Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks

Neil Band*† Tim G. J. Rudner*† Qixuan Feng† Angelos Filos† Zachary Nado‡
Michael W. Dusenberry‡ Ghassen Jerfel‡ Dustin Tran‡ Yarin Gal†

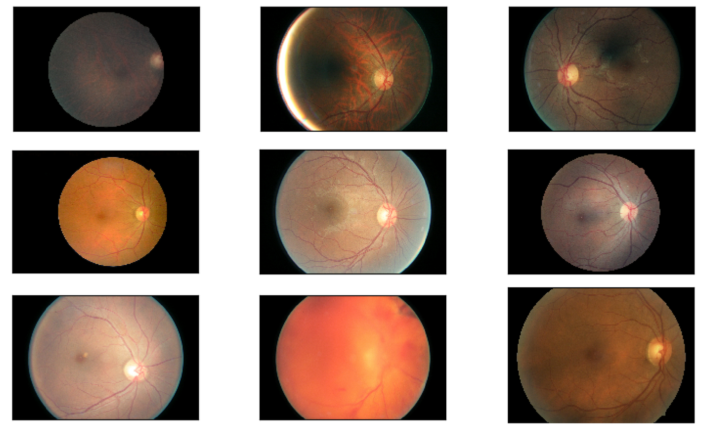
*Equal Contribution †University of Oxford ‡Google Research Correspondence to: {neil.band,tim.rudner}@cs.ox.ac.uk. @neilbband @timrudner

TL;DR

- We introduce two tasks motivated by real distributional shifts in diabetic retinopathy detection.
- We use downstream metrics to evaluate BDL methods, and:
  (i) Find that methods that capture both aleatoric and epistemic uncertainty outperform deterministic neural networks;
  (ii) Identify the failure of uncertainty quantification methods in a safety-critical automated diagnosis pipeline.

Domain: Diabetic Retinopathy Detection

- **BDL benchmark** desiderata:
  (i) Accurately reflect a **real-world setting**;
  (ii) Be usable **without** extensive **domain expertise**;
  (iii) Account for aleatoric **and** epistemic uncertainty.



Label	Diagnosis
0	No DR
1	Mild DR
2	Moderate DR
3	Severe DR
4	Proliferative DR

Figure 1 & Table 1: Left: Raw retina images from the unprocessed EyePACS dataset; **Right:** Clinical severity labels of EyePACS and APTOS retina images.

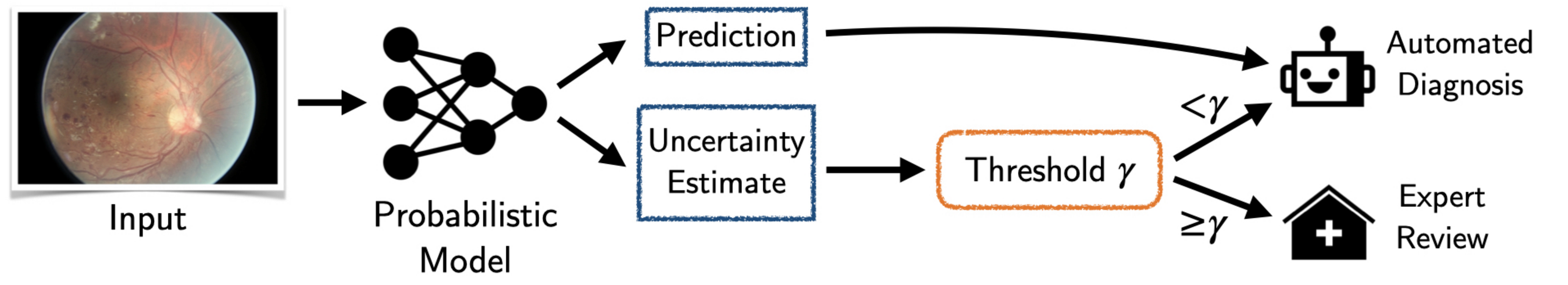


Figure 2: Automated Diagnosis Pipeline. For each input, a model provides a prediction and an uncertainty estimate; if the estimate is below γ (indicating low uncertainty) the diagnosis is processed without further review; else, it is referred to an expert.

