Supplementary Material

Table of Contents

Appendix A Implementation, Training, and Evaluation Details

A.1 RETINA 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 RETINA 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. RETINA is framework-agnostic. For example, FSVI is implemented in JAX, a variant of MC Dropout is in PyTorch [49] (though we use in this work a TensorFlow variant to simplify TPU tuning), and other models in raw TensorFlow [1]. This interoperability means

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

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 sweep 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. Wang et al. [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 RETINA 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]

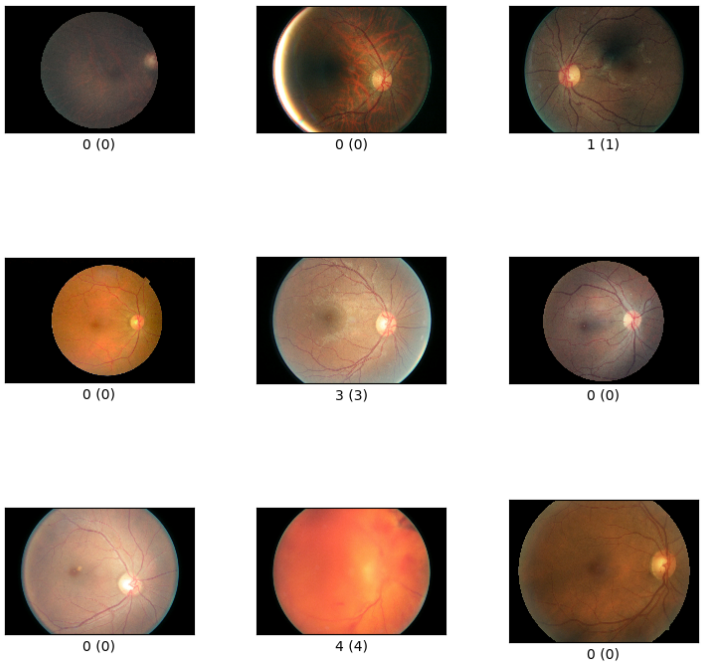
We tuned each model with a quasi-random search on several hyperparameters including learning rate, momentum, ℓ_2 regularization, and method-specific variables including dropout rate and variational posterior initializations. We used a minimum of 32 trials per model. Because of the large size of the input 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 Monte Carlo samples at training time. With a single Monte Carlo 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 Monte Carlo 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. As described in Section 6.1, for evaluation, we use five Monte Carlo samples per model to estimate predictive means (e.g., the MC DROPOUT ENSEMBLE with $K = 3$ ensemble members uses a total of $S = 15$ Monte Carlo samples).

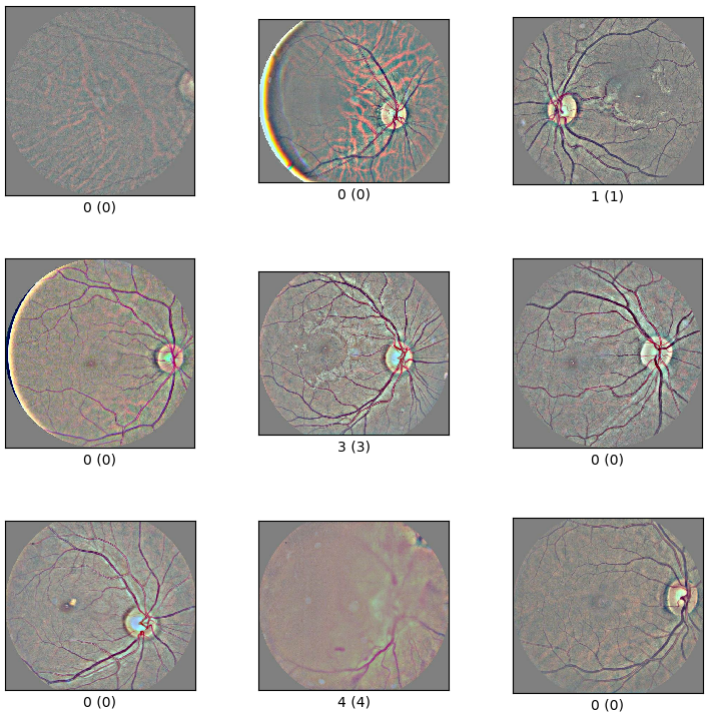
Compute Resources. The majority of methods were tuned on TPU v2-8 nodes. MFVI had particularly high memory requirements which required the use of TPU v3-8 nodes to achieve a reasonable batch size and stable training. Evaluation was performed on NVIDIA A100 GPUs with 40 GB memory, though GPUs with standard 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).

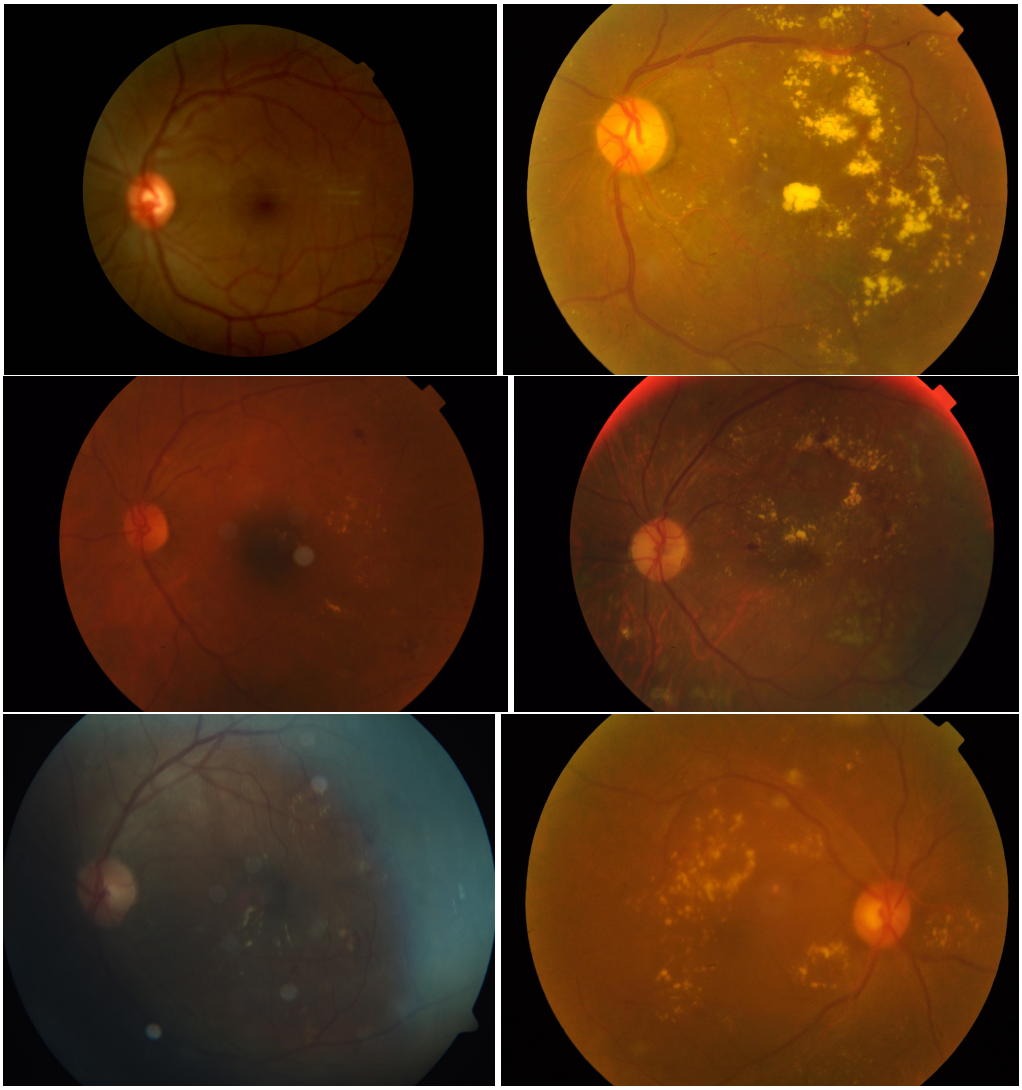


Figure 7: Illustrative examples of retina images in the APTOS dataset. The images are collected using different measurement devices than the EyePACS dataset. Note the artifacts present in the images including blur, low background lighting, and effects around the edges of the retina.

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.

