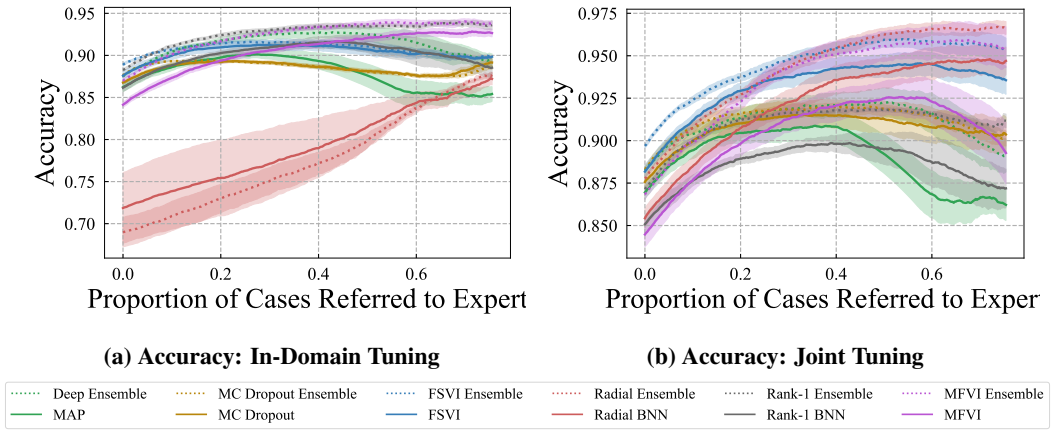
B.3 Tuning in the Presence of Distributionally Shifted Data

In prior work in Bayesian deep learning, little emphasis has been placed on the standardization of a training and evaluation protocol; in particular, the assumption of whether a model has access to distributionally shifted validation data for hyperparameter tuning is often changed on an ad-hoc basis across studies.

This is a significant assumption, and researchers in Bayesian deep learning should be expected to outwardly declare their tuning procedure—in particular access to distributionally shifted data—as is done in works such as Prior Networks [39, 40]. This will permit researchers and practitioners to more fairly compare the performance of methods based on results reported in their respective papers.

We investigate what impact this assumption—access to distributionally shifted validation data—has on downstream performance across all our tasks, and on held-out in-domain, distributionally shifted, and joint (in-domain combined with distributionally shifted) evaluation datasets. We find that it has a significant impact on metrics commonly used to assess robustness and uncertainty quantification, including area under referral curves (Figure 15) and expected calibration error.

Joint Validation Metric. To consider the performance of our baseline models under this assumption, we construct a metric that conveys both in-domain and distributionally shifted performance. In particular, we construct an accuracy referral curve on a combined set of in-domain and distributionally shifted validation examples. Because the in-domain validation dataset is significantly larger than the distributionally shifted dataset for both of the tasks, we upsample the shifted dataset to avoid the signal from the in-domain examples overwhelming that from the shifted examples. We construct an *upsampled shifted dataset* by first duplicating the shifted validation dataset as many times as possible without exceeding the size of the in-domain validation dataset, and then randomly sampling examples from the shifted validation dataset without replacement until the upsampled shifted dataset contains the same number of examples as the in-domain validation dataset. We construct the "balanced" joint validation dataset as the union of the in-domain validation and upsampled shifted datasets. We construct a "balanced" accuracy referral curve using this balanced joint validation dataset, sweeping over τ to obtain all possible partitions of the dataset into "referral" and "non-referral". We then tune on the area under this curve.



(a) Accuracy: In-Domain Tuning **(b) Accuracy: Joint Tuning**

Figure 15: Hyperparameter Tuning on Distributionally Shifted Data. Accuracy referral curve on the distributionally shifted APTOS dataset in the Country Shift task. **Left:** Performance of various methods when using the in-domain validation AUC for hyperparameter tuning. **Right:** The same methods when using the proposed balanced referral metric evaluated over a combination of in-domain and distributionally shifted validation data. Even without permitting a model to explicitly train on distributionally shifted data, the model selection process results in significantly improved predictive performance and quality of uncertainty estimates, as demonstrated by curves for respective methods shifted upwards, and steeper slopes in each curve as the first $\approx 50\%$ of cases are referred to an expert, respectively.

B.4 Complete Tabular Results

We report additional tabular results for standard predictive performance and robustness (expected calibration error), referral metrics, and out-of-distribution detection across the *Severity* and *Country Shift* tasks, considering hyperparameter tuning on either in-domain validation AUC or the joint validation metric (cf. Appendix B.3), in Tables 2-11.

Table 2: Severity Shift. Prediction and uncertainty quality of baseline methods in terms of area under the receiver operating characteristic curve (AUC) and classification accuracy, as a function of the proportion of data referred to a medical expert for further review.

	No Referral		50% Data Referred		70% Data Referred	
Method	**AUC (%) ↑**	**Accuracy (%) ↑**	**AUC (%) ↑**	**Accuracy (%) ↑**	**AUC (%) ↑**	**Accuracy ↑**
In-Domain (No, Mild, or Moderate DR, Clinical Labels {0,1,2})						
MAP (Deterministic)	82.0±1.0	87.9±0.4	83.1±1.9	95.2±0.3	88.4±1.9	96.0±0.2
DEEP ENSEMBLE	85.1±0.7	89.3±0.2	82.0±0.9	96.3±0.2	85.3±0.9	97.3±0.2
MC DROPOUT	89.2±0.2	90.5±0.1	92.8±0.6	97.2±0.0	95.4±0.4	97.8±0.0
MC DROPOUT ENSEMBLE	**90.6±0.0**	**91.4±0.1**	**93.1±0.2**	**97.8±0.0**	**95.7±0.2**	**98.2±0.0**
FSVI	83.2±0.3	89.5±0.1	81.2±0.9	95.6±0.1	86.4±0.7	96.4±0.1
FSVI ENSEMBLE	86.2±0.1	90.0±0.0	81.2±0.3	96.4±0.0	86.1±0.3	97.3±0.0
RADIAL-MFVI	76.9±1.6	86.7±0.4	69.0±4.1	93.5±0.5	70.1±4.8	94.6±0.5
RADIAL-MFVI ENSEMBLE	81.3±1.2	87.4±0.3	66.3±2.3	95.1±0.4	66.2±3.0	96.1±0.4
RANK-1	81.6±1.5	88.3±0.5	79.4±2.9	95.1±0.4	82.9±2.9	96.0±0.4
RANK-1 ENSEMBLE	85.1±1.1	89.3±0.4	75.6±1.0	96.1±0.3	79.1±1.3	96.9±0.2
MFVI	81.3±1.4	87.8±0.6	79.5±2.4	95.0±0.4	82.6±2.6	95.9±0.3
MFVI ENSEMBLE	85.2±0.6	89.4±0.3	77.7±0.9	96.1±0.2	80.3±1.0	96.8±0.1
Severity Shift (Severe or Proliferate DR, Clinical Labels {3, 4})						
MAP (Deterministic)	–	74.4±1.9	–	93.2±2.6	–	98.6±1.1
DEEP ENSEMBLE	–	74.5±1.2	–	89.8±1.0	–	97.0±0.7
MC DROPOUT	–	86.4±1.3	–	**99.5±0.2**	–	**100.0±0.0**
MC DROPOUT ENSEMBLE	–	**87.4±0.3**	–	99.4±0.1	–	**100.0±0.0**
FSVI	–	68.6±1.0	–	88.5±1.0	–	99.6±0.2
FSVI ENSEMBLE	–	69.3±0.3	–	86.3±0.5	–	99.4±0.2
RADIAL-MFVI	–	52.0±7.6	–	59.3±10.8	–	63.9±11.1
RADIAL-MFVI ENSEMBLE	–	54.4±4.7	–	58.0±7.6	–	60.6±8.3
RANK-1	–	67.5±3.5	–	82.6±4.2	–	92.7±2.2
RANK-1 ENSEMBLE	–	69.7±1.9	–	81.6±1.6	–	92.0±1.3
MFVI	–	71.5±2.3	–	86.7±3.1	–	94.1±2.0
MFVI ENSEMBLE	–	73.5±1.2	–	87.4±0.7	–	94.2±0.7

Table 3: OOD Detection Metrics. We assess model uncertainty quantification across both shift tasks by using predictive entropy to detect out-of-distribution data.

	Country Shift		Severity Shift	
Method	**AUROC (%) ↑**	**AUPRC (%) ↑**	**AUROC (%) ↑**	**AUPRC (%) ↑**
MAP (Deterministic)	37.6±1.3	5.2±0.2	44.0±2.7	9.3±0.6
DEEP ENSEMBLE	41.7±1.0	5.6±0.1	56.8±0.9	12.4±0.3
MC DROPOUT	37.6±0.7	5.1±0.1	34.9±1.1	7.1±0.4
MC DROPOUT ENSEMBLE	39.5±0.2	5.3±0.0	38.3±0.9	7.7±0.2
FSVI	42.2±0.7	5.7±0.1	49.0±0.8	11.6±0.3
FSVI ENSEMBLE	43.8±0.4	5.9±0.1	54.5±0.4	14.5±0.2
RADIAL-MFVI	39.2±2.1	5.3±0.3	66.8±4.8	19.9±2.6
RADIAL-MFVI ENSEMBLE	36.5±0.6	4.9±0.1	**79.7±2.7**	**28.0±2.3**
RANK-1	44.3±1.9	6.0±0.3	54.5±3.4	12.8±1.1
RANK-1 ENSEMBLE	48.9±1.0	6.4±0.2	65.6±0.7	17.4±0.5
MFVI	51.2±0.7	6.7±0.1	51.3±2.8	10.4±0.7
MFVI ENSEMBLE	**52.4±0.3**	**6.9±0.1**	60.4±0.8	13.5±0.4

Table 4: Standard Metrics, Country Shift. We assess model predictive performance via standard metrics, and evaluate uncertainty quantification using expected calibration error on in-domain, shifted, and joint datasets (composed of the in-domain and shifted dataset, with no explicit balancing).

