AVOIDING CATASTROPHIC REFERRAL FAILURES IN MEDICAL IMAGES UNDER DOMAIN SHIFT

Anuj Srivastava[1], Pradeep Shenoy[2], Devarajan Sridharan[1] *
[1] Center for Neuroscience and Computer Science and Automation, Indian Institute of Science
asrivastava.g@gmail.com, sridhar@iisc.ac.in
[2] Google Research India
shenoypradeep@google.com

ABSTRACT

Developing robust approaches for domain generalization is critical for the real world deployment of deep learning models. Here, we address a particular domain generalization challenge: selective classification for automated medical image diagnosis. In this setting, models must learn to abstain from making predictions when label confidence is low, especially when tested with samples that deviate significantly from the training set (covariate shift). Using the example of diabetic retinopathy detection we show that even state-of-the-art deep learning models, including Bayesian networks, fail during selective classification under covariate shift. Bayesian estimates of predictive uncertainty do not generalize well under covariate shift yielding catastrophic performance drops during referral. We identify the source of these failures and propose several *post hoc* referral solutions that enable reliable selective classification under covariate shift.

1 INTRODUCTION

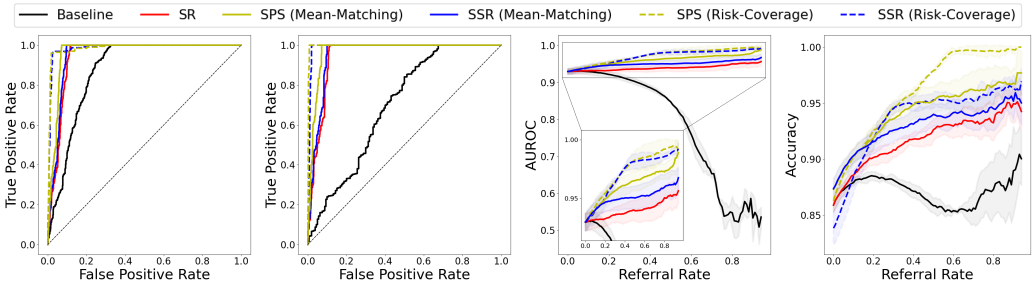
Deep learning models that are deployed in real-world situations must be able to readily generalize when they encounter novel inputs. Yet, deep discriminative models do not perform well, typically, when tested on data that is is far removed from the training distribution (domain shift). Previous domain generalization studies have explored at least one of two complementary types of domain shift: covariate shift and semantic shift. In covariate shift, the distribution of the input data changes between training and testing phases whereas the target labels remain the same (e.g. a classifier trained on photographs deployed to categorize paintings, Li et al. (2017)). By contrast, in semantic shift the distribution of input data and target labels assume additional degrees of freedom or meaning (e.g. a classifier trained to distinguish cars from boats deployed on truck and ship images).

Here, we address a particular challenge with domain generalization for automated disease diagnosis with domain-shifted medical images. Specifically, we focus challenges with selective classification under covariate shift. In this setting, the model must abstain from making predictions when it is not confident about label assignment, especially when the input (test) data deviates significantly from its training set. Such "uncertain" instances are typically referred to an expert (e.g. clinician) for further diagnosis and decision-making. In an ideal scenario, this approach would reduce the burden on the clinician while also allowing the model to maximize its accuracy by making predictions only for the most confident samples. Yet – using the example of diabetic retinopathy (DR) diagnosis with retinal fundus images – we show that even state-of-the-art deep models, including Bayesian networks, do not generalize well under covariate shift, yielding catastrophic performance failures during referral. We analyze these failures, and propose *post hoc* referral solutions that enable robust selective classification under domain shift. Related work is presented in Appendix E.

2 THE DOMAIN SHIFT CHALLENGE: DIABETIC RETINOPATHY

We demonstrate our solutions focusing on one medical imaging use case: diabetic retinopathy (DR) detection using retinal fundus images (Band et al. (2021)). In the Appendix, we apply our solutions to a different use case: pneumonia detection with chest X-ray images (Appendix F).

*Correspondence to sridhar@iisc.ac.in



(a) Referral rate: 0.5 (b) Referral rate: 0.7 (c) AUROC-Coverage (d) Accuracy-Coverage

Figure 1: Referral performance of a Bayesian network (MCD) at baseline (black curves), or following referral strategies proposed here (colored curves; SR, SSR, SPS). **(a-b)** AUROC for different referral rates. **(c-d)** AUROC or accuracy for increasing referral rates (decreasing coverage).