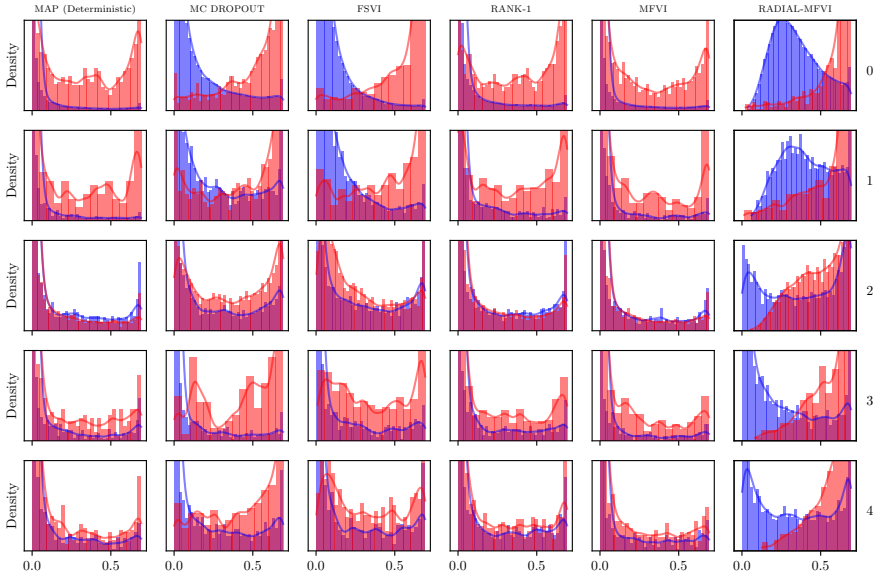


Figure 8: Clinical Label Binning – Severity Shift, Single Models. We analyze predictive uncertainty for each underlying clinical severity label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider both the in-domain and distributionally shifted Severity Shift 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.

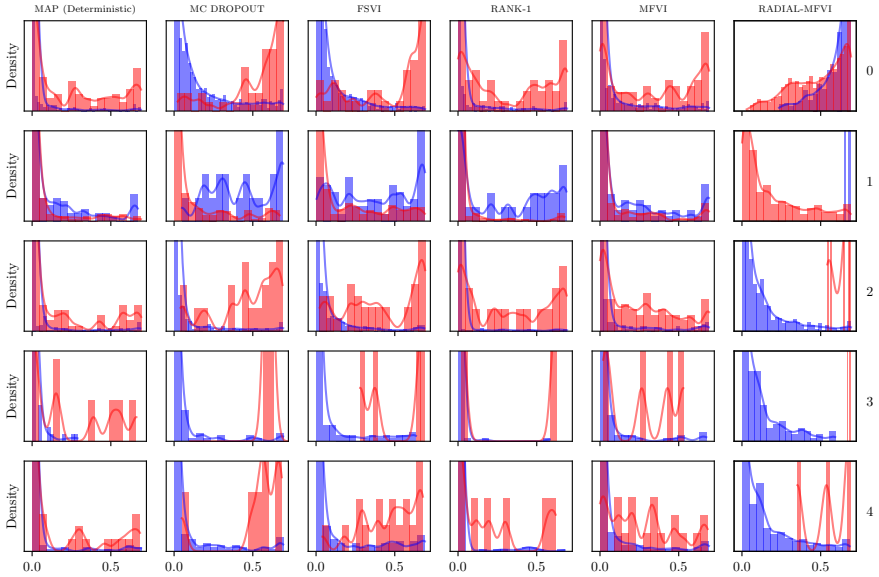


Figure 9: Clinical Label Binning – Country Shift (Shifted), Single Models. We analyze predictive uncertainty for each underlying clinical severity label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider the distributionally shifted Country Shift evaluation 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.

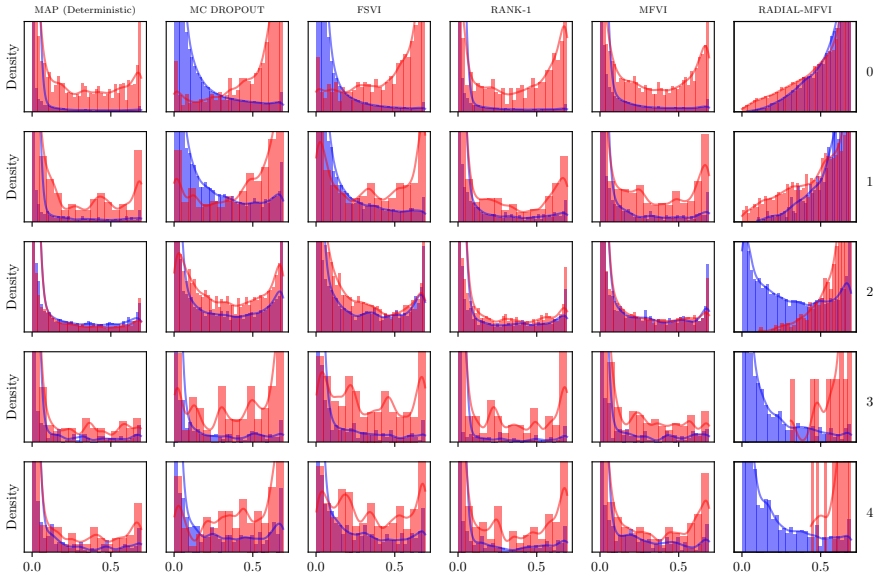


Figure 10: Clinical Label Binning – Country Shift (In-Domain), Single Models. We analyze predictive uncertainty for each ground-truth clinical label (rows) and each uncertainty quantification method (columns). Here, we consider the in-domain Country Shift evaluation dataset, and single models ($K = 1$).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

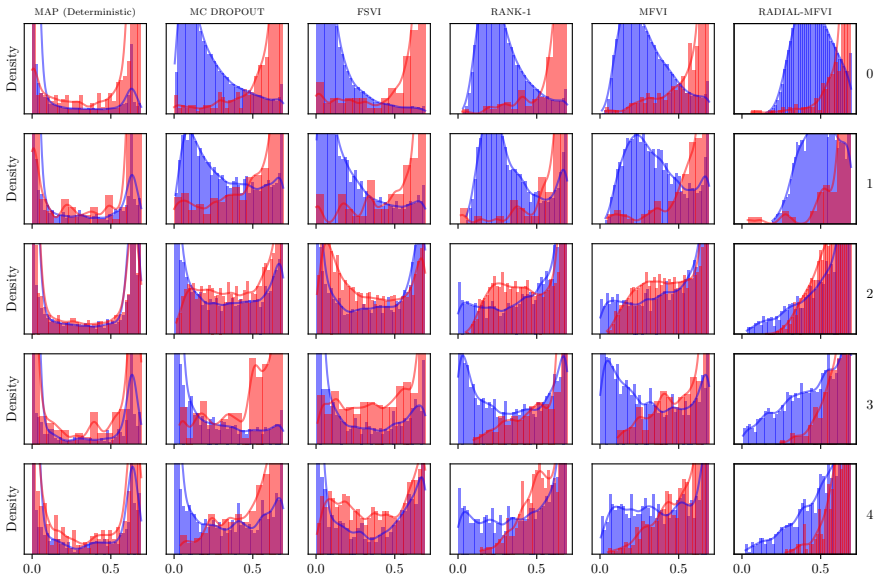


Figure 11: Clinical Label Binning – Severity Shift, Ensembles. We analyze predictive uncertainty for each ground-truth clinical label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider both the in-domain and distributionally shifted Severity Shift evaluation 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.

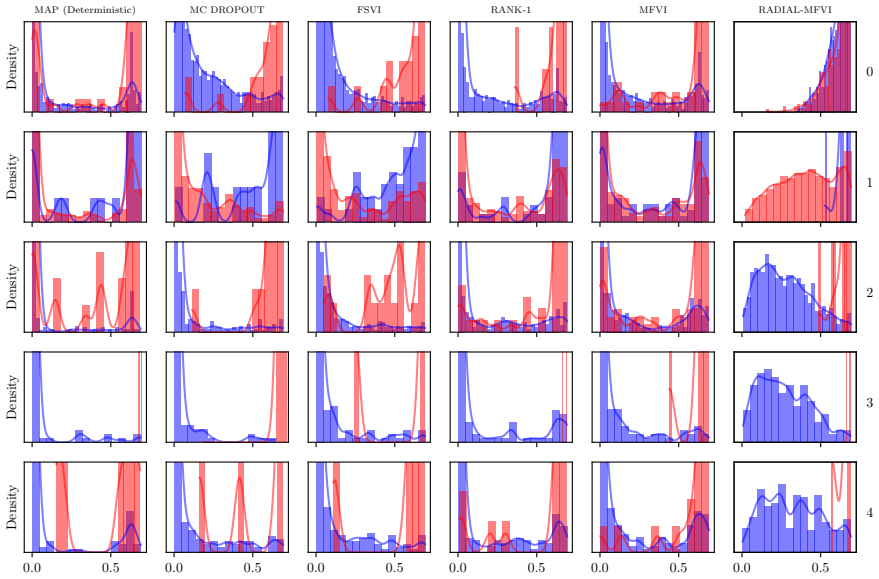


Figure 12: Clinical Label Binning – Country Shift (Shifted), Ensembles. We analyze predictive uncertainty for each ground-truth clinical label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider the distributionally shifted Country Shift evaluation dataset (APTOS), 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.