	NLL ↓			Accuracy (%) ↑			AUPRC (%) ↑		
Method	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**
MAP (Deterministic)	1.27±0.06	2.68±0.14	1.36±0.05	88.6±0.5	86.2±0.4	88.5±0.5	75.2±1.7	89.7±0.2	77.2±1.4
DEEP ENSEMBLE	0.60±0.00	1.60±0.12	0.67±0.01	90.3±0.2	87.5±0.1	90.1±0.2	79.9±0.4	**91.1±0.1**	81.0±0.3
MC DROPOUT	0.29±0.00	1.07±0.02	0.34±0.00	90.9±0.0	86.8±0.2	90.6±0.0	82.6±0.2	88.8±0.4	82.9±0.2
MC DROPOUT ENSEMBLE	**0.25±0.00**	0.92±0.02	**0.29±0.00**	**91.6±0.0**	87.6±0.0	**91.4±0.0**	**84.4±0.0**	88.3±0.3	**84.3±0.1**
FSVI	0.35±0.01	0.72±0.04	0.38±0.01	89.8±0.0	87.6±0.3	89.6±0.0	77.7±0.1	88.3±0.4	78.9±0.0
FSVI ENSEMBLE	0.28±0.01	**0.58±0.01**	0.30±0.01	90.6±0.0	**88.9±0.1**	90.5±0.0	80.7±0.1	88.9±0.2	81.3±0.0
RADIAL-MFVI	0.56±0.05	0.70±0.07	0.57±0.05	74.2±3.5	71.8±3.2	74.1±3.5	66.0±0.7	84.8±0.6	69.0±0.6
RADIAL-MFVI ENSEMBLE	0.55±0.01	0.65±0.03	0.56±0.01	74.2±1.1	69.0±1.3	73.8±1.1	68.9±0.3	86.1±0.1	71.6±0.2
RANK-1	0.99±0.05	1.85±0.15	1.05±0.04	87.7±0.5	86.2±0.4	87.6±0.5	71.6±1.9	88.8±0.4	74.1±1.6
RANK-1 ENSEMBLE	0.49±0.03	0.96±0.05	0.52±0.02	89.3±0.3	88.3±0.1	89.2±0.3	78.0±1.0	89.6±0.2	79.3±0.8
MFVI	0.91±0.01	1.26±0.05	0.93±0.01	85.7±0.1	84.1±0.2	85.6±0.1	66.7±0.2	85.9±0.2	69.7±0.2
MFVI ENSEMBLE	0.53±0.00	0.72±0.02	0.54±0.00	87.8±0.0	87.0±0.2	87.7±0.0	71.2±0.1	87.4±0.1	73.7±0.0
	AUROC (%) ↑			ECE ↓					
MAP (Deterministic)	87.4±0.9	92.2±0.2	88.3±0.8	0.10±0.00	0.13±0.00	0.10±0.00			
DEEP ENSEMBLE	90.3±0.1	94.2±0.1	90.9±0.1	0.06±0.00	0.08±0.00	0.06±0.00			
MC DROPOUT	91.4±0.1	94.0±0.1	91.9±0.1	0.03±0.00	0.09±0.00	0.03±0.00			
MC DROPOUT ENSEMBLE	**92.5±0.0**	94.1±0.1	**92.9±0.0**	**0.02±0.00**	0.09±0.00	**0.02±0.00**			
FSVI	88.5±0.1	94.1±0.1	89.4±0.0	0.05±0.01	0.08±0.00	0.06±0.01			
FSVI ENSEMBLE	90.3±0.1	**94.6±0.0**	90.9±0.0	0.03±0.00	0.07±0.00	0.03±0.00			
RADIAL-MFVI	83.2±0.4	90.7±0.5	84.3±0.3	0.09±0.02	0.14±0.03	0.09±0.02			
RADIAL-MFVI ENSEMBLE	84.9±0.1	91.8±0.1	85.9±0.1	0.06±0.01	0.10±0.01	0.05±0.01			
RANK-1	85.6±1.0	92.5±0.2	86.7±0.9	0.10±0.00	0.11±0.00	0.10±0.00			
RANK-1 ENSEMBLE	89.5±0.6	94.1±0.1	90.2±0.5	0.05±0.00	0.06±0.00	0.05±0.00			
MFVI	83.3±0.1	91.4±0.2	84.6±0.1	0.11±0.00	0.13±0.00	0.12±0.00			
MFVI ENSEMBLE	85.4±0.0	93.2±0.1	86.6±0.0	0.06±0.00	**0.06±0.00**	0.06±0.00			

Table 5: Standard Metrics, Severity Shift. We assess model predictive performance and expected calibration error on in-domain, shifted, and joint datasets (composed of the in-domain and shifted dataset, with no explicit balancing).

	NLL ↓			Accuracy (%) ↑			AUPRC (%) ↑		
Method	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**
MAP (Deterministic)	1.27±0.06	2.27±0.12	1.35±0.06	87.9±0.4	74.4±1.7	86.8±0.5	60.8±1.9	–	75.2±1.3
DEEP ENSEMBLE	0.62±0.02	1.03±0.05	0.65±0.02	89.3±0.2	74.5±1.1	88.1±0.3	65.6±1.2	–	79.2±0.8
MC DROPOUT	0.29±0.00	0.33±0.01	0.29±0.00	90.5±0.0	86.4±1.1	90.1±0.1	74.8±0.5	–	85.1±0.2
MC DROPOUT ENSEMBLE	**0.25±0.00**	**0.28±0.00**	**0.25±0.00**	**91.4±0.1**	**87.4±0.2**	**91.1±0.0**	**77.0±0.0**	–	**86.7±0.1**
FSVI	0.36±0.01	0.92±0.04	0.41±0.01	89.5±0.1	68.6±0.9	87.8±0.2	64.7±0.7	–	77.6±0.4
FSVI ENSEMBLE	0.31±0.00	0.76±0.01	0.34±0.00	90.0±0.0	69.3±0.2	88.4±0.1	70.0±0.1	–	81.6±0.1
RADIAL-MFVI	0.37±0.01	0.76±0.09	0.40±0.02	86.7±0.3	52.0±7.0	83.9±0.8	49.1±2.7	–	66.9±2.2
RADIAL-MFVI ENSEMBLE	0.35±0.01	0.73±0.05	0.38±0.01	87.4±0.3	54.4±4.3	84.8±0.6	56.2±2.0	–	73.5±1.5
RANK-1	0.56±0.05	1.14±0.12	0.61±0.05	88.3±0.5	67.5±3.2	86.6±0.7	59.4±2.8	–	74.1±1.9
RANK-1 ENSEMBLE	0.29±0.01	0.60±0.03	0.32±0.01	89.3±0.3	69.7±1.7	87.7±0.4	66.5±1.9	–	80.0±1.3
MFVI	0.66±0.08	1.26±0.16	0.71±0.09	87.8±0.5	71.5±2.1	86.5±0.6	59.0±2.5	–	73.7±1.8
MFVI ENSEMBLE	0.29±0.01	0.55±0.02	0.31±0.01	89.4±0.3	73.5±1.1	88.2±0.3	66.4±1.3	–	79.7±0.9
	AUROC (%) ↑			ECE ↓					
MAP (Deterministic)	82.0±0.9	–	86.3±0.7	0.11±0.00	0.23±0.01	0.12±0.00			
DEEP ENSEMBLE	85.1±0.6	–	88.9±0.5	0.06±0.00	0.15±0.01	0.07±0.00			
MC DROPOUT	89.2±0.2	–	92.0±0.1	0.02±0.00	0.06±0.00	0.02±0.00			
MC DROPOUT ENSEMBLE	**90.6±0.0**	–	**93.1±0.0**	**0.01±0.00**	**0.03±0.00**	**0.01±0.00**			
FSVI	83.2±0.3	–	86.9±0.2	0.06±0.00	0.23±0.01	0.07±0.00			
FSVI ENSEMBLE	86.2±0.1	–	89.4±0.0	0.04±0.00	0.19±0.00	0.06±0.00			
RADIAL-MFVI	76.9±1.4	–	82.2±1.2	0.05±0.01	0.23±0.06	0.04±0.01			
RADIAL-MFVI ENSEMBLE	81.3±1.1	–	86.2±0.9	0.07±0.01	0.15±0.03	0.06±0.01			
RANK-1	81.6±1.4	–	85.8±1.1	0.06±0.01	0.22±0.02	0.07±0.01			
RANK-1 ENSEMBLE	85.1±1.0	–	89.1±0.7	0.02±0.00	0.12±0.01	0.03±0.00			
MFVI	81.3±1.2	–	85.4±1.0	0.07±0.01	0.19±0.02	0.08±0.01			
MFVI ENSEMBLE	85.2±0.6	–	88.9±0.4	0.02±0.00	0.10±0.01	0.02±0.00			

Table 6: Expert Referral Metrics, Country Shift. We assess model predictive performance and uncertainty quantification in the context of expert referral. We construct referral curves on a variety of metrics—AUC, Accuracy, NLL and AUPRC—by sweeping over the referral thresholds τ, obtaining a point for each possible partition of the dataset into "referred" and "non-referred". We report the area under the referral curve for metric X as R-X AUC. All methods are tuned according to the area under the ROC curve on the in-domain dataset. The Balanced evaluation dataset is constructed using the procedure described in Appendix B.3.