Benchmarking Tasks and Setup

Task Construction

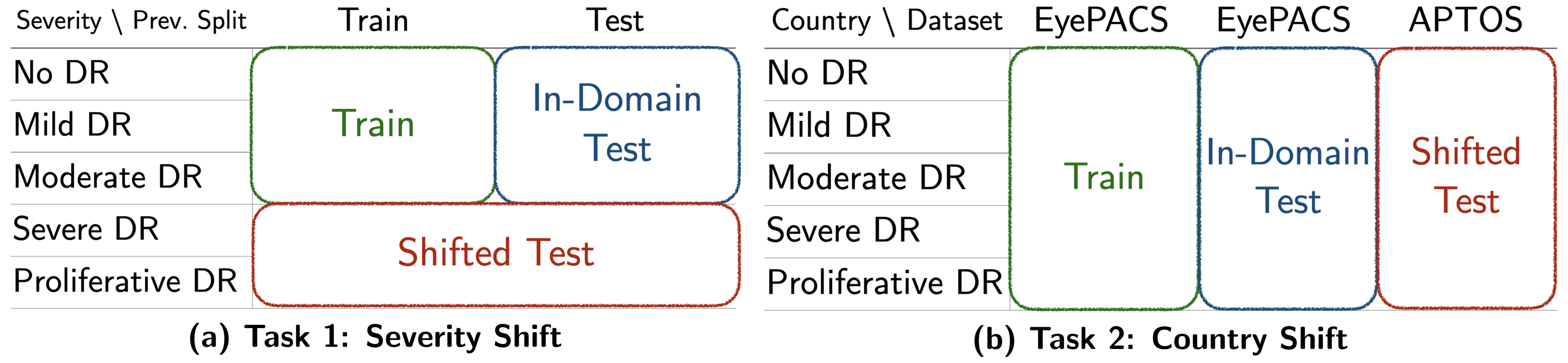


Figure 3: (a) Task 1: Severity Shift. Partitioning of the EyePACS dataset. *Goal:* evaluate reliability for rare inputs. **(b) Task 2: Country Shift.** Partitioning of the EyePACS (United States) and APTOS (India) datasets. *Goal:* evaluate reliability under different patient populations and different collection devices.

Uncertainty Quantification Methods

- **Deterministic Baselines**:
  - Maximum A Posteriori (MAP)
  - Deep Ensembles [Lakshminarayanan et al., 2017]
- **Established VI Methods for BNNs**:
  - Gaussian Mean-Field VI [Blundell et al., 2015]
  - MC Dropout [Gal and Ghahramani, 2016]
- **Improved VI Methods for BNNs**:
  - Radial Gaussian Mean-Field VI [Farquhar et al., 2020]
  - Function-Space VI [Rudner et al., 2021]
  - Rank-1 BNNs [Dusenberry et al., 2020]

Downstream Metric: Selective Prediction

- For referral rate τ, refer all images with predictive uncertainty $\geq \tau$ to an expert. Assess model on remaining images to obtain performance p. Plot p w.r.t. all possible τ.

Full paper: rebrand.ly/bdl-retinopathy

Empirical Evaluation

Severity Shift

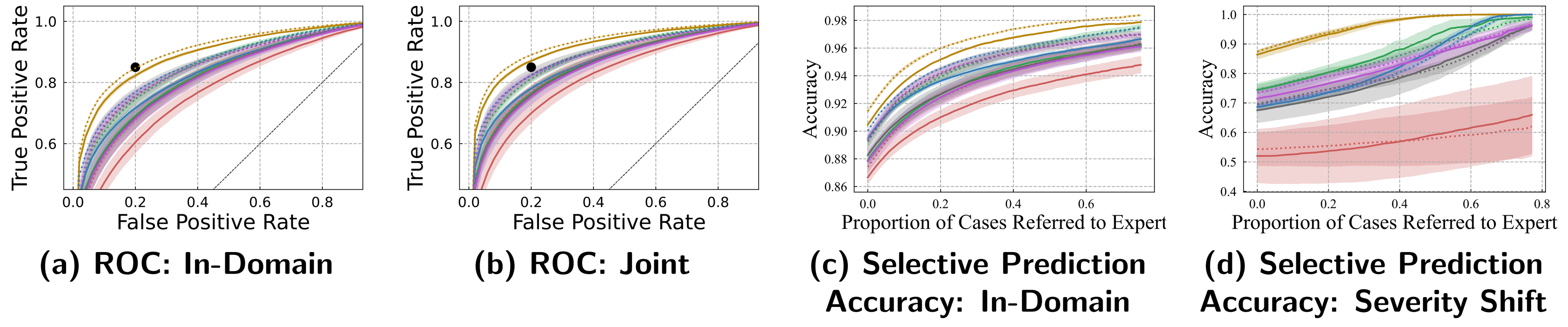


Figure 4: Left: The ROC curve for in-domain diagnosis (**a**) and for a joint dataset composed of examples from both the in-domain and *Severity Shift* evaluation sets (**b**). **Right:** Selective prediction in the in-domain (**c**) and *Severity Shift* (**d**) settings.