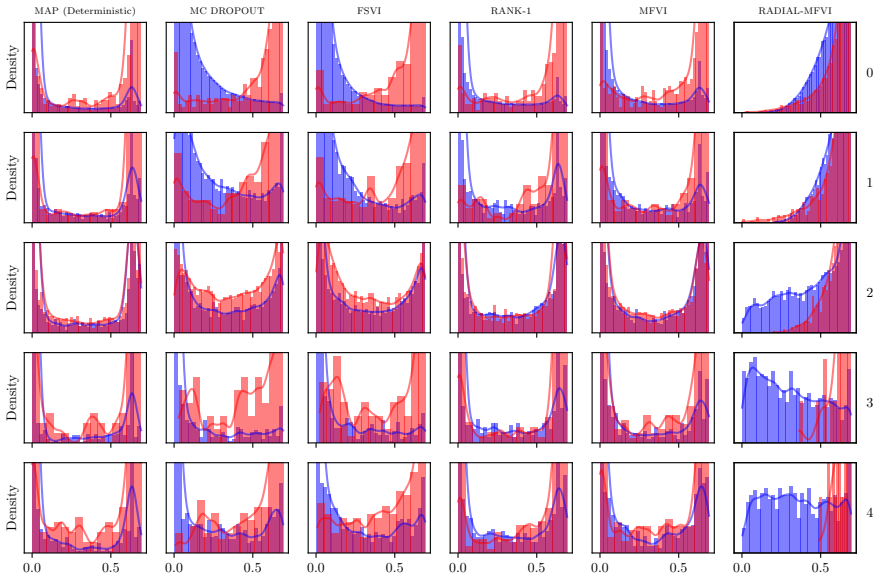
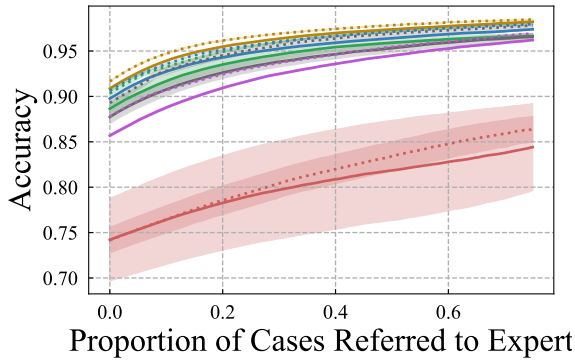


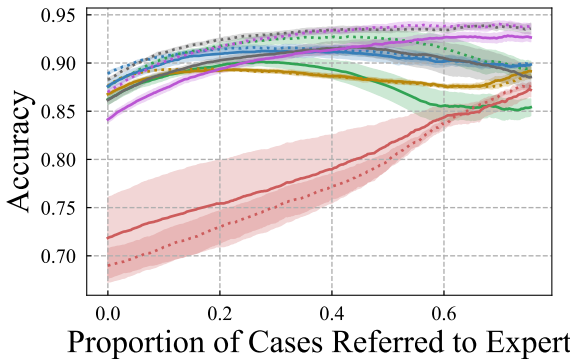
Figure 13: Clinical Label Binning – Country Shift (In-Domain), Ensembles. We analyze predictive uncertainty for each ground-truth clinical label (rows, label on right) and each uncertainty quantification method (columns). Here, we consider the in-domain Country Shift evaluation dataset (APTOS), and ensembles ($K = 3$). 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.

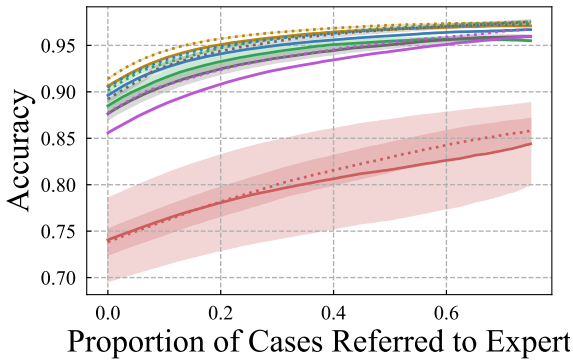
We provide referral curves on accuracy for *Country Shift* with in-domain validation tuning in Figure 14.



(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.

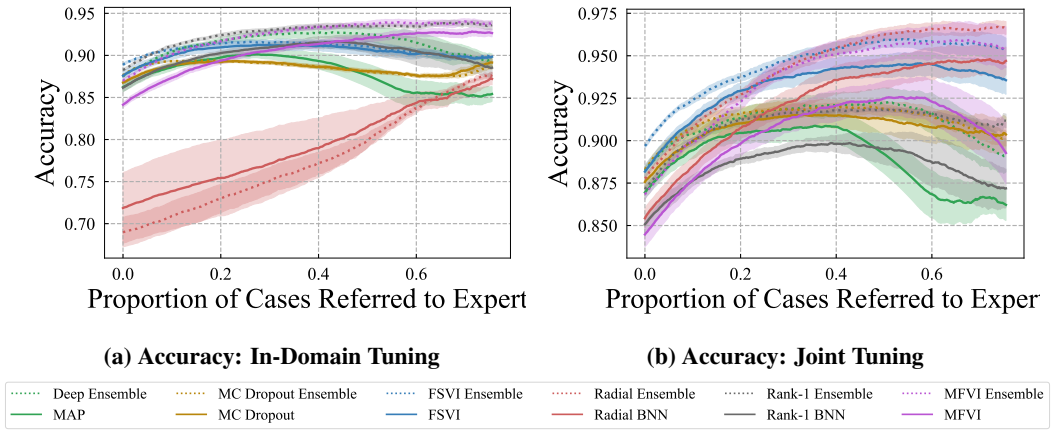
B.3 Tuning in the Presence of Distributionally Shifted Data

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±1.3	87.9±0.5	83.1±2.4	95.2±0.4	88.4±2.5	96.0±0.3
DEEP ENSEMBLE	85.1±0.9	89.3±0.3	82.0±1.1	96.3±0.3	85.3±1.2	97.3±0.2
MC DROPOUT	89.2±0.3	90.5±0.1	92.8±0.7	97.2±0.0	95.4±0.5	97.8±0.0
MC DROPOUT ENSEMBLE	**90.6±0.0**	**91.4±0.1**	**93.1±0.3**	**97.8±0.0**	**95.7±0.2**	**98.2±0.1**
FSVI	83.2±0.4	89.5±0.2	81.2±1.1	95.6±0.1	86.4±0.9	96.4±0.2
FSVI ENSEMBLE	86.2±0.1	90.0±0.1	81.2±0.4	96.4±0.0	86.1±0.4	97.3±0.0
RADIAL-MFVI	76.9±2.0	86.7±0.5	69.0±5.2	93.5±0.6	70.1±6.2	94.6±0.6
RADIAL-MFVI ENSEMBLE	81.3±1.6	87.4±0.4	66.3±3.0	95.1±0.5	66.2±3.9	96.1±0.5
RANK-1	81.6±2.0	88.3±0.7	79.4±3.7	95.1±0.6	82.9±3.7	96.0±0.5
RANK-1 ENSEMBLE	85.1±1.4	89.3±0.5	75.6±1.3	96.1±0.4	79.1±1.7	96.9±0.3
MFVI	81.3±1.7	87.8±0.7	79.5±3.1	95.0±0.5	82.6±3.4	95.9±0.4
MFVI ENSEMBLE	85.2±0.8	89.4±0.4	77.7±1.1	96.1±0.2	80.3±1.3	96.8±0.2
Severity Shift (Severe or Proliferate DR, Clinical Labels {3, 4})						
MAP (Deterministic)	–	74.4±2.5	–	93.2±3.3	–	98.6±1.4
DEEP ENSEMBLE	–	74.5±1.6	–	89.8±1.3	–	97.0±0.9
MC DROPOUT	–	86.4±1.6	–	**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±1.2	–	88.5±1.3	–	99.6±0.3
FSVI ENSEMBLE	–	69.3±0.3	–	86.3±0.6	–	99.4±0.2
RADIAL-MFVI	–	52.0±9.9	–	59.3±13.9	–	63.9±14.3
RADIAL-MFVI ENSEMBLE	–	54.4±6.1	–	58.0±9.8	–	60.6±10.7
RANK-1	–	67.5±4.5	–	82.6±5.5	–	92.7±2.9
RANK-1 ENSEMBLE	–	69.7±2.4	–	81.6±2.1	–	92.0±1.7
MFVI	–	71.5±3.0	–	86.7±4.0	–	94.1±2.6
MFVI ENSEMBLE	–	73.5±1.6	–	87.4±0.9	–	94.2±0.8