	R-AUROC AUC ↑				R-Accuracy AUC ↑			
Method	**In-Domain**	**Shifted**	**Joint**	**Balanced**	**In-Domain**	**Shifted**	**Joint**	**Balanced**
MAP (Deterministic)	94.7±0.5	91.4±0.6	95.1±0.4	95.1±0.4	97.0±0.2	94.1±0.2	96.8±0.2	95.4±0.1
DEEP ENSEMBLE	95.5±0.1	94.3±0.5	96.1±0.1	96.1±0.1	97.7±0.0	95.6±0.2	97.6±0.0	96.6±0.1
MC DROPOUT	96.9±0.1	93.6±0.1	97.1±0.1	97.1±0.1	97.9±0.0	94.3±0.1	97.6±0.0	95.9±0.0
MC DROPOUT ENSEMBLE	**97.2±0.0**	93.4±0.1	**97.4±0.0**	**97.4±0.0**	**98.2±0.0**	94.4±0.0	**97.9±0.0**	96.1±0.0
FSVI	94.9±0.1	94.9±0.2	95.6±0.1	95.6±0.1	97.4±0.0	95.2±0.2	97.2±0.0	96.2±0.1
FSVI ENSEMBLE	95.5±0.1	94.9±0.1	96.1±0.0	96.1±0.0	97.8±0.0	95.4±0.1	97.6±0.0	96.5±0.1
RADIAL-MFVI	93.4±0.3	91.9±0.5	93.8±0.3	93.8±0.3	89.8±2.0	89.1±1.3	89.8±1.9	89.3±1.6
RADIAL-MFVI ENSEMBLE	94.4±0.1	91.4±0.1	94.7±0.1	94.7±0.1	90.2±0.6	88.2±0.5	90.0±0.6	88.9±0.5
RANK-1	93.3±0.6	94.1±0.6	94.1±0.5	94.1±0.5	96.7±0.2	94.9±0.2	96.5±0.2	95.7±0.0
RANK-1 ENSEMBLE	94.4±0.4	96.6±0.2	95.3±0.3	95.3±0.3	97.5±0.1	**96.2±0.1**	97.4±0.1	**96.8±0.1**
MFVI	92.5±0.2	96.1±0.1	93.4±0.1	93.4±0.1	96.0±0.0	94.9±0.1	95.9±0.0	95.4±0.1
MFVI ENSEMBLE	93.1±0.1	**97.0±0.1**	94.0±0.1	94.0±0.1	96.7±0.0	96.0±0.1	96.6±0.0	96.3±0.0
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	60.5±3.6	179.5±11.4	69.4±2.5	126.0±5.2	90.9±0.9	95.8±0.1	91.9±0.7	91.9±0.7
DEEP ENSEMBLE	28.7±0.2	110.5±9.7	34.7±0.8	73.0±5.5	91.7±0.3	**96.6±0.1**	92.8±0.2	92.8±0.2
MC DROPOUT	10.5±0.2	74.4±2.0	15.2±0.3	44.9±1.1	94.5±0.1	95.1±0.2	94.4±0.1	94.4±0.1
MC DROPOUT ENSEMBLE	**8.2±0.1**	63.6±1.4	12.2±0.2	38.0±0.8	**94.9±0.1**	94.8±0.1	**94.7±0.0**	**94.7±0.0**
FSVI	13.4±0.5	45.1±3.1	15.6±0.6	30.2±1.8	91.1±0.2	94.9±0.2	91.9±0.1	91.9±0.1
FSVI ENSEMBLE	9.8±0.3	35.5±0.8	**11.6±0.3**	**23.5±0.5**	91.9±0.1	95.2±0.1	92.5±0.1	92.5±0.1
RADIAL-MFVI	26.2±3.7	34.6±4.0	26.9±3.8	30.9±3.9	87.5±0.4	93.5±0.3	88.6±0.4	88.6±0.4
RADIAL-MFVI ENSEMBLE	24.4±0.8	**31.1±1.6**	24.9±0.9	28.3±1.3	88.9±0.1	93.9±0.1	89.8±0.1	89.8±0.1
RANK-1	48.0±3.0	123.5±12.0	53.6±1.9	89.2±5.5	88.1±1.2	95.8±0.1	89.6±1.0	89.6±1.0
RANK-1 ENSEMBLE	22.2±1.7	64.3±3.7	25.1±1.4	44.2±1.8	89.3±0.7	96.1±0.1	90.8±0.6	90.8±0.6
MFVI	42.5±0.6	76.4±3.9	44.7±0.8	59.8±2.2	86.3±0.3	94.6±0.1	87.9±0.2	87.9±0.2
MFVI ENSEMBLE	24.9±0.2	44.6±1.8	26.2±0.3	34.9±1.0	87.3±0.2	94.9±0.0	88.9±0.1	88.9±0.1

Table 7: Expert Referral Metrics, Severity Shift. We assess model predictive performance and uncertainty quantification in the context of expert referral. We construct referral curves on a variety of metrics—AUC, Accuracy, NLL and AUPRC—by sweeping over the referral thresholds τ, obtaining a point for each possible partition of the dataset into "referred" and "non-referred". We report the area under the referral curve for metric X as R-X AUC. All methods are tuned according to the area under the ROC curve on the in-domain dataset. The Balanced evaluation dataset is constructed using the procedure described in Appendix B.3.

	R-AUROC AUC ↑				R-Accuracy AUC ↑			
Method	**In-Domain**	**Shifted**	**Joint**	**Balanced**	**In-Domain**	**Shifted**	**Joint**	**Balanced**
MAP (Deterministic)	91.5±0.7	–	94.0±0.5	94.0±0.5	96.6±0.2	92.7±0.9	96.4±0.2	94.9±0.4
DEEP ENSEMBLE	91.5±0.4	–	94.5±0.3	94.5±0.3	97.3±0.1	92.2±0.5	97.0±0.1	95.2±0.2
MC DROPOUT	95.7±0.2	–	97.2±0.1	97.2±0.1	97.8±0.0	97.5±0.3	97.8±0.0	97.7±0.1
MC DROPOUT ENSEMBLE	**96.1±0.1**	–	**97.5±0.0**	**97.5±0.0**	**98.1±0.0**	**97.7±0.1**	**98.1±0.0**	**98.0±0.0**
FSVI	90.7±0.3	–	93.6±0.2	93.6±0.2	97.0±0.1	90.3±0.4	96.7±0.1	94.2±0.2
FSVI ENSEMBLE	91.0±0.1	–	94.2±0.1	94.2±0.1	97.5±0.0	90.0±0.1	97.1±0.0	94.5±0.1
RADIAL-MFVI	85.6±1.5	–	88.5±1.5	88.5±1.5	95.9±0.2	78.4±4.4	95.2±0.3	89.5±1.6
RADIAL-MFVI ENSEMBLE	85.3±1.0	–	88.4±1.1	88.4±1.1	96.5±0.2	78.5±3.0	95.9±0.3	90.2±1.1
RANK-1	90.0±1.1	–	92.9±0.8	92.9±0.8	96.7±0.2	88.7±1.7	96.3±0.3	93.5±0.7
RANK-1 ENSEMBLE	89.3±0.4	–	93.1±0.4	93.1±0.4	97.3±0.2	89.4±0.8	96.9±0.2	94.3±0.4
MFVI	90.1±0.9	–	93.1±0.7	93.1±0.7	96.5±0.2	90.6±1.2	96.2±0.3	94.0±0.6
MFVI ENSEMBLE	90.0±0.3	–	93.6±0.2	93.6±0.2	97.3±0.1	91.4±0.4	97.0±0.1	94.9±0.2
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	61.6±3.3	89.5±7.7	63.2±3.4	73.3±4.7	83.1±1.3	–	90.2±0.9	90.2±0.9
DEEP ENSEMBLE	29.7±1.1	49.2±2.9	30.7±1.2	37.5±1.8	82.0±0.7	–	90.2±0.5	90.2±0.5
MC DROPOUT	10.3±0.2	7.8±0.5	10.0±0.2	8.9±0.1	91.2±0.5	–	95.4±0.3	95.4±0.3
MC DROPOUT ENSEMBLE	**7.8±0.2**	**6.6±0.1**	**7.7±0.2**	**7.0±0.1**	**91.6±0.2**	–	**95.8±0.1**	**95.8±0.1**
FSVI	14.2±0.5	36.0±2.1	15.6±0.6	24.0±1.1	81.8±0.6	–	89.6±0.4	89.6±0.4
FSVI ENSEMBLE	11.0±0.1	30.9±0.4	12.1±0.1	19.3±0.2	81.3±0.3	–	89.8±0.2	89.8±0.2
RADIAL-MFVI	14.4±0.7	37.5±6.7	15.3±0.9	22.9±2.8	69.8±3.1	–	78.1±3.2	78.1±3.2
RADIAL-MFVI ENSEMBLE	13.3±0.4	36.9±4.0	14.2±0.5	21.9±1.6	66.1±2.1	–	75.1±2.5	75.1±2.5
RANK-1	25.1±2.8	50.4±7.7	26.4±3.0	35.5±4.5	79.7±2.2	–	87.7±1.5	87.7±1.5
RANK-1 ENSEMBLE	10.1±0.5	25.2±1.7	10.8±0.6	15.9±0.9	77.5±0.9	–	87.3±0.7	87.3±0.7
MFVI	30.0±4.5	52.8±8.0	31.4±4.6	39.9±5.7	80.4±1.8	–	88.4±1.2	88.4±1.2
MFVI ENSEMBLE	10.3±0.3	22.3±1.0	11.0±0.3	15.3±0.6	79.6±0.6	–	88.8±0.4	88.8±0.4

Table 8: Standard Metrics, Country Shift, Tuned on Joint Dataset. Here all methods are tuned according to the joint validation metric (Appendix B.3): area under the retention–accuracy curve constructed on the balanced joint validation dataset (composed of the in-domain and upsampled shifted validation datasets). Ensembles have $K = 3$ constituent models. We assess model predictive performance and expected calibration error on in-domain, shifted, and joint (union of in-domain and shifted, without explicit balancing) evaluation datasets.