2.1 Bayesian predictive uncertainty under Covariate Shift

The Retina Benchmark (Band et al. (2021)) explores selective classification under two kinds of domain shift: i) semantic shift and ii) covariate shift. In the semantic shift case, also termed "Severity shift", a model trained on retinal fundus images with at most moderate DR is evaluated on images with severe or proliferative DR. In the covariate shift case, also termed "Country shift", a model trained on a DR dataset collected in United States (EyePACS (2015)) is evaluated on a DR dataset collected in India (APTOS (2019)). For both tasks the underlying task is posed as binary classification: whether a given image exhibits symptoms of diabetic retinopathy, or not. Here, we illustrate selective classification challenges in the Country shift case.

For selective classification, uncertainty with predicting a test sample's label is typically evaluated using the model's predictive entropy (equation 1).

$$\mathcal{H}\left(p_\theta\left(Y|x\right)\right) = -\sum_{y=1}^{C} p_\theta(Y=y|x)\log p_\theta(Y=y|x) \tag{1}$$

where x is a test sample, Y is a random variable denoting the model output, C is the number of output classes, and θ are the model parameters. However, conventional training objectives for deterministic models can result in overconfident, yet incorrect predictions especially when the test data are far removed from the training distribution. To address this challenge, Band et al. (2021) proposed various Bayesian deep learning models, including Monte-Carlo dropout (MCD, Gal & Ghahramani (2016)) and mean-field variational inference (MFVI, Xing et al. (2012)), which enable estimating predictive uncertainties reliably. Specifically, taking into account both aleatoric uncertainty – uncertainty due to inherent ambiguity and noise in the data – and epistemic uncertainty – uncertainty due to model constraints or biases in the training process – was shown to be important for robust referral.

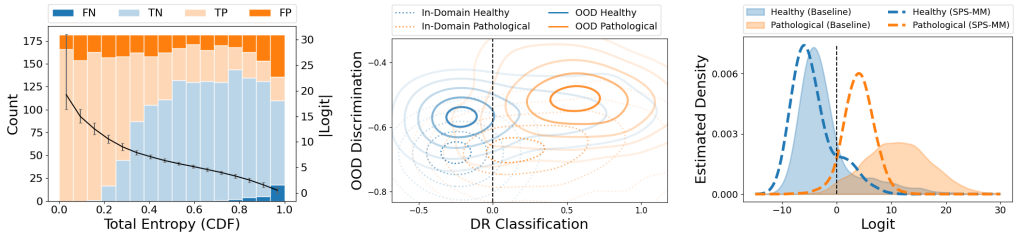
Briefly, the total uncertainty for Bayesian networks, where the network parameters θ are stochastic, is given as the sum of its aleatoric and epistemic uncertainties:

$$\underbrace{\mathcal{H}\left(\mathbb{E}\left[p_\theta\left(y|x\right)\right]\right)}_{\text{total}} = \underbrace{\mathbb{E}\left(\mathcal{H}\left[p_\theta\left(y|x\right)\right]\right)}_{\text{aleatoric}} + \underbrace{\mathcal{I}(y;\theta)}_{\text{epistemic}} \tag{2}$$

where expectation is taken over network parameters, and $\mathcal{I}(y;\theta)$ represents the mutual information between the outputs and the network parameters. Using this metric (Band et al. (2021)) showed consistent improvement over a deterministic baseline, especially for in-domain data.

2.2 Bayesian uncertainty fails to generalize under Covariate Shift

We illustrate a surprising failure of the Bayesian uncertainty metrics to generalize under covariate shift. To illustrate this, we plot the area under the ROC curve (AUROC) values for the MCD model for different referral rates – proportion of cases referred to the expert – for the Country shift case with the out-of-domain (OOD/APTOS) data (Fig. 1c). Note that the proportion of samples retained for evaluation by the model decreases along the x-axis: these represent samples with progressively lower predictive uncertainties, or greater model confidence. Therefore, AUROC should increase monotonically with the proportion of referred cases.



(a) Entropy partition (b) Feature visualization (c) Logit distributions

Figure 2: Analyzing referral failures. **(a)** Total entropy (CDF) for True and False Positives (TP/FP, orange), and True and False Negatives (TN/FN, blue), for the MCD model. Black curve: Logit magnitude for each total entropy bin. **(b)** Visualizing the in-domain (dashed contours) and out-of-domain (solid contours) data distributions for healthy (blue) and pathological (orange) classes (see text for details). **(c)** Distribution of OOD logits for the MCD model at baseline (filled histograms) and following SPS with mean-matching (dashed, open histograms).