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±1.7	5.2±0.2	44.0±3.5	9.3±0.8
DEEP ENSEMBLE	41.7±1.2	5.6±0.2	56.8±1.2	12.4±0.4
MC DROPOUT	37.6±0.9	5.1±0.1	34.9±1.4	7.1±0.5
MC DROPOUT ENSEMBLE	39.5±0.3	5.3±0.0	38.3±1.2	7.7±0.3
FSVI	42.2±0.9	5.7±0.1	49.0±1.0	11.6±0.4
FSVI ENSEMBLE	43.8±0.6	5.9±0.1	54.5±0.5	14.5±0.3
RADIAL-MFVI	39.2±2.7	5.3±0.3	66.8±6.2	19.9±3.4
RADIAL-MFVI ENSEMBLE	36.5±0.8	4.9±0.1	**79.7±3.4**	**28.0±2.9**
RANK-1	44.3±2.4	6.0±0.3	54.5±4.4	12.8±1.5
RANK-1 ENSEMBLE	48.9±1.3	6.4±0.2	65.6±0.9	17.4±0.7
MFVI	51.2±0.8	6.7±0.1	51.3±3.6	10.4±0.9
MFVI ENSEMBLE	**52.4±0.4**	**6.9±0.1**	60.4±1.0	13.5±0.6

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±0.08	2.68±0.18	1.36±0.07	88.6±0.7	86.2±0.5	88.5±0.6	75.2±2.2	89.7±0.3	77.2±1.9
DEEP ENSEMBLE	0.60±0.00	1.60±0.16	0.67±0.01	90.3±0.3	87.5±0.1	90.1±0.2	79.9±0.5	**91.1±0.1**	81.0±0.4
MC DROPOUT	0.29±0.00	1.07±0.03	0.34±0.00	90.9±0.1	86.8±0.2	90.6±0.1	82.6±0.2	88.8±0.5	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±0.1	**91.4±0.0**	**84.4±0.0**	88.3±0.3	**84.3±0.1**
FSVI	0.35±0.01	0.72±0.05	0.38±0.01	89.8±0.0	87.6±0.4	89.6±0.0	77.7±0.1	88.3±0.5	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±0.07	0.70±0.09	0.57±0.07	74.2±4.5	71.8±4.2	74.1±4.5	66.0±0.9	84.8±0.8	69.0±0.8
RADIAL-MFVI ENSEMBLE	0.55±0.02	0.65±0.03	0.56±0.02	74.2±1.4	69.0±1.7	73.8±1.4	68.9±0.4	86.1±0.1	71.6±0.3
RANK-1	0.99±0.07	1.85±0.20	1.05±0.05	87.7±0.7	86.2±0.5	87.6±0.7	71.6±2.5	88.8±0.5	74.1±2.1
RANK-1 ENSEMBLE	0.49±0.04	0.96±0.06	0.52±0.03	89.3±0.4	88.3±0.1	89.2±0.4	78.0±1.3	89.6±0.3	79.3±1.1
MFVI	0.91±0.02	1.26±0.07	0.93±0.02	85.7±0.1	84.1±0.3	85.6±0.1	66.7±0.3	85.9±0.2	69.7±0.3
MFVI ENSEMBLE	0.53±0.00	0.72±0.03	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±0.1
	AUROC (%) ↑			ECE ↓					
MAP (Deterministic)	87.4±1.2	92.2±0.2	88.3±1.1	0.10±0.01	0.13±0.00	0.10±0.01			
DEEP ENSEMBLE	90.3±0.2	94.2±0.2	90.9±0.2	0.06±0.00	0.08±0.00	0.06±0.00			
MC DROPOUT	91.4±0.1	94.0±0.2	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±0.1**	90.9±0.0	0.03±0.00	0.07±0.00	0.03±0.00			
RADIAL-MFVI	83.2±0.5	90.7±0.6	84.3±0.4	0.09±0.03	0.14±0.04	0.09±0.03			
RADIAL-MFVI ENSEMBLE	84.9±0.1	91.8±0.1	85.9±0.1	0.06±0.01	0.10±0.02	0.05±0.01			
RANK-1	85.6±1.3	92.5±0.2	86.7±1.2	0.10±0.01	0.11±0.00	0.10±0.01			
RANK-1 ENSEMBLE	89.5±0.8	94.1±0.2	90.2±0.7	0.05±0.00	0.06±0.00	0.05±0.00			
MFVI	83.3±0.2	91.4±0.2	84.6±0.2	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±0.07	2.27±0.15	1.35±0.08	87.9±0.5	74.4±2.3	86.8±0.6	60.8±2.4	–	75.2±1.7
DEEP ENSEMBLE	0.62±0.02	1.03±0.06	0.65±0.03	89.3±0.3	74.5±1.5	88.1±0.4	65.6±1.5	–	79.2±1.1
MC DROPOUT	0.29±0.00	0.33±0.02	0.29±0.00	90.5±0.1	86.4±1.5	90.1±0.1	74.8±0.6	–	85.1±0.3
MC DROPOUT ENSEMBLE	**0.25±0.00**	**0.28±0.00**	**0.25±0.00**	**91.4±0.1**	**87.4±0.3**	**91.1±0.1**	**77.0±0.1**	–	**86.7±0.1**
FSVI	0.36±0.01	0.92±0.05	0.41±0.02	89.5±0.1	68.6±1.1	87.8±0.2	64.7±0.8	–	77.6±0.5
FSVI ENSEMBLE	0.31±0.00	0.76±0.01	0.34±0.00	90.0±0.1	69.3±0.3	88.4±0.1	70.0±0.2	–	81.6±0.1
RADIAL-MFVI	0.37±0.01	0.76±0.12	0.40±0.02	86.7±0.4	52.0±9.0	83.9±1.1	49.1±3.5	–	66.9±2.9
RADIAL-MFVI ENSEMBLE	0.35±0.01	0.73±0.07	0.38±0.01	87.4±0.4	54.4±5.5	84.8±0.8	56.2±2.5	–	73.5±2.0
RANK-1	0.56±0.06	1.14±0.15	0.61±0.07	88.3±0.6	67.5±4.1	86.6±0.9	59.4±3.7	–	74.1±2.5
RANK-1 ENSEMBLE	0.29±0.01	0.60±0.04	0.32±0.01	89.3±0.4	69.7±2.2	87.7±0.5	66.5±2.5	–	80.0±1.6
MFVI	0.66±0.11	1.26±0.21	0.71±0.11	87.8±0.7	71.5±2.7	86.5±0.8	59.0±3.2	–	73.7±2.3
MFVI ENSEMBLE	0.29±0.01	0.55±0.02	0.31±0.01	89.4±0.4	73.5±1.4	88.2±0.4	66.4±1.6	–	79.7±1.1
	AUROC (%) ↑			ECE ↓					
MAP (Deterministic)	82.0±1.2	–	86.3±1.0	0.11±0.00	0.23±0.02	0.12±0.01			
DEEP ENSEMBLE	85.1±0.8	–	88.9±0.6	0.06±0.00	0.15±0.01	0.07±0.00			
MC DROPOUT	89.2±0.3	–	92.0±0.2	0.02±0.00	0.06±0.01	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±0.4	–	86.9±0.3	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±1.8	–	82.2±1.6	0.05±0.01	0.23±0.07	0.04±0.01			
RADIAL-MFVI ENSEMBLE	81.3±1.4	–	86.2±1.2	0.07±0.01	0.15±0.04	0.06±0.01			
RANK-1	81.6±1.8	–	85.8±1.4	0.06±0.01	0.22±0.03	0.07±0.02			
RANK-1 ENSEMBLE	85.1±1.3	–	89.1±0.9	0.02±0.00	0.12±0.02	0.03±0.00			
MFVI	81.3±1.6	–	85.4±1.3	0.07±0.01	0.19±0.03	0.08±0.02			
MFVI ENSEMBLE	85.2±0.7	–	88.9±0.6	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)	89.2±0.9	74.7±1.7	88.8±0.5	88.8±0.5	94.9±0.4	87.3±0.8	94.0±0.3	89.9±0.5
DEEP ENSEMBLE	91.7±0.2	80.7±1.5	91.8±0.1	91.8±0.1	96.5±0.1	91.0±0.7	95.9±0.1	92.9±0.5
MC DROPOUT	94.7±0.2	79.7±0.3	93.9±0.2	93.9±0.2	96.8±0.0	88.9±0.2	95.9±0.0	91.7±0.1
MC DROPOUT ENSEMBLE	**95.4±0.1**	79.4±0.1	**94.4±0.1**	**94.4±0.1**	**97.3±0.0**	89.0±0.2	**96.4±0.0**	92.0±0.1
FSVI	91.6±0.2	83.7±1.0	92.0±0.1	92.0±0.1	95.9±0.0	90.6±0.2	95.3±0.1	92.5±0.2
FSVI ENSEMBLE	92.6±0.1	83.2±0.4	92.9±0.1	92.9±0.1	96.6±0.0	90.9±0.1	95.9±0.0	93.0±0.1
RADIAL-MFVI	87.9±1.0	77.7±1.3	88.0±1.1	88.0±1.1	82.4±5.0	81.9±2.7	82.3±4.8	81.8±3.6
RADIAL-MFVI ENSEMBLE	89.5±0.3	76.2±0.3	89.1±0.3	89.1±0.3	83.4±1.4	80.8±0.9	83.0±1.4	81.2±1.2
RANK-1	87.8±1.3	81.2±2.2	88.4±0.9	88.4±0.9	94.6±0.5	89.7±0.7	94.0±0.3	91.4±0.3
RANK-1 ENSEMBLE	90.3±0.9	88.3±1.0	91.5±0.7	91.5±0.7	96.2±0.3	**92.8±0.2**	95.9±0.3	**94.1±0.2**
MFVI	86.9±0.4	88.7±0.8	88.2±0.3	88.2±0.3	93.7±0.1	90.5±0.4	93.5±0.1	91.9±0.2
MFVI ENSEMBLE	88.1±0.3	**92.0±0.4**	89.7±0.2	89.7±0.2	94.8±0.0	92.4±0.2	94.6±0.0	93.5±0.1
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	1.22±0.09	4.10±0.33	1.54±0.04	3.09±0.22	86.4±2.2	92.3±0.3	87.7±1.6	87.7±1.6
DEEP ENSEMBLE	0.54±0.01	2.61±0.30	0.79±0.04	1.87±0.21	87.6±0.8	**94.0±0.3**	89.1±0.5	89.1±0.5
MC DROPOUT	0.19±0.01	1.87±0.08	0.36±0.01	1.22±0.04	92.1±0.3	91.6±0.6	91.5±0.2	91.5±0.2
MC DROPOUT ENSEMBLE	**0.14±0.00**	1.61±0.05	0.29±0.01	1.04±0.03	**92.8±0.1**	91.0±0.4	**91.9±0.1**	**91.9±0.1**
FSVI	0.24±0.01	1.13±0.12	0.33±0.02	0.79±0.09	86.7±0.4	91.0±0.6	87.6±0.3	87.6±0.3
FSVI ENSEMBLE	0.17±0.01	0.90±0.06	**0.24±0.01**	0.60±0.03	87.8±0.2	91.4±0.3	88.6±0.2	88.6±0.2
RADIAL-MFVI	0.50±0.11	0.68±0.11	0.51±0.11	0.61±0.11	80.6±1.1	88.7±0.7	82.2±1.0	82.2±1.0
RADIAL-MFVI ENSEMBLE	0.44±0.02	**0.59±0.05**	0.46±0.03	**0.54±0.04**	83.0±0.4	89.3±0.1	84.1±0.2	84.1±0.2
RANK-1	0.93±0.08	2.92±0.35	1.16±0.03	2.22±0.22	81.3±3.1	92.7±0.3	83.8±2.4	83.8±2.4
Rank1 Ensemble	0.41±0.05	1.58±0.11	0.54±0.04	1.15±0.07	82.7±1.8	93.5±0.2	85.3±1.4	85.3±1.4
MFVI	0.79±0.02	1.92±0.13	0.89±0.03	1.44±0.08	77.9±0.9	91.1±0.1	80.6±0.7	80.6±0.7
MFVI ENSEMBLE	0.47±0.01	1.22±0.05	0.53±0.01	0.89±0.03	79.3±0.6	91.1±0.1	82.0±0.4	82.0±0.4