	NLL ↓			Accuracy (%) ↑			AUPRC (%) ↑		
Method	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**
MAP (Deterministic)	1.02±0.07	2.41±0.14	1.11±0.05	89.3±0.2	87.0±0.3	89.2±0.2	77.5±0.9	90.5±0.2	79.2±0.7
DEEP ENSEMBLE	0.54±0.00	1.65±0.13	0.61±0.01	90.8±0.0	88.3±0.1	90.7±0.0	81.1±0.1	**91.3±0.2**	82.0±0.1
MC DROPOUT	0.31±0.01	0.77±0.06	0.34±0.01	90.0±0.2	87.6±0.3	89.9±0.2	81.1±0.3	87.7±0.4	82.0±0.2
MC DROPOUT ENSEMBLE	**0.25±0.00**	0.58±0.03	**0.28±0.00**	**91.2±0.0**	88.3±0.1	**91.0±0.0**	**83.3±0.1**	87.7±0.3	**83.7±0.1**
FSVI	0.52±0.04	0.67±0.05	0.53±0.04	88.7±0.4	88.2±0.3	88.7±0.3	75.8±0.6	88.1±0.8	77.5±0.5
FSVI ENSEMBLE	0.39±0.02	0.42±0.02	0.39±0.01	89.2±0.2	**89.7±0.1**	89.3±0.2	79.5±0.3	88.9±0.4	80.5±0.2
RADIAL-MFVI	0.60±0.08	0.72±0.16	0.61±0.08	85.9±0.2	85.4±0.4	85.9±0.2	66.2±0.6	87.9±0.4	69.6±0.4
RADIAL-MFVI ENSEMBLE	0.38±0.00	**0.34±0.01**	0.38±0.00	87.2±0.2	87.8±0.1	87.2±0.2	69.7±0.3	89.3±0.2	72.6±0.2
RANK-1	0.88±0.06	1.95±0.21	0.95±0.07	87.0±0.6	85.1±0.4	86.9±0.6	71.3±1.9	88.4±0.3	73.8±1.6
RANK-1 ENSEMBLE	0.40±0.01	1.02±0.09	0.44±0.02	89.1±0.3	87.1±0.2	89.0±0.3	77.2±1.0	89.4±0.1	78.7±0.8
MFVI	1.09±0.10	1.69±0.20	1.12±0.11	85.9±0.4	84.5±0.6	85.8±0.4	67.1±1.5	87.6±0.7	70.3±1.2
MFVI ENSEMBLE	0.46±0.03	0.71±0.12	0.48±0.04	88.4±0.1	86.8±0.2	88.3±0.1	73.5±0.7	89.6±0.4	75.7±0.6

	AUROC (%) ↑			ECE ↓		
MAP (Deterministic)	88.6±0.5	93.2±0.2	89.5±0.4	0.09±0.00	0.12±0.00	0.09±0.00
DEEP ENSEMBLE	90.6±0.0	94.5±0.2	91.3±0.0	0.05±0.00	0.09±0.00	0.05±0.00
MC DROPOUT	90.7±0.2	93.9±0.2	91.4±0.1	0.03±0.00	0.08±0.00	0.04±0.00
MC DROPOUT ENSEMBLE	**91.9±0.0**	94.2±0.1	**92.5±0.0**	**0.02±0.00**	0.06±0.00	**0.02±0.00**
FSVI	87.4±0.3	94.0±0.3	88.5±0.3	0.08±0.01	0.08±0.00	0.08±0.01
FSVI ENSEMBLE	89.6±0.2	**94.6±0.1**	90.4±0.1	0.06±0.00	0.05±0.00	0.06±0.00
RADIAL-MFVI	83.0±0.3	92.7±0.3	84.3±0.2	0.09±0.01	0.07±0.01	0.09±0.01
RADIAL-MFVI ENSEMBLE	84.8±0.2	94.1±0.1	86.0±0.1	0.05±0.00	**0.03±0.00**	0.05±0.00
RANK-1	85.4±1.0	92.0±0.2	86.5±0.9	0.10±0.01	0.12±0.00	0.10±0.00
RANK-1 ENSEMBLE	89.0±0.6	94.0±0.1	89.8±0.5	0.05±0.00	0.07±0.00	0.05±0.00
MFVI	83.4±0.7	91.7±0.5	84.7±0.6	0.11±0.01	0.12±0.01	0.11±0.01
MFVI ENSEMBLE	86.8±0.4	94.0±0.3	87.9±0.3	0.05±0.00	0.06±0.01	0.05±0.00

Table 9: Standard Metrics, Severity Shift, Tuned on Joint Dataset. Here all methods are tuned according to the joint validation metric (Appendix B.3): area under the retention–accuracy curve constructed on the balanced joint validation dataset (composed of the in-domain and upsampled shifted validation datasets). Ensembles have $K = 3$ constituent models. We assess model predictive performance and expected calibration error on in-domain, shifted, and joint (union of in-domain and shifted, without explicit balancing) evaluation datasets.

	NLL ↓			Accuracy (%) ↑			AUPRC (%) ↑		
Method	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**	**In-Domain**	**Shifted**	**Joint**
MAP (Deterministic)	1.05±0.11	1.48±0.20	1.09±0.12	87.6±0.6	81.5±0.9	87.1±0.6	63.6±2.0	–	77.5±1.3
DEEP ENSEMBLE	0.39±0.04	0.49±0.07	0.40±0.04	89.6±0.3	83.1±0.4	89.1±0.3	68.1±1.1	–	81.3±0.6
MC DROPOUT	0.32±0.02	0.31±0.03	0.32±0.02	89.0±0.6	87.5±0.9	88.9±0.6	72.6±1.7	–	83.5±1.2
MC DROPOUT ENSEMBLE	**0.26±0.00**	**0.24±0.00**	**0.26±0.00**	**90.9±0.1**	**89.2±0.1**	**90.8±0.0**	**76.9±0.1**	–	**86.6±0.1**
FSVI	0.40±0.02	0.57±0.03	0.41±0.02	87.8±0.6	79.8±0.8	87.1±0.5	63.3±1.6	–	77.1±1.1
FSVI ENSEMBLE	0.29±0.00	0.41±0.01	0.30±0.00	90.0±0.1	81.5±0.4	89.4±0.1	68.7±0.6	–	81.4±0.3
RADIAL-MFVI	0.37±0.01	0.76±0.09	0.40±0.02	86.7±0.3	52.0±7.0	83.9±0.8	49.1±2.7	–	66.9±2.2
RADIAL-MFVI ENSEMBLE	0.35±0.01	0.73±0.05	0.38±0.01	87.4±0.3	54.4±4.3	84.8±0.6	56.2±2.0	–	73.5±1.5
RANK-1	0.56±0.05	1.14±0.12	0.61±0.05	88.3±0.5	67.5±3.2	86.6±0.7	59.4±2.8	–	74.1±1.9
RANK-1 ENSEMBLE	0.29±0.01	0.60±0.03	0.32±0.01	89.3±0.3	69.7±1.7	87.7±0.4	66.5±1.9	–	80.0±1.3
MFVI	0.56±0.06	0.75±0.16	0.57±0.07	83.7±0.2	79.8±1.8	83.4±0.1	55.2±0.4	–	71.0±0.6
MFVI ENSEMBLE	0.35±0.00	0.37±0.01	0.36±0.00	86.2±0.3	81.6±0.6	85.8±0.2	59.9±0.2	–	75.3±0.1

	AUROC (%) ↑			ECE ↓		
MAP (Deterministic)	83.7±0.9	–	87.8±0.7	0.09±0.01	0.15±0.02	0.09±0.01
DEEP ENSEMBLE	86.3±0.4	–	90.0±0.3	0.03±0.01	0.07±0.01	0.03±0.01
MC DROPOUT	88.2±0.8	–	91.1±0.7	0.02±0.00	0.06±0.01	0.02±0.00
MC DROPOUT ENSEMBLE	**90.6±0.1**	–	**93.1±0.0**	0.02±0.00	**0.02±0.00**	0.02±0.00
FSVI	82.8±0.7	–	86.8±0.6	0.06±0.00	0.14±0.01	0.06±0.00
FSVI ENSEMBLE	86.1±0.2	–	89.7±0.1	0.03±0.00	0.08±0.00	0.03±0.00
RADIAL-MFVI	76.9±1.4	–	82.2±1.2	0.05±0.01	0.23±0.06	0.04±0.01
RADIAL-MFVI ENSEMBLE	81.3±1.1	–	86.2±0.9	0.07±0.01	0.15±0.03	0.06±0.01
RANK-1	81.6±1.4	–	85.8±1.1	0.06±0.01	0.22±0.02	0.07±0.01
RANK-1 ENSEMBLE	85.1±1.0	–	89.1±0.7	**0.02±0.00**	0.12±0.01	0.03±0.00
MFVI	79.8±0.4	–	84.3±0.4	0.07±0.01	0.12±0.03	0.07±0.01
MFVI ENSEMBLE	82.3±0.1	–	86.8±0.1	0.02±0.00	0.05±0.01	**0.02±0.00**

Table 10: Expert Referral Metrics, Country Shift, Tuned on Joint Dataset. We assess model predictive performance and uncertainty quantification in the context of expert referral. Here all methods are tuned according to the joint validation metric (Appendix B.3): area under the retention–accuracy curve constructed on the balanced joint validation dataset (composed of the in-domain and upsampled shifted validation datasets). We construct referral curves on a variety of metrics—AUC, Accuracy, NLL and AUPRC—by sweeping over the referral thresholds τ, obtaining a point for each possible partition of the dataset into "referred" and "non-referred". The Balanced evaluation dataset is constructed using the procedure described in Appendix B.3.