Country Shift

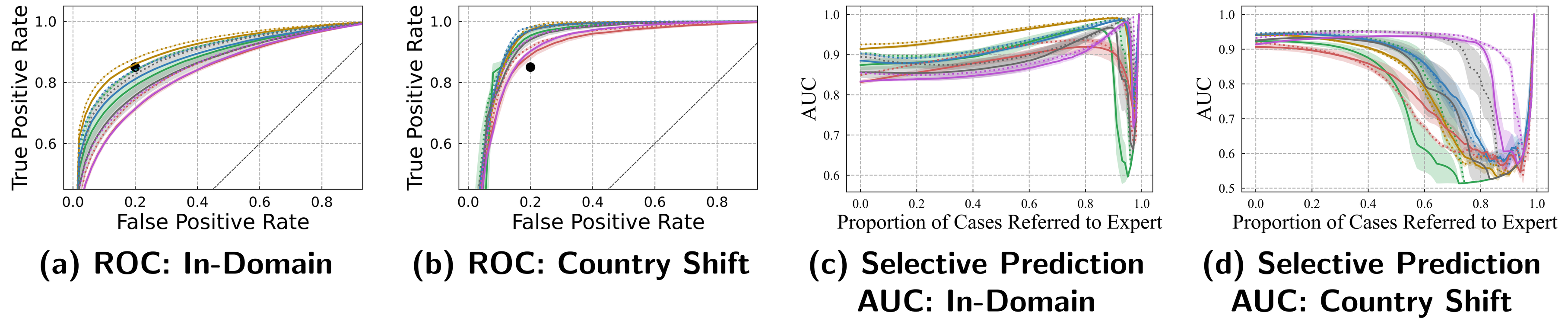


Figure 5: Left: The ROC curve for in-population diagnosis on the EyePACS test set (**a**) and for changing medical equipment and patient populations on the APTOS test set (**b**). **Right:** *selective prediction* on AUC in the EyePACS (**c**) and APTOS (**d**) settings.

Predictive Uncertainty Distributions

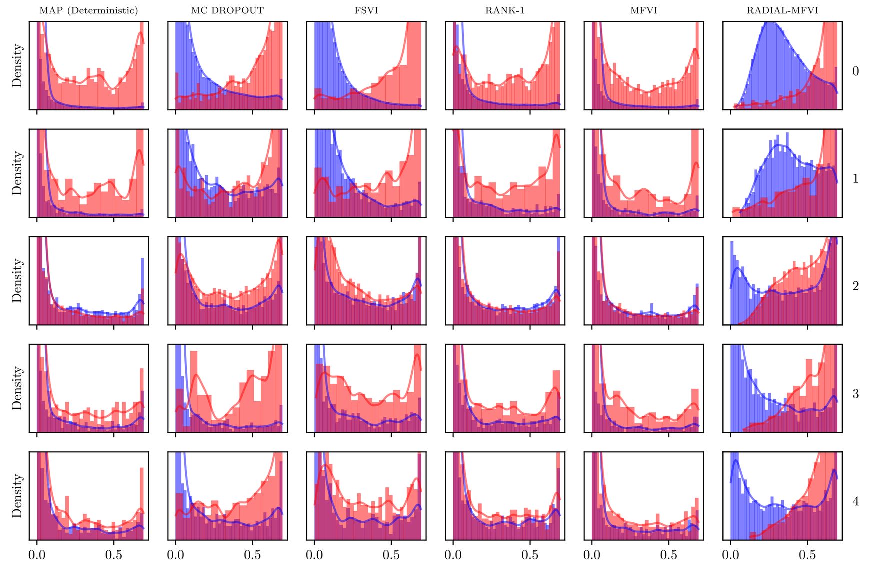


Figure 6: Severity Shift. Predictive uncertainty for each clinical severity label (rows) and method (columns), for both in-domain and shifted datasets.

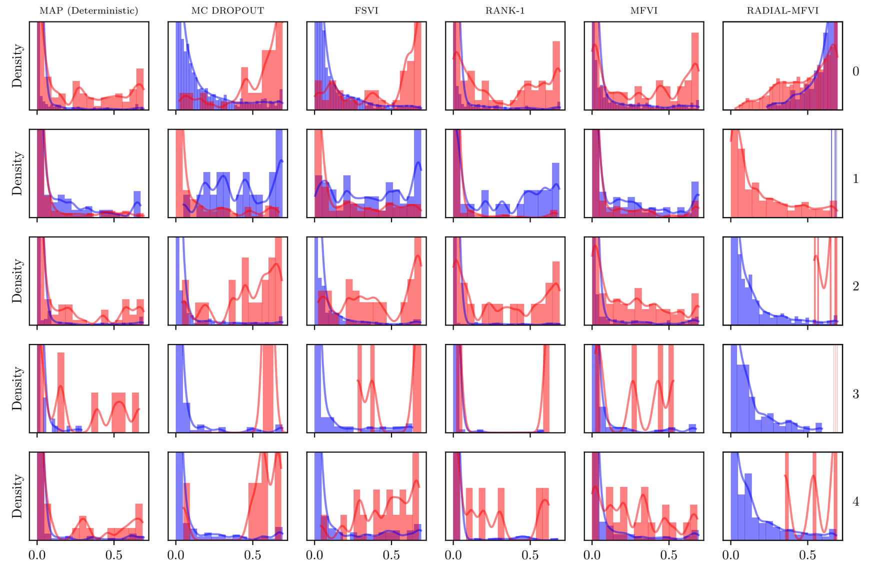


Figure 7: Country Shift. Predictive uncertainty for each clinical severity label (rows) and method (columns), for the distributionally shifted dataset (APTOS).