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)	84.9±1.6	–	88.3±0.8	88.3±0.8	94.2±0.4	90.4±1.9	94.2±0.4	93.1±1.0
DEEP ENSEMBLE	85.1±1.0	–	90.2±0.6	90.2±0.6	95.7±0.3	89.3±1.0	95.5±0.3	93.4±0.5
MC DROPOUT	93.2±0.6	–	95.2±0.3	95.2±0.3	96.5±0.0	97.1±0.6	96.7±0.0	97.0±0.3
MC DROPOUT ENSEMBLE	**93.6±0.2**	–	**95.7±0.1**	**95.7±0.1**	**97.1±0.1**	**97.3±0.2**	**97.2±0.0**	**97.4±0.0**
FSVI	84.4±0.8	–	89.5±0.5	89.5±0.5	95.2±0.1	87.2±0.9	94.8±0.2	92.0±0.4
FSVI ENSEMBLE	84.8±0.3	–	90.4±0.1	90.4±0.1	96.0±0.0	86.5±0.3	95.6±0.0	92.5±0.1
RADIAL-MFVI	72.3±4.8	–	78.2±4.8	78.2±4.8	92.9±0.6	61.7±12.7	92.0±0.9	83.8±3.9
RADIAL-MFVI ENSEMBLE	70.6±3.2	–	76.8±3.6	76.8±3.6	94.4±0.5	60.3±8.9	93.5±0.6	85.2±2.5
RANK-1	82.3±2.7	–	87.4±1.5	87.4±1.5	94.5±0.5	83.9±3.7	94.1±0.6	90.8±1.6
RANK-1 ENSEMBLE	80.7±1.2	–	88.1±1.0	88.1±1.0	95.6±0.4	84.5±1.7	95.3±0.4	92.0±0.8
MFVI	82.2±2.5	–	87.4±1.4	87.4±1.4	94.3±0.5	86.5±2.8	93.9±0.6	91.2±1.4
MFVI ENSEMBLE	81.7±1.0	–	88.9±0.7	88.9±0.7	95.6±0.2	87.6±0.9	95.2±0.2	92.6±0.4
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	1.26±0.11	1.23±0.18	1.19±0.09	1.10±0.11	73.2±3.5	–	85.2±2.2	85.2±2.2
DEEP ENSEMBLE	0.57±0.03	0.76±0.07	0.56±0.03	0.60±0.05	70.2±1.8	–	84.5±1.2	84.5±1.2
MC DROPOUT	0.19±0.01	0.10±0.01	0.17±0.01	0.12±0.00	86.8±1.2	–	93.5±0.5	93.5±0.5
MC DROPOUT ENSEMBLE	**0.14±0.01**	**0.08±0.00**	**0.13±0.01**	**0.10±0.00**	**87.4±0.4**	–	**94.0±0.2**	**94.0±0.2**
FSVI	0.26±0.01	0.50±0.05	0.26±0.01	0.35±0.02	70.8±1.7	–	84.2±0.9	84.2±0.9
FSVI ENSEMBLE	0.19±0.00	0.44±0.01	0.20±0.00	0.28±0.00	69.6±0.9	–	84.4±0.4	84.4±0.4
RADIAL-MFVI	0.26±0.02	0.70±0.21	0.27±0.03	0.38±0.08	43.5±9.8	–	59.2±9.7	59.2±9.7
RADIAL-MFVI ENSEMBLE	0.24±0.01	0.72±0.13	0.25±0.01	0.37±0.04	33.7±7.0	–	50.8±8.0	50.8±8.0
RANK-1	0.49±0.09	0.77±0.20	0.48±0.08	0.56±0.11	65.9±5.9	–	80.2±3.6	80.2±3.6
Rank1 Ensemble	0.18±0.01	0.39±0.04	0.18±0.01	0.24±0.02	60.9±2.5	–	79.0±1.8	79.0±1.8
MFVI	0.60±0.14	0.79±0.17	0.58±0.13	0.62±0.12	66.6±5.2	–	81.1±3.2	81.1±3.2
MFVI ENSEMBLE	0.18±0.01	0.35±0.03	0.19±0.01	0.24±0.02	63.7±1.8	–	81.1±1.1	81.1±1.1

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±0.09	2.41±0.18	1.11±0.07	89.3±0.3	87.0±0.3	89.2±0.3	77.5±1.2	90.5±0.2	79.2±1.0
DEEP ENSEMBLE	0.54±0.01	1.65±0.17	0.61±0.01	90.8±0.0	88.3±0.2	90.7±0.0	81.1±0.2	**91.3±0.2**	82.0±0.2
MC DROPOUT	0.31±0.01	0.77±0.08	0.34±0.01	90.0±0.2	87.6±0.4	89.9±0.2	81.1±0.4	87.7±0.5	82.0±0.3
MC DROPOUT ENSEMBLE	**0.25±0.00**	0.58±0.04	**0.28±0.00**	**91.2±0.0**	88.3±0.2	**91.0±0.0**	**83.3±0.1**	87.7±0.4	**83.7±0.1**
FSVI	0.52±0.05	0.67±0.06	0.53±0.05	88.7±0.5	88.2±0.4	88.7±0.4	75.8±0.8	88.1±1.0	77.5±0.6
FSVI ENSEMBLE	0.39±0.02	0.42±0.02	0.39±0.02	89.2±0.3	**89.7±0.1**	89.3±0.3	79.5±0.4	88.9±0.5	80.5±0.3
RADIAL-MFVI	0.60±0.10	0.72±0.20	0.61±0.11	85.9±0.3	85.4±0.6	85.9±0.3	66.2±0.8	87.9±0.5	69.6±0.6
RADIAL-MFVI ENSEMBLE	0.38±0.00	**0.34±0.02**	0.38±0.00	87.2±0.2	87.8±0.1	87.2±0.2	69.7±0.4	89.3±0.2	72.6±0.3
RANK-1	0.88±0.08	1.95±0.27	0.95±0.09	87.0±0.8	85.1±0.5	86.9±0.7	71.3±2.4	88.4±0.3	73.8±2.0
RANK-1 ENSEMBLE	0.40±0.02	1.02±0.11	0.44±0.02	89.1±0.4	87.1±0.3	89.0±0.4	77.2±1.2	89.4±0.1	78.7±1.0
MFVI	1.09±0.13	1.69±0.26	1.12±0.14	85.9±0.5	84.5±0.7	85.8±0.5	67.1±1.9	87.6±0.9	70.3±1.5
MFVI ENSEMBLE	0.46±0.04	0.71±0.16	0.48±0.05	88.4±0.2	86.8±0.3	88.3±0.1	73.5±0.9	89.6±0.6	75.7±0.7