	R-AUROC AUC ↑				R-Accuracy AUC ↑			
Method	**In-Domain**	**Shifted**	**Joint**	**Balanced**	**In-Domain**	**Shifted**	**Joint**	**Balanced**
MAP (Deterministic)	95.1±0.5	92.1±0.5	95.6±0.3	95.6±0.3	97.3±0.1	94.6±0.3	97.1±0.1	95.8±0.1
DEEP ENSEMBLE	95.6±0.2	94.3±0.5	96.2±0.1	96.2±0.1	97.8±0.0	95.5±0.2	97.6±0.0	96.5±0.1
MC DROPOUT	96.5±0.2	95.4±0.4	96.9±0.2	96.9±0.2	97.7±0.0	95.3±0.3	97.5±0.1	96.4±0.1
MC DROPOUT ENSEMBLE	**96.9±0.1**	95.8±0.2	**97.2±0.0**	**97.2±0.0**	**98.1±0.0**	95.6±0.1	**97.9±0.0**	96.7±0.1
FSVI	93.0±0.4	97.0±0.3	94.0±0.3	94.0±0.3	97.1±0.1	96.4±0.2	97.0±0.1	96.7±0.1
FSVI ENSEMBLE	93.2±0.3	97.9±0.2	94.3±0.3	94.3±0.3	97.4±0.1	**97.0±0.1**	97.4±0.0	**97.2±0.0**
RADIAL-MFVI	90.0±0.7	96.9±0.3	91.3±0.6	91.3±0.6	96.0±0.1	95.8±0.3	96.0±0.1	95.8±0.1
RADIAL-MFVI ENSEMBLE	89.6±0.4	**98.0±0.1**	91.1±0.4	91.1±0.4	96.4±0.1	96.8±0.1	96.4±0.0	96.5±0.0
RANK-1	93.9±0.5	93.3±0.5	94.5±0.4	94.5±0.4	96.4±0.2	94.2±0.2	96.2±0.2	95.2±0.1
RANK-1 ENSEMBLE	95.1±0.3	95.5±0.1	95.8±0.2	95.8±0.2	97.4±0.1	95.4±0.1	97.3±0.1	96.4±0.1
MFVI	92.1±0.8	95.0±0.6	93.1±0.6	93.1±0.6	96.0±0.2	94.9±0.3	96.0±0.1	95.5±0.1
MFVI ENSEMBLE	91.8±0.8	97.4±0.4	93.3±0.6	93.3±0.6	96.9±0.1	96.5±0.2	96.9±0.1	96.7±0.1
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	46.3±2.8	161.5±10.9	54.8±1.9	109.5±5.1	91.4±1.0	96.1±0.1	92.4±0.7	92.4±0.7
DEEP ENSEMBLE	25.0±0.2	113.0±9.8	31.4±0.9	72.4±5.6	92.0±0.5	**96.6±0.1**	93.0±0.3	93.0±0.3
MC DROPOUT	11.3±0.3	48.6±5.2	13.9±0.6	31.1±2.9	93.6±0.4	94.6±0.2	93.8±0.3	93.8±0.3
MC DROPOUT ENSEMBLE	**8.3±0.1**	34.9±2.5	**10.2±0.3**	22.4±1.4	**94.3±0.2**	94.5±0.2	**94.4±0.1**	**94.4±0.1**
FSVI	20.7±1.4	39.1±4.1	22.0±1.3	30.3±2.1	86.9±1.1	95.0±0.5	88.8±0.7	88.8±0.7
FSVI ENSEMBLE	14.6±0.7	21.2±1.7	15.1±0.6	18.1±0.7	86.3±0.9	95.4±0.3	88.5±0.6	88.5±0.6
RADIAL-MFVI	25.8±4.2	38.7±11.0	26.8±4.7	32.8±7.6	80.5±1.7	95.3±0.1	83.4±1.4	83.4±1.4
RADIAL-MFVI ENSEMBLE	14.9±0.2	**13.0±1.0**	14.8±0.2	**14.3±0.5**	79.0±1.0	95.9±0.1	82.4±0.9	82.4±0.9
RANK-1	42.8±3.6	132.7±15.8	49.2±3.8	91.1±9.4	89.4±0.9	95.5±0.0	90.6±0.7	90.6±0.7
RANK-1 ENSEMBLE	17.0±0.9	69.9±7.2	20.6±1.2	44.7±4.1	91.1±0.4	95.8±0.1	92.0±0.3	92.0±0.3
MFVI	51.7±5.4	106.8±14.6	55.5±5.9	80.6±9.8	85.2±1.9	95.5±0.3	87.4±1.4	87.4±1.4
MFVI ENSEMBLE	20.5±2.1	42.4±10.3	22.0±2.6	31.8±6.3	84.0±1.7	96.3±0.2	87.2±1.2	87.2±1.2

Table 11: Expert Referral Metrics, Severity Shift, Tuned on Joint Dataset. We assess model predictive performance and uncertainty quantification in the context of expert referral. Here all methods are tuned according to the joint validation metric (Appendix B.3): area under the retention–accuracy curve constructed on the balanced joint validation dataset (composed of the in-domain and upsampled shifted validation datasets). We construct referral curves on a variety of metrics—AUC, Accuracy, NLL and AUPRC—by sweeping over the referral thresholds τ, obtaining a point for each possible partition of the dataset into "referred" and "non-referred". The Balanced evaluation dataset is constructed using the procedure described in Appendix B.3.

	R-AUROC AUC ↑				R-Accuracy AUC ↑			
Method	**In-Domain**	**Shifted**	**Joint**	**Balanced**	**In-Domain**	**Shifted**	**Joint**	**Balanced**
MAP (Deterministic)	93.0±0.4	–	95.2±0.3	95.2±0.3	96.6±0.2	95.9±0.4	96.6±0.2	96.3±0.2
DEEP ENSEMBLE	93.4±0.2	–	95.9±0.1	95.9±0.1	97.4±0.1	96.4±0.1	97.3±0.1	97.0±0.1
MC DROPOUT	95.7±0.4	–	97.0±0.3	97.0±0.3	97.3±0.2	97.9±0.3	97.3±0.2	97.6±0.3
MC DROPOUT ENSEMBLE	**96.5±0.1**	–	**97.7±0.0**	**97.7±0.0**	**98.0±0.0**	**98.3±0.1**	**98.0±0.0**	**98.2±0.0**
FSVI	92.7±0.4	–	94.9±0.2	94.9±0.2	96.5±0.2	95.7±0.3	96.4±0.2	96.0±0.2
FSVI ENSEMBLE	93.4±0.1	–	95.8±0.1	95.8±0.1	97.4±0.0	95.8±0.1	97.3±0.0	96.7±0.1
RADIAL-MFVI	85.6±1.5	–	88.5±1.5	88.5±1.5	95.9±0.2	78.4±4.4	95.2±0.3	89.5±1.6
RADIAL-MFVI ENSEMBLE	85.3±1.0	–	88.4±1.1	88.4±1.1	96.5±0.2	78.5±3.0	95.9±0.3	90.2±1.1
RANK-1	90.0±1.1	–	92.9±0.8	92.9±0.8	96.7±0.2	88.7±1.7	96.3±0.3	93.5±0.7
RANK-1 ENSEMBLE	89.3±0.4	–	93.1±0.4	93.1±0.4	97.3±0.2	89.4±0.8	96.9±0.2	94.3±0.4
MFVI	91.5±0.3	–	94.1±0.3	94.1±0.3	95.3±0.0	94.4±0.8	95.2±0.0	94.9±0.4
MFVI ENSEMBLE	92.7±0.0	–	95.2±0.0	95.2±0.0	96.2±0.1	95.9±0.2	96.2±0.1	96.1±0.1
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	50.2±6.2	49.0±8.0	49.9±6.3	48.8±6.9	85.9±0.9	–	92.2±0.5	92.2±0.5
DEEP ENSEMBLE	16.2±2.5	14.8±2.7	15.9±2.5	15.1±2.6	86.5±0.6	–	93.2±0.3	93.2±0.3
MC DROPOUT	11.9±0.8	7.1±1.3	11.4±0.9	9.4±1.1	91.0±0.7	–	95.2±0.5	95.2±0.5
MC DROPOUT ENSEMBLE	**8.3±0.1**	**4.7±0.2**	**8.0±0.1**	**6.3±0.1**	**92.4±0.2**	–	**96.2±0.1**	**96.2±0.1**
FSVI	16.8±1.1	15.6±1.4	16.8±1.1	16.4±1.1	85.9±0.8	–	92.2±0.5	92.2±0.5
FSVI ENSEMBLE	10.4±0.1	11.6±0.3	10.5±0.1	10.8±0.2	86.6±0.4	–	93.1±0.2	93.1±0.2
RADIAL-MFVI	14.4±0.7	37.5±6.7	15.3±0.9	22.9±2.8	69.8±3.1	–	78.1±3.2	78.1±3.2
RADIAL-MFVI ENSEMBLE	13.3±0.4	36.9±4.0	14.2±0.5	21.9±1.6	66.1±2.1	–	75.1±2.5	75.1±2.5
RANK-1	25.1±2.8	50.4±7.7	26.4±3.0	35.5±4.5	79.7±2.2	–	87.7±1.5	87.7±1.5
RANK-1 ENSEMBLE	10.1±0.5	25.2±1.7	10.8±0.6	15.9±0.9	77.5±0.9	–	87.3±0.7	87.3±0.7
MFVI	24.4±3.4	30.7±8.4	24.9±3.8	27.4±5.9	82.3±0.5	–	89.8±0.4	89.8±0.4
MFVI ENSEMBLE	12.9±0.1	10.3±0.4	12.6±0.1	11.4±0.1	84.6±0.0	–	91.7±0.0	91.7±0.0

B.5 Effect of Class Balancing the APTOS Dataset (Figure 16 and 17).