	AUROC (%) ↑			ECE ↓		
MAP (Deterministic)	88.6±0.6	93.2±0.2	89.5±0.5	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±0.2	0.03±0.00	0.08±0.00	0.04±0.00
MC DROPOUT ENSEMBLE	**91.9±0.1**	94.2±0.2	**92.5±0.0**	**0.02±0.00**	0.06±0.00	**0.02±0.00**
FSVI	87.4±0.4	94.0±0.4	88.5±0.4	0.08±0.01	0.08±0.00	0.08±0.01
FSVI ENSEMBLE	89.6±0.2	**94.6±0.2**	90.4±0.2	0.06±0.01	0.05±0.00	0.06±0.01
RADIAL-MFVI	83.0±0.4	92.7±0.4	84.3±0.3	0.09±0.01	0.07±0.02	0.09±0.01
RADIAL-MFVI ENSEMBLE	84.8±0.2	94.1±0.1	86.0±0.2	0.05±0.00	**0.03±0.01**	0.05±0.00
RANK-1	85.4±1.3	92.0±0.3	86.5±1.2	0.10±0.01	0.12±0.01	0.10±0.01
RANK-1 ENSEMBLE	89.0±0.8	94.0±0.2	89.8±0.7	0.05±0.00	0.07±0.00	0.05±0.00
MFVI	83.4±0.9	91.7±0.6	84.7±0.8	0.11±0.01	0.12±0.02	0.11±0.01
MFVI ENSEMBLE	86.8±0.5	94.0±0.3	87.9±0.5	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±0.15	1.48±0.26	1.09±0.15	87.6±0.8	81.5±1.2	87.1±0.8	63.6±2.5	–	77.5±1.7
DEEP ENSEMBLE	0.39±0.05	0.49±0.09	0.40±0.05	89.6±0.4	83.1±0.5	89.1±0.4	68.1±1.4	–	81.3±0.8
MC DROPOUT	0.32±0.02	0.31±0.03	0.32±0.02	89.0±0.8	87.5±1.1	88.9±0.8	72.6±2.1	–	83.5±1.5
MC DROPOUT ENSEMBLE	**0.26±0.00**	**0.24±0.01**	**0.26±0.00**	**90.9±0.1**	**89.2±0.2**	**90.8±0.1**	**76.9±0.2**	–	**86.6±0.1**
FSVI	0.40±0.03	0.57±0.03	0.41±0.02	87.8±0.7	79.8±1.1	87.1±0.7	63.3±2.1	–	77.1±1.5
FSVI ENSEMBLE	0.29±0.00	0.41±0.01	0.30±0.00	90.0±0.2	81.5±0.5	89.4±0.2	68.7±0.7	–	81.4±0.4
RADIAL-MFVI	0.37±0.01	0.76±0.12	0.40±0.02	86.7±0.4	52.0±9.0	83.9±1.1	49.1±3.5	–	66.9±2.9
RADIAL-MFVI ENSEMBLE	0.35±0.01	0.73±0.07	0.38±0.01	87.4±0.4	54.4±5.5	84.8±0.8	56.2±2.5	–	73.5±2.0
RANK-1	0.56±0.06	1.14±0.15	0.61±0.07	88.3±0.6	67.5±4.1	86.6±0.9	59.4±3.7	–	74.1±2.5
RANK-1 ENSEMBLE	0.29±0.01	0.60±0.04	0.32±0.01	89.3±0.4	69.7±2.2	87.7±0.5	66.5±2.5	–	80.0±1.6
MFVI	0.56±0.08	0.75±0.20	0.57±0.09	83.7±0.3	79.8±2.3	83.4±0.1	55.2±0.6	–	71.0±0.8
MFVI ENSEMBLE	0.35±0.00	0.37±0.01	0.36±0.00	86.2±0.3	81.6±0.7	85.8±0.3	59.9±0.3	–	75.3±0.2

	AUROC (%) ↑			ECE ↓		
MAP (Deterministic)	83.7±1.1	–	87.8±0.8	0.09±0.01	0.15±0.03	0.09±0.01
DEEP ENSEMBLE	86.3±0.5	–	90.0±0.4	0.03±0.01	0.07±0.01	0.03±0.01
MC DROPOUT	88.2±1.1	–	91.1±0.9	0.02±0.00	0.06±0.01	0.02±0.01
MC DROPOUT ENSEMBLE	**90.6±0.1**	–	**93.1±0.1**	0.02±0.00	**0.02±0.00**	0.02±0.00
FSVI	82.8±1.0	–	86.8±0.7	0.06±0.01	0.14±0.01	0.06±0.01
FSVI ENSEMBLE	86.1±0.3	–	89.7±0.2	0.03±0.00	0.08±0.01	0.03±0.00
RADIAL-MFVI	76.9±1.8	–	82.2±1.6	0.05±0.01	0.23±0.07	0.04±0.01
RADIAL-MFVI ENSEMBLE	81.3±1.4	–	86.2±1.2	0.07±0.01	0.15±0.04	0.06±0.01
RANK-1	81.6±1.8	–	85.8±1.4	0.06±0.01	0.22±0.03	0.07±0.02
RANK-1 ENSEMBLE	85.1±1.3	–	89.1±0.9	**0.02±0.00**	0.12±0.02	0.03±0.00
MFVI	79.8±0.5	–	84.3±0.5	0.07±0.02	0.12±0.03	0.07±0.02
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)	90.1±0.9	76.0±1.6	89.9±0.4	89.9±0.4	95.6±0.2	88.2±0.9	94.8±0.1	90.8±0.6
DEEP ENSEMBLE	91.8±0.3	80.7±1.6	92.0±0.1	92.0±0.1	96.6±0.0	90.7±0.8	96.0±0.1	92.9±0.6
MC DROPOUT	94.4±0.5	86.1±2.1	94.5±0.3	94.5±0.3	96.6±0.1	90.8±0.8	96.0±0.1	93.0±0.5
MC DROPOUT ENSEMBLE	**95.2±0.1**	86.9±1.0	**95.3±0.0**	**95.3±0.0**	**97.2±0.0**	91.1±0.4	**96.6±0.1**	93.5±0.3
FSVI	88.4±1.1	90.5±1.3	90.0±0.7	90.0±0.7	95.4±0.2	92.9±0.7	95.2±0.1	93.9±0.3
FSVI ENSEMBLE	88.5±0.9	93.8±0.9	90.3±0.7	90.3±0.7	96.1±0.1	94.3±0.5	95.9±0.1	**95.1±0.2**
RADIAL-MFVI	82.0±2.1	91.7±1.8	84.4±1.8	84.4±1.8	93.7±0.2	92.7±0.8	93.5±0.2	92.9±0.4
RADIAL-MFVI ENSEMBLE	80.6±1.3	**95.3±0.7**	83.7±1.2	83.7±1.2	94.4±0.1	**94.6±0.4**	94.3±0.1	94.2±0.1
RANK-1	88.7±0.8	79.5±1.8	89.0±0.5	89.0±0.5	94.1±0.5	88.3±0.7	93.5±0.4	90.4±0.4
RANK-1 ENSEMBLE	91.7±0.6	84.4±0.3	92.3±0.5	92.3±0.5	96.0±0.3	90.9±0.3	95.5±0.3	93.0±0.2
MFVI	85.7±2.1	84.3±2.5	87.0±1.3	87.0±1.3	93.6±0.4	89.9±1.2	93.3±0.3	91.3±0.6
MFVI ENSEMBLE	85.0±2.4	91.5±1.8	88.0±1.4	88.0±1.4	95.2±0.3	93.7±0.9	95.0±0.2	94.2±0.5
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	0.91±0.06	3.78±0.33	1.20±0.03	2.73±0.20	87.2±2.3	92.8±0.4	88.5±1.6	88.5±1.6
DEEP ENSEMBLE	0.48±0.01	2.69±0.29	0.72±0.04	1.86±0.21	87.9±1.2	**93.9±0.4**	89.4±0.7	89.4±0.7
MC DROPOUT	0.20±0.01	1.23±0.19	0.30±0.02	0.81±0.12	90.7±1.1	90.5±0.7	90.7±0.7	90.7±0.7
MC DROPOUT ENSEMBLE	**0.14±0.00**	0.87±0.09	**0.21±0.01**	0.57±0.06	**91.7±0.4**	90.1±0.6	**91.5±0.2**	**91.5±0.2**
FSVI	0.36±0.03	1.06±0.17	0.43±0.04	0.77±0.10	79.2±2.7	91.0±1.4	82.2±1.7	82.2±1.7
FSVI ENSEMBLE	0.25±0.02	0.55±0.09	0.28±0.01	0.42±0.04	77.6±2.3	91.8±0.8	81.2±1.7	81.2±1.7
RADIAL-MFVI	0.47±0.12	0.87±0.37	0.52±0.14	0.74±0.28	66.3±4.9	92.1±0.3	71.7±4.1	71.7±4.1
RADIAL-MFVI ENSEMBLE	0.26±0.01	**0.27±0.03**	0.26±0.01	**0.28±0.02**	61.9±3.0	92.9±0.2	68.7±2.6	68.7±2.6
RANK-1	0.88±0.11	3.14±0.47	1.10±0.13	2.29±0.35	83.8±2.1	92.1±0.2	85.5±1.6	85.5±1.6
Rank1 Ensemble	0.32±0.03	1.83±0.26	0.46±0.04	1.19±0.16	86.4±1.0	92.9±0.1	87.7±0.8	87.7±0.8
MFVI	1.01±0.16	2.64±0.52	1.15±0.18	1.97±0.36	76.0±5.3	92.2±0.6	79.9±3.6	79.9±3.6
MFVI ENSEMBLE	0.37±0.06	1.08±0.40	0.44±0.09	0.80±0.26	71.6±5.4	93.8±0.6	78.1±3.4	78.1±3.4

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)	87.7±0.8	–	90.2±0.5	90.2±0.5	94.3±0.6	94.9±0.7	94.6±0.5	95.1±0.5
DEEP ENSEMBLE	88.9±0.5	–	92.6±0.2	92.6±0.2	95.7±0.2	95.6±0.2	95.9±0.2	96.1±0.2
MC DROPOUT	92.9±0.7	–	94.4±0.4	94.4±0.4	95.7±0.4	97.5±0.6	96.0±0.4	96.9±0.5
MC DROPOUT ENSEMBLE	**94.2±0.2**	–	**95.6±0.2**	**95.6±0.2**	**96.9±0.1**	**98.1±0.1**	**97.1±0.0**	**97.7±0.1**
FSVI	87.7±0.8	–	90.9±0.4	90.9±0.4	94.1±0.4	94.8±0.7	94.4±0.4	94.7±0.5
FSVI ENSEMBLE	89.1±0.3	–	92.6±0.2	92.6±0.2	95.9±0.1	94.8±0.3	96.0±0.1	95.7±0.1
RADIAL-MFVI	72.3±4.8	–	78.2±4.8	78.2±4.8	92.9±0.6	61.7±12.7	92.0±0.9	83.8±3.9
RADIAL-MFVI ENSEMBLE	70.6±3.2	–	76.8±3.6	76.8±3.6	94.4±0.5	60.3±8.9	93.5±0.6	85.2±2.5
RANK-1	82.3±2.7	–	87.4±1.5	87.4±1.5	94.5±0.5	83.9±3.7	94.1±0.6	90.8±1.6
RANK-1 ENSEMBLE	80.7±1.2	–	88.1±1.0	88.1±1.0	95.6±0.4	84.5±1.7	95.3±0.4	92.0±0.8
MFVI	86.0±0.9	–	90.4±0.8	90.4±0.8	92.6±0.1	92.6±1.7	92.8±0.1	92.9±0.9
MFVI ENSEMBLE	88.2±0.1	–	92.3±0.1	92.3±0.1	94.1±0.1	94.9±0.3	94.3±0.1	94.8±0.1
	R-NLL AUC ↓				R-AUPRC AUC ↑			
MAP (Deterministic)	1.06±0.20	0.62±0.15	0.96±0.18	0.72±0.15	78.2±2.2	–	88.6±1.2	88.6±1.2
DEEP ENSEMBLE	0.32±0.08	0.19±0.05	0.29±0.07	0.22±0.05	78.7±1.4	–	89.9±0.6	89.9±0.6
MC DROPOUT	0.23±0.02	0.09±0.03	0.20±0.02	0.13±0.02	86.6±1.7	–	93.3±1.0	93.3±1.0
MC DROPOUT ENSEMBLE	**0.15±0.01**	**0.06±0.00**	**0.13±0.00**	**0.09±0.00**	**88.8±0.5**	–	**94.6±0.2**	**94.6±0.2**
FSVI	0.34±0.03	0.20±0.03	0.31±0.03	0.24±0.03	78.4±2.1	–	88.8±1.1	88.8±1.1
FSVI ENSEMBLE	0.19±0.00	0.15±0.01	0.18±0.00	0.15±0.00	79.2±0.9	–	89.9±0.4	89.9±0.4
RADIAL-MFVI	0.26±0.02	0.70±0.21	0.27±0.03	0.38±0.08	43.5±9.8	–	59.2±9.7	59.2±9.7
RADIAL-MFVI ENSEMBLE	0.24±0.01	0.72±0.13	0.25±0.01	0.37±0.04	33.7±7.0	–	50.8±8.0	50.8±8.0
RANK-1	0.49±0.09	0.77±0.20	0.48±0.08	0.56±0.11	65.9±5.9	–	80.2±3.6	80.2±3.6
Rank1 Ensemble	0.18±0.01	0.39±0.04	0.18±0.01	0.24±0.02	60.9±2.5	–	79.0±1.8	79.0±1.8
MFVI	0.45±0.09	0.46±0.19	0.44±0.10	0.43±0.14	72.0±1.4	–	84.7±1.1	84.7±1.1
MFVI ENSEMBLE	0.22±0.00	0.13±0.01	0.21±0.00	0.16±0.00	76.1±0.1	–	87.8±0.1	87.8±0.1

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

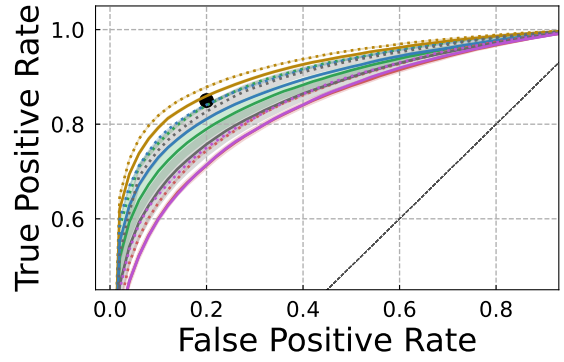
We additionally investigated to what extent the change in class distribution—in terms of the ground-truth clinical labels ranging from 0 (No DR) to 4 (Proliferative DR)—contributed to the higher performance of models in AUC, and weaker performance of models in selective prediction on the APTOS dataset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

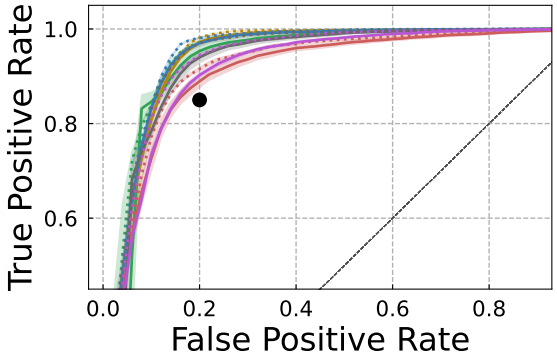
In Figure 16, we see that the ROC curves of models on the rebalanced APTOS dataset is shifted further towards the upper left as compared to the original APTOS dataset. This suggests that the class proportions of the original APTOS dataset were not the reason why models obtained stronger ROC performance 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.

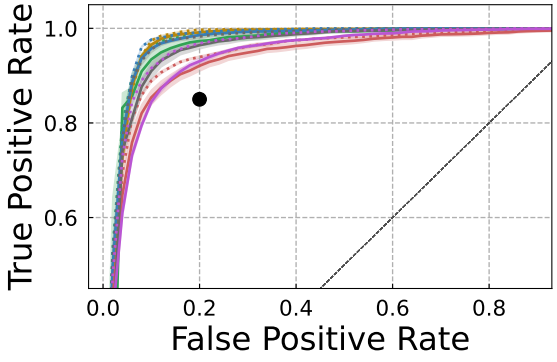
This supports the claim that factors other than simply a changed class distribution, such as meaningful shifts in equipment or patient demographics, result in both stronger predictive performance at 0% of data referred and poor quality of uncertainty estimates in the shifted setting.