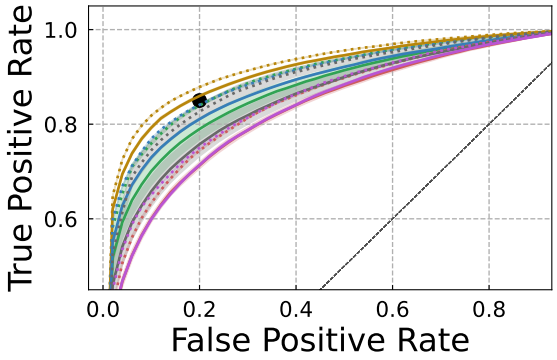
We additionally investigated to what extent the change in class distribution—in terms of the ground-truth clinical labels ranging from 0 (No DR) to 4 (Proliferative DR)—contributed to the higher performance of models in AUC, and weaker performance of models in selective prediction on the APTOS dataset (the distributionally shifted dataset in the *Country Shift* task) than the in-domain test dataset.

In order to normalize for the change in class distribution, we constructed a variant of the APTOS dataset with the same clinical class proportions as the in-domain EyePACS dataset. This was done by randomly sampling APTOS examples from each class, weighted by the empirical class probability of the EyePACS dataset, until reaching 10,000 samples.

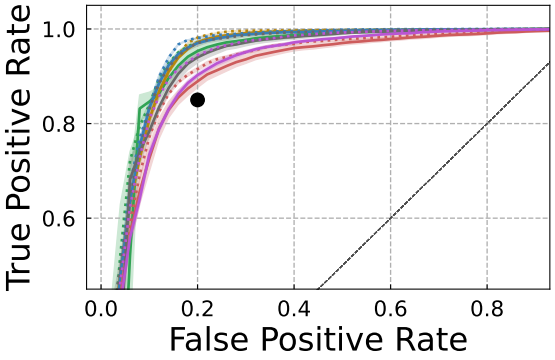
In Figure 16, we see that the ROC curves of models on the rebalanced APTOS dataset is shifted further towards the upper left as compared to the original APTOS dataset. This suggests that the class proportions of the original APTOS dataset were not the reason why models obtained stronger ROC performance on APTOS than the in-domain test set—on the contrary, introducing the in-domain class proportions in the class-balanced dataset improves model performance.

In Figure 17, we observe that the selective prediction performance of models on this rebalanced APTOS dataset is slightly better than on the original APTOS dataset, but the ordering of models does not notably change, and performance is still significantly worse at high referral thresholds than on the in-domain data.

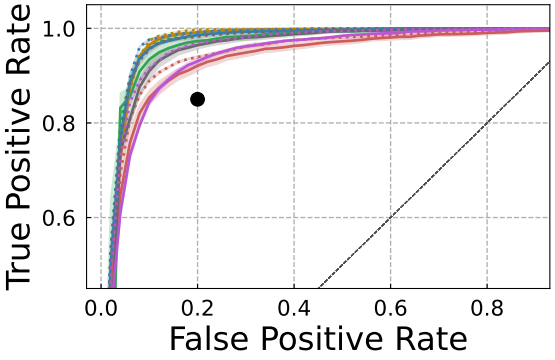
This supports the claim that factors other than simply a changed class distribution, such as meaningful shifts in equipment or patient demographics, result in both stronger predictive performance at 0% of data referred and poor quality of uncertainty estimates in the shifted setting.



(a) ROC: In-Domain

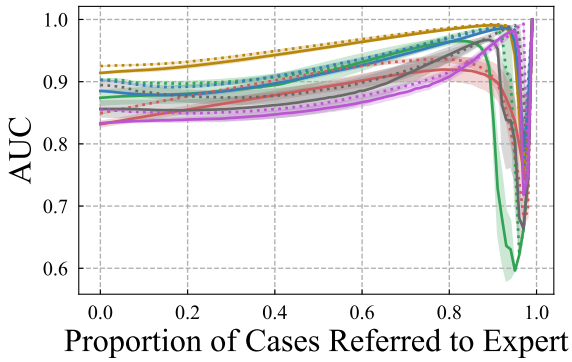


(b) ROC: Country Shift

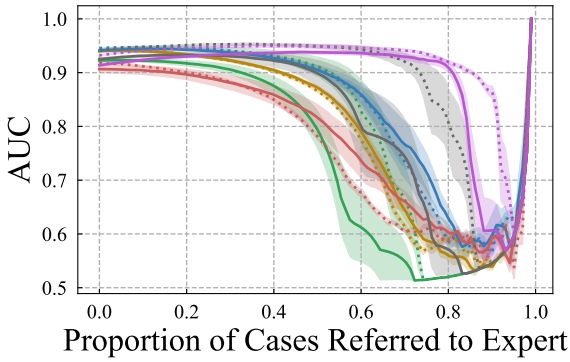


(c) ROC: Class-Balanced Country Shift

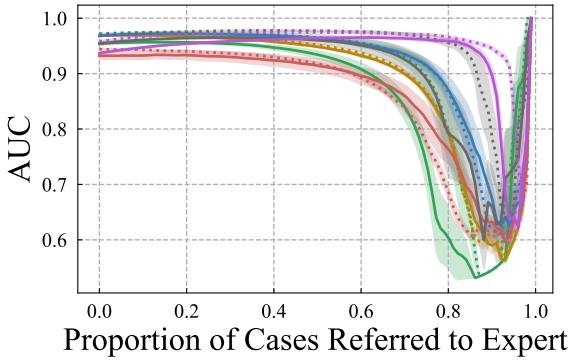
Figure 16: Class Balancing the Country Shift Dataset (ROC Curves). We consider how balancing the proportions of the ground-truth clinical class labels—ranging from 0 (No DR) to 4 (Proliferative DR)—affects performance on the *Country Shift* receiver-operating characteristic (ROC) curve. (**a**): ROC curve on in-domain test data. (**b**): ROC curve for changing medical equipment and patient populations on the shifted APTOS [3] test set. (**c**): ROC curve on the class rebalanced APTOS dataset. Shading denotes one standard error.



(a) Selective Prediction AUC: In-Domain



(b) Selective Prediction AUC: Country Shift



(c) Selective Prediction AUC: Class-Balanced Country Shift

Figure 17: Class Balancing the Country Shift Dataset (Selective Prediction). We consider how balancing the proportions of the ground-truth clinical class labels—ranging from 0 (No DR) to 4 (Proliferative DR)—affects performance on the *Country Shift* selective prediction over AUC. (**a**): selective prediction AUC on in-domain test data. (**b**): selective prediction AUC for changing medical equipment and patient populations on the shifted APTOS [3] test set. (**c**): selective prediction AUC on the class rebalanced APTOS dataset. Shading denotes one standard error.

B.6 Effect of Preprocessing on Downstream Tasks

Preprocessing played an important role in the EyePACS Kaggle challenge [13]. Here, we investigate how changes in preprocessing affect downstream predictive performance and uncertainty quantification.

In the above experiments, we used the preprocessing procedure of the Kaggle competition winner which consisted of the following steps:

1. Rescaling the images such that the retinas have a radius of 300 pixels,
2. Subtracting the local average color, computed using Gaussian blur, and finally,
3. Clipping the images to 90% size to remove "boundary effects".

While (1) and (3) are (somewhat) standard techniques used to make the data more amenable for use in non-convex optimization, the standard deviation hyperparameter of the Gaussian blur kernel in (2) presupposes some amount of expert knowledge as the size of the standard deviation governs how visible certain visual artifacts are. As such, varying it has a dramatic visual effect on the preprocessed image, and likely required significant tuning.

In the preprocessing procedure, the standard deviation of the kernel is computed as $\sigma = (\texttt{target\_radius}/\texttt{blur\_constant})$, where by default, $\texttt{target\_radius} = 300$ and $\texttt{blur\_constant} = 30$.

Decreasing the `blur_constant` results in a larger kernel standard deviation, and hence the local average color at each pixel location is computed using a larger window. This ultimately results in the preservation of more signal as well as more noise in the input image (because lower-frequency patterns are subtracted). See Figure 18 for examples of unprocessed retina images along with processed images with various blur constants.

We test the downstream performance of MAP estimation (a deterministic model), a DEEP ENSEMBLE, MC DROPOUT, and an MC DROPOUT ENSEMBLE on the Country and Severity Shift prediction tasks, varying the $\texttt{blur\_constant} \in \{5, 10, 20, 30\}$.

Severity Shift: Varying Blur Constant (Figure 19, Table 12). On the in-domain evaluation dataset, higher `blur_constant` (corresponding to stronger smoothing) tends to perform better across MAP and MC DROPOUT, single and ensembled models, and the various referral thresholds. However, on the Severity Shift (distributionally shifted evaluation dataset), the MC DROPOUT variants perform better with *lower* `blur_constant`. This highlights the importance for practitioners to test changes in experimental settings, including preprocessing, across a variety of uncertainty quantification methods.

Country Shift: Varying Blur Constant (Figure 20, Table 13). Similarly to the Severity Shift results, higher `blur_constant` tends to perform better on the in-domain evaluation data across methods and referral rates. Notably, on the distributionally shifted APTOS data, DEEP ENSEMBLE outperforms MC DROPOUT ENSEMBLE, and $\texttt{blur\_constant} = 20$ significantly improves performance from the default $\texttt{blur\_constant} = 30$ for DEEP ENSEMBLE between referral rates 0.4 and 0.7. For example, for DEEP ENSEMBLE at $\tau = 0.7$, we observe 82.2 ± 2.5 AUC with $\texttt{blur\_constant} = 20$ versus 67.4 ± 5.6 AUC with $\texttt{blur\_constant} = 30$.