(a) ROC: In-Domain

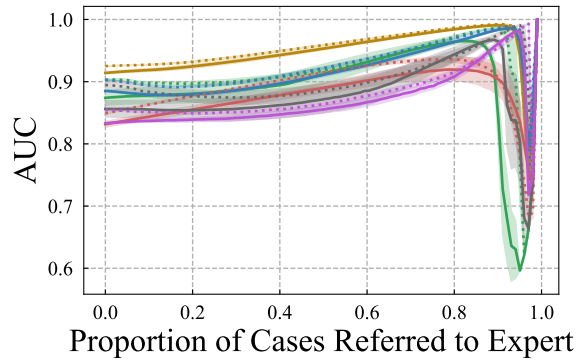


(b) ROC: Country Shift

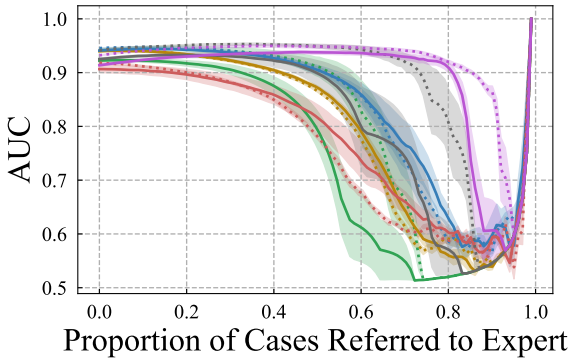


(c) ROC: Class-Balanced Country Shift

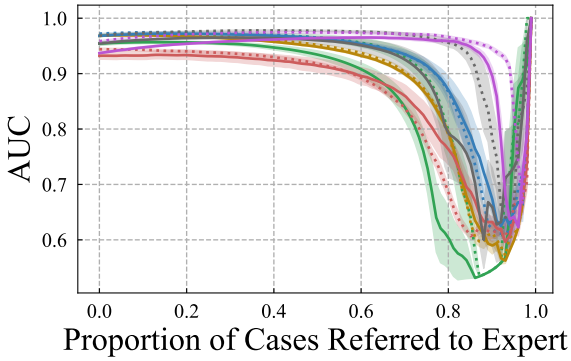
Figure 16: Class Balancing the Country Shift Dataset (ROC Curves). We consider how balancing the proportions of the ground-truth clinical class labels—ranging from 0 (No DR) to 4 (Proliferative DR)—affects performance on the *Country Shift* receiver-operating characteristic (ROC) curve. (**a**): ROC 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 $\texttt{blur\_constant} = 30$.

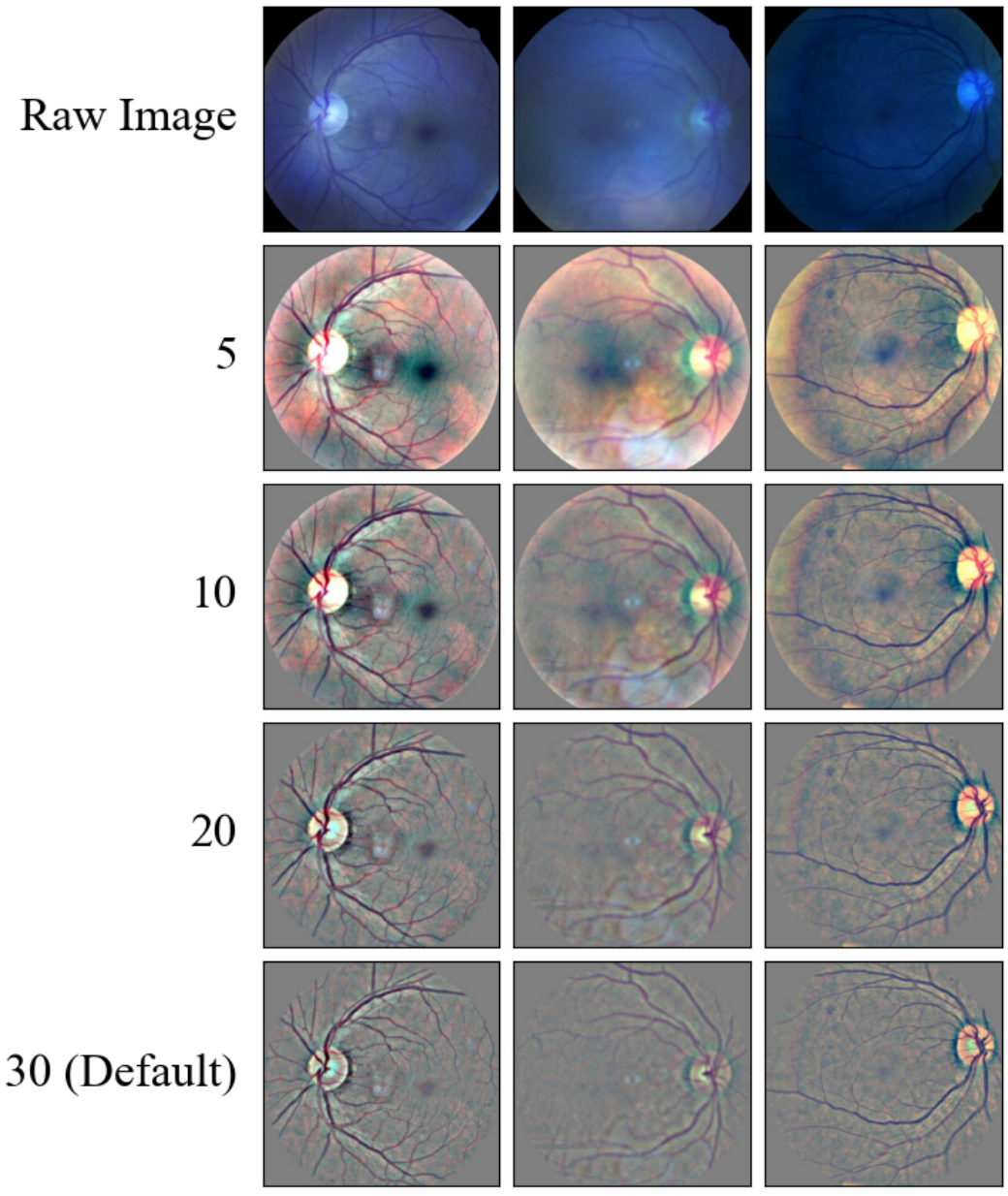
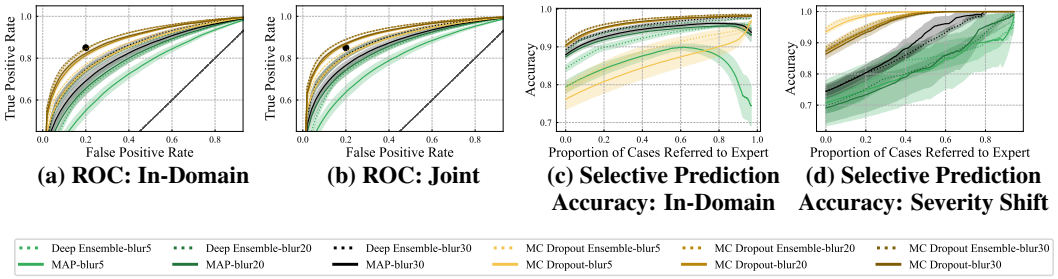
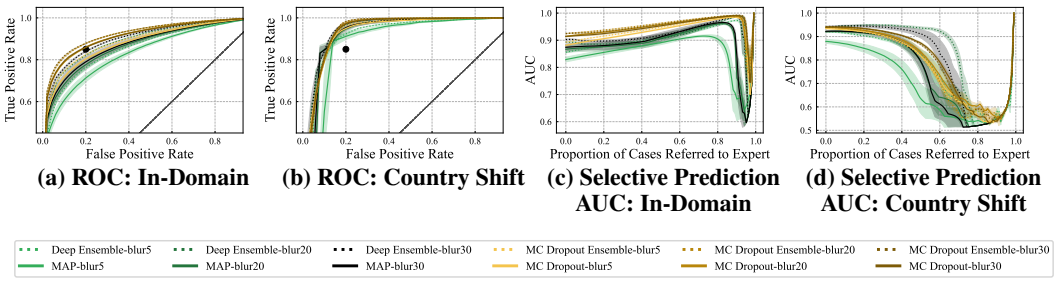


Figure 18: Preprocessing Examples. Input unprocessed EyePACS images (top row), and images processed with varying `blur_constant` (labeled on left side of grid). Higher `blur_constant` corresponds to stronger smoothing.



(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.6. 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.6. 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.6. 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.6. 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