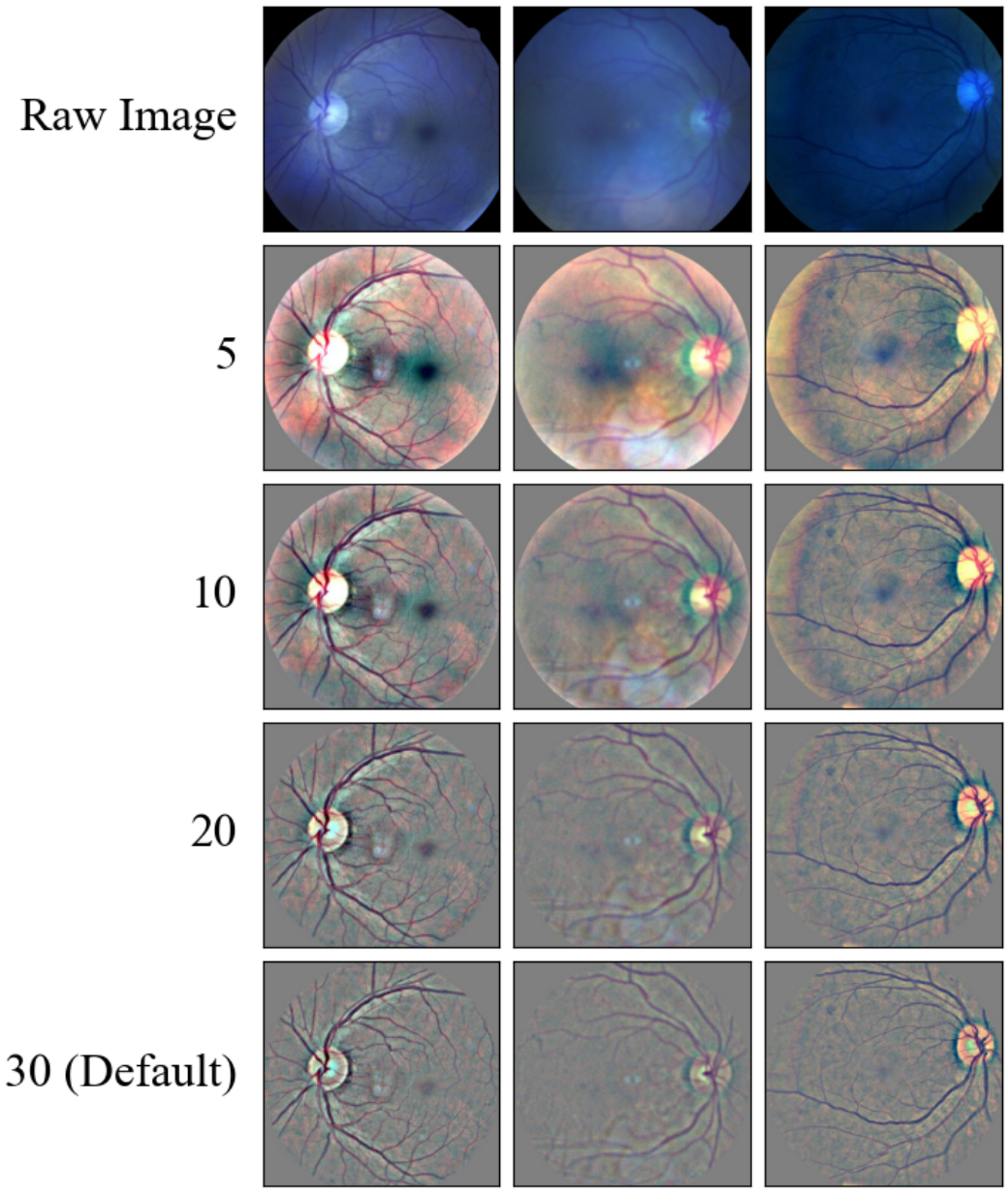
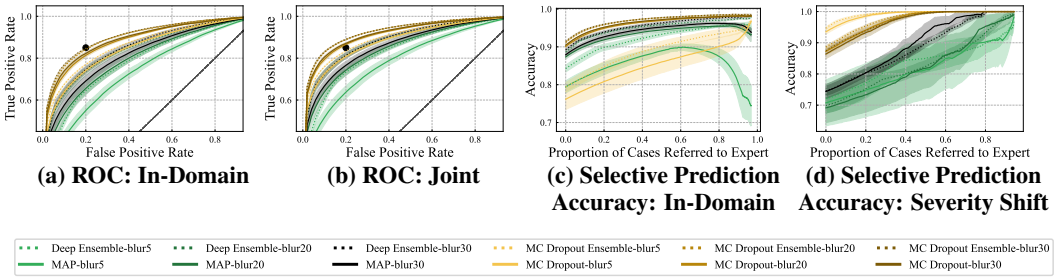
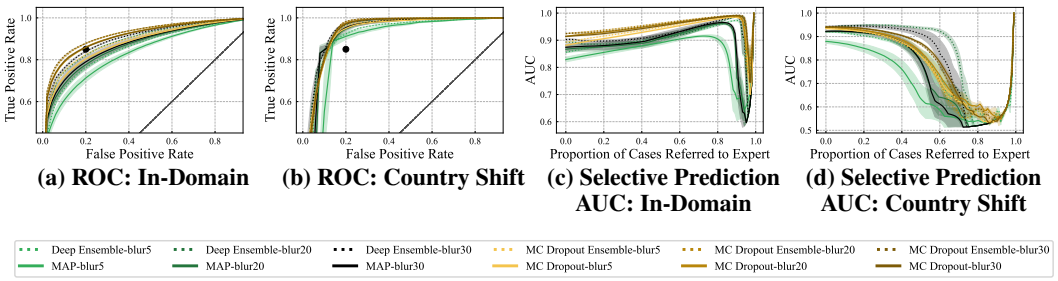


Figure 18: Preprocessing Examples. Input unprocessed EyePACS images (top row), and images processed with varying `blur_constant` (labeled on left side of grid). Higher `blur_constant` corresponds to stronger smoothing.



(a) ROC: In-Domain **(b) ROC: Joint** **(c) Selective Prediction Accuracy: In-Domain** **(d) Selective Prediction Accuracy: Severity Shift**

Figure 19: Severity Shift, Varying Blur Constant. We consider how preprocessing affects model predictive performance and uncertainty quantification on the in-domain test dataset composed only of cases with either no, mild, or moderate diabetic retinopathy, and the *Severity Shift* evaluation set composed only of severe and proliferate cases. **Left:** The receiver operating characteristic curve (ROC) for in-domain diagnosis (**a**) and for a joint dataset composed of examples from both the in-domain and *Severity Shift* evaluation sets (**b**). The dot in **black** denotes the NHS-recommended 85% sensitivity and 80% specificity ratios [63]. **Right:** Selective prediction on accuracy in the in-domain (**c**) and *Severity Shift* (**d**) settings. Shading denotes standard error computed over six random seeds. We vary the standard deviation hyperparameter of the Gaussian blur kernel through a `blur_constant` (e.g., blur5 below corresponds to `blur_constant` $= 5$). A higher `blur_constant` results in a stronger smoothing of the image as per the preprocessing procedure outlined in Appendix B.4. The default `blur_constant` used in other experiments throughout this work is 30.



(a) ROC: In-Domain **(b) ROC: Country Shift** **(c) Selective Prediction AUC: In-Domain** **(d) Selective Prediction AUC: Country Shift**

Figure 20: Country Shift, Varying Blur Constant. We consider how preprocessing affects model predictive performance and uncertainty quantification on both in-domain and distributionally shifted data. **Left:** The *receiver operating characteristic curve (ROC)* for in-population diagnosis on the EyePACS [13] test set (**a**) and for changing medical equipment and patient populations on the APTOS [3] test set (**b**). The dot in **black** denotes the NHS-recommended 85% sensitivity and 80% specificity ratios [63]. **Right:** *selective prediction* on AUC in the EyePACS [13] (**c**) and the APTOS [3] (**d**) settings. Shading denotes standard error computed over six random seeds. We vary the standard deviation hyperparameter of the Gaussian blur kernel through a `blur_constant` (e.g., blur5 below corresponds to `blur_constant` $= 5$). A higher `blur_constant` results in a stronger smoothing of the image as per the preprocessing procedure outlined in Appendix B.4. The default `blur_constant` used in other experiments throughout this work is 30.

Table 12: Severity Shift, Varying Blur Constant. We consider how preprocessing affects downstream prediction and uncertainty quality of baseline methods in terms of the area under the receiver operating characteristic curve (AUC) and classification accuracy, as a function of the proportion of data referred to a medical expert for further review. All methods are tuned on in-domain validation AUC, and ensembles have $K = 3$ constituent models. We vary the standard deviation hyperparameter of the Gaussian blur kernel through a `blur_constant` (e.g., blur5 below corresponds to `blur_constant` $= 5$). A higher `blur_constant` results in a stronger smoothing of the image as per the preprocessing procedure outlined in Appendix B.4. The default `blur_constant` used in other experiments is 30.

	No Referral		50% Data Referred		70% Data Referred	
Method	**AUC (%)** ↑	**Accuracy (%)** ↑	**AUC (%)** ↑	**Accuracy (%)** ↑	**AUC (%)** ↑	**Accuracy** ↑
In-Domain (No, Mild, or Moderate DR, Clinical Labels {0,1,2})						
MAP (Deterministic)-blur5	73.7±1.3	79.4±1.4	75.5±3.1	89.0±0.9	79.1±3.4	89.3±1.0
MAP (Deterministic)-blur10	78.7±1.1	84.6±0.6	80.0±2.3	93.4±0.3	84.5±2.2	94.1±0.4
MAP (Deterministic)-blur20	79.9±1.3	87.3±0.5	77.2±3.4	94.5±0.4	80.9±4.1	95.3±0.3
MAP (Deterministic)-blur30	82.0±1.0	87.9±0.4	83.1±1.9	95.2±0.3	88.4±1.9	96.0±0.2
MC DROPOUT-blur5	84.8±0.4	76.1±2.3	91.4±0.3	86.0±2.4	94.1±0.4	88.6±2.2
MC DROPOUT-blur10	86.3±0.1	84.2±1.3	92.4±0.4	93.5±0.8	95.2±0.2	95.1±0.6
MC DROPOUT-blur20	88.7±0.3	90.1±0.2	92.5±0.5	97.0±0.1	95.3±0.3	97.7±0.1
MC DROPOUT-blur30	89.2±0.2	90.5±0.1	92.8±0.6	97.2±0.0	95.4±0.4	97.8±0.0
DEEP ENSEMBLE-blur5	78.6±0.6	84.3±0.8	75.0±2.6	93.3±0.5	75.9±3.3	94.8±0.3
DEEP ENSEMBLE-blur10	82.4±0.3	87.7±0.1	80.9±1.3	95.1±0.1	84.1±1.3	96.1±0.1
DEEP ENSEMBLE-blur20	84.2±0.8	88.6±0.3	70.9±1.1	95.8±0.2	71.4±1.4	96.7±0.2
DEEP ENSEMBLE-blur30	85.1±0.7	89.3±0.2	82.0±0.9	96.3±0.2	85.3±0.9	97.3±0.2
MC DROPOUT ENSEMBLE-blur5	86.5±0.1	79.4±1.0	93.2±0.1	90.2±1.1	95.7±0.2	92.5±0.9
MC DROPOUT ENSEMBLE-blur10	87.5±0.0	86.7±0.6	**93.4±0.2**	95.4±0.3	**96.0±0.2**	96.5±0.3
MC DROPOUT ENSEMBLE-blur20	90.3±0.0	91.1±0.1	**93.5±0.2**	97.6±0.0	**96.0±0.1**	**98.2±0.0**
MC DROPOUT ENSEMBLE-blur30	**90.6±0.0**	**91.4±0.1**	93.1±0.2	**97.8±0.0**	95.7±0.2	**98.2±0.0**
Severity Shift (Severe or Proliferate DR, Clinical Labels {3, 4})						
MAP (Deterministic)-blur5	–	70.8±6.2	–	81.4±7.9	–	87.7±7.9
MAP (Deterministic)-blur10	–	77.3±2.2	–	91.9±2.8	–	97.2±1.5
MAP (Deterministic)-blur20	–	69.1±4.0	–	81.8±5.3	–	88.8±4.4
MAP (Deterministic)-blur30	–	74.4±1.9	–	93.2±2.6	–	98.6±1.1
MC DROPOUT-blur5	–	93.5±0.6	–	**100.0±0.0**	–	**100.0±0.0**
MC DROPOUT-blur10	–	91.0±1.3	–	99.9±0.0	–	**100.0±0.0**
MC DROPOUT-blur20	–	87.2±0.9	–	99.7±0.1	–	**100.0±0.0**
MC DROPOUT-blur30	–	86.4±1.3	–	99.5±0.2	–	**100.0±0.0**
DEEP ENSEMBLE-blur5	–	72.0±3.9	–	85.1±3.7	–	87.5±3.3
DEEP ENSEMBLE-blur10	–	80.0±1.2	–	94.0±1.0	–	97.8±0.5
DEEP ENSEMBLE-blur20	–	69.8±2.1	–	82.4±1.5	–	89.1±1.5
DEEP ENSEMBLE-blur30	–	74.5±1.2	–	89.8±1.0	–	97.0±0.7
MC DROPOUT ENSEMBLE-blur5	–	**94.7±0.3**	–	**100.0±0.0**	–	**100.0±0.0**
MC DROPOUT ENSEMBLE-blur10	–	91.9±0.7	–	**100.0±0.0**	–	**100.0±0.0**
MC DROPOUT ENSEMBLE-blur20	–	88.6±0.4	–	99.8±0.0	–	**100.0±0.0**
MC DROPOUT ENSEMBLE-blur30	–	87.4±0.3	–	99.4±0.1	–	**100.0±0.0**

Table 13: Country Shift, Varying Blur Constant. We consider how preprocessing affects downstream prediction and uncertainty quality of baseline methods in terms of the area under the receiver operating characteristic curve (AUC) and classification accuracy, as a function of the proportion of data referred to a medical expert for further review. All methods are tuned on in-domain validation AUC, and ensembles have $K = 3$ constituent models. We vary the standard deviation hyperparameter of the Gaussian blur kernel through a `blur_constant` (e.g., blur5 below corresponds to `blur_constant` $= 5$). A higher `blur_constant` results in a stronger smoothing of the image as per the preprocessing procedure outlined in Appendix B.4. The default `blur_constant` used in other experiments is 30.

	No Referral		50% Data Referred		70% Data Referred	
Method	**AUC (%)** ↑	**Accuracy (%)** ↑	**AUC (%)** ↑	**Accuracy (%)** ↑	**AUC (%)** ↑	**Accuracy** ↑
EyePACS Dataset (In-Domain)						
MAP (Deterministic)-blur5	82.9±0.7	80.3±0.8	89.1±0.6	91.0±0.7	91.4±0.3	91.6±0.6
MAP (Deterministic)-blur10	87.1±0.1	85.6±0.3	92.6±0.1	95.0±0.2	95.0±0.2	94.9±0.3
MAP (Deterministic)-blur20	86.7±1.0	88.0±0.5	90.5±1.4	95.6±0.3	94.4±0.9	96.3±0.2
MAP (Deterministic)-blur30	87.4±1.0	88.6±0.6	91.1±1.4	95.9±0.3	94.9±0.8	96.5±0.2
MC DROPOUT-blur5	88.1±0.2	85.9±0.4	94.0±0.1	95.0±0.2	96.5±0.1	96.4±0.1
MC DROPOUT-blur10	89.0±0.2	85.5±0.5	94.7±0.2	94.9±0.3	96.9±0.1	96.3±0.2
MC DROPOUT-blur20	91.4±0.1	90.2±0.2	95.7±0.2	97.3±0.1	97.5±0.1	98.0±0.1
MC DROPOUT-blur30	91.4±0.1	90.9±0.0	95.3±0.2	97.4±0.0	97.4±0.1	98.1±0.0
DEEP ENSEMBLE-blur5	85.6±0.2	84.6±0.1	90.9±0.3	94.3±0.0	93.6±0.3	95.8±0.2
DEEP ENSEMBLE-blur10	88.8±0.0	88.0±0.1	94.2±0.1	96.2±0.0	96.4±0.1	97.3±0.0
DEEP ENSEMBLE-blur20	89.2±0.2	89.5±0.2	90.5±0.3	96.9±0.1	93.8±0.3	97.7±0.0
DEEP ENSEMBLE-blur30	90.3±0.1	90.3±0.2	91.7±0.5	97.2±0.0	95.0±0.4	97.9±0.0
MC DROPOUT ENSEMBLE-blur5	89.3±0.0	87.3±0.1	94.7±0.0	95.7±0.1	97.1±0.0	96.9±0.0
MC DROPOUT ENSEMBLE-blur10	90.1±0.0	87.4±0.1	95.4±0.0	96.0±0.0	97.3±0.0	97.0±0.1
MC DROPOUT ENSEMBLE-blur20	92.4±0.0	91.2±0.0	**96.2±0.1**	97.7±0.0	**97.9±0.0**	98.3±0.0
MC DROPOUT ENSEMBLE-blur30	**92.5±0.0**	**91.6±0.0**	95.8±0.1	**97.8±0.0**	97.7±0.1	**98.4±0.0**
APTOS 2019 Dataset (Shifted)						
MAP (Deterministic)-blur5	87.9±0.7	69.9±1.4	64.0±5.3	78.6±1.7	55.3±3.2	78.9±1.9
MAP (Deterministic)-blur10	90.2±0.2	77.0±0.7	63.1±2.0	81.1±0.6	51.1±0.0	80.0±0.6
MAP (Deterministic)-blur20	92.1±0.2	85.2±0.3	79.8±3.8	87.9±1.5	60.0±4.6	86.0±1.2
MAP (Deterministic)-blur30	92.2±0.2	86.2±0.4	80.1±2.8	87.6±1.1	55.4±3.3	85.4±0.9
MC DROPOUT-blur5	93.4±0.2	78.4±0.7	82.2±0.4	84.6±0.1	62.5±0.6	88.0±0.5
MC DROPOUT-blur10	93.3±0.2	77.3±0.9	79.7±0.3	83.3±0.3	59.6±1.0	87.2±0.4
MC DROPOUT-blur20	93.9±0.1	84.9±0.4	83.8±1.2	86.2±0.6	63.8±2.4	87.9±0.2
MC DROPOUT-blur30	94.0±0.2	86.8±0.2	87.4±0.3	88.1±0.2	65.3±1.3	88.2±0.3
DEEP ENSEMBLE-blur5	92.1±0.1	70.8±0.6	82.5±1.7	85.0±0.3	63.2±4.2	87.1±0.6
DEEP ENSEMBLE-blur10	91.8±0.0	78.8±0.3	73.5±0.3	84.5±0.1	51.1±0.0	82.0±0.1
DEEP ENSEMBLE-blur20	**94.1±0.0**	87.0±0.1	**93.7±0.4**	**93.6±0.3**	**82.2±2.5**	**91.7±0.5**
DEEP ENSEMBLE-blur30	94.2±0.2	87.5±0.1	91.2±1.4	92.4±0.7	67.4±5.6	90.1±0.9
MC DROPOUT ENSEMBLE-blur5	93.7±0.1	80.1±0.3	81.9±0.2	84.7±0.1	63.2±0.4	87.2±0.2
MC DROPOUT ENSEMBLE-blur10	93.6±0.1	78.7±0.4	79.1±0.1	83.4±0.1	59.6±0.2	87.3±0.3
MC DROPOUT ENSEMBLE-blur20	94.0±0.0	86.4±0.3	83.3±0.7	85.7±0.3	58.3±0.9	87.6±0.1
MC DROPOUT ENSEMBLE-blur30	94.1±0.1	**87.6±0.1**	86.8±0.2	88.0±0.1	62.3±0.3	87.7±0.2