Yet, this is not the case. Despite a high AUROC (> 0.9) for zero referral, it decreases systematically for higher referral rates (Fig. 1c, black). Plotting the ROCs for different referral rates (50%, Fig. 1a, black; and 70%, Fig. 1b,black) confirm these observations. Moreover, we observe a less pronounced, albeit similarly non-monotonic behavior with the accuracy-referral rate curve also (Fig. 1d, black) – accuracy decreases with increasing referral rates, upto around 0.6. Such failures are largely absent for in-domain (ID/EyePACS) data (Appendix D.2).

We have illustrated this pathological behavior in the AUROC-referral curve with the MCD model, among the best performing of the Bayesian deep learning models. Nonetheless, a similar pattern of catastrophic failures occurs for the other Bayesian models also, albeit to varying degrees, and for the deterministic (MAP) model also (see Appendix D.1), replicating observations in the original study (Band et al. (2021), see *their* Figure 5d).

3 Ameliorating Referral Failures under Covariate Shift

3.1 Identifying the Source of Referral Failures

We analyze the reasons for this failure by visualizing the entropy distributions for the Country shift test data (Fig. 2a). While true positives (TP) and true negatives (TN) are predicted with high confidence (low entropy), we find a significant imbalance in entropies among two error types. While false negatives are generally predicted with low confidence (high entropy), a significant fraction of false positives are predicted with high confidence (low entropy).

To understand the reason behind this trend we visualize the model outputs (logits, $f(x)$) using a two-dimensional representation (Fig. 2b): the x-axis represents the model logits themselves, whereas the y-axis represents a dimension that maximally discriminates the ID from the OOD data (detailed description in Appendix D.3). The dashed vertical line in the figure denotes the model's classification boundary estimated from the ID data, such that positive (diseased) and negative (healthy) label assignments are made to either side of this boundary. Note that there exists a monotonic relationship between the total entropy ($\mathcal{H}$) and the logit magnitude ($|L|$, Fig. 2a, black curve): for any referral rate based on an entropy threshold ($\mathcal{H}_r$), there is a unique threshold L_r on the logit magnitude, such that all the test samples with $f(x) \in [-L_r, L_r]$ are referred.

From the visualization, it is apparent that healthy class OOD distribution has a long tail to the right of the classification boundary (false positives, Fig. 2b, solid blue contours to the right of the classification boundary). Thus, some of these samples would be predicted, albeit incorrectly, with even greater confidence than the most confident true negatives! The long tail in the distribution of the healthy class logits (Fig. 2c, filled blue histogram) highlights this failure case even more starkly. As a result, a significant proportion of the most confident predictions are also incorrect. Consequently, these incorrectly labeled samples are retained with high confidence at higher referral rates, leading to the catastrophic drop in AUROC.

3.2 OUTPUT TRANSFORMATIONS AND REFERRAL STRATEGIES

We propose three novel strategies for referral, including transformations to model outputs. All approaches are applied *post hoc* and do not require retraining the model.

3.2.1 SPLIT REFERRAL

In this approach, we rank OOD test samples based on their transformed logit values (total entropies), separately for each predicted class. Then, referrals are made in the order of the least to the most confident predictions, proportionately for each class label. Mathematically,

$$t_{\text{SR}}(z) = \begin{cases} \Phi_{\mathcal{N}}(z) - 1, & \text{if } z < 0 \\ \Phi_{\mathcal{P}}(z), & \text{otherwise} \end{cases}$$

where $z = f(x)$ are the model outputs (logits), $\Phi_{\mathcal{N}}$ and $\Phi_{\mathcal{P}}$ are the empirical CDFs of the sets of negative $f(\mathcal{N})$ and positive $f(\mathcal{P})$ predictions respectively, where $\mathcal{N} = \{x \in \mathcal{D}_{\text{test}} : z < 0\}$, and $\mathcal{P} = \{x \in \mathcal{D}_{\text{test}} : z \geq 0\}$. Thus, for a given referral rate τ, $\tau|\mathcal{N}|$ and $\tau|\mathcal{P}|$ samples are referred from $\mathcal{N}$ and $\mathcal{P}$ respectively, and the proportion of predicted labels is maintained at every rate of referral. This approach does not require validation data or parameter tuning.

3.2.2 SHIFTED SPLIT REFERRAL

Shifted Split Referral (SSR) is essentially identical with the previous approach, except that model outputs are transformed by "shifting" before split referral.

$$t_{\text{SSR}}(z; b) = \begin{cases} \Phi_{\mathcal{N}}(z + b) - 1, & \text{if } z + b < 0 \\ \Phi_{\mathcal{P}}(z + b), & \text{otherwise} \end{cases}$$

where b is a tunable parameter, $\mathcal{N} = \{x \in \mathcal{D}_{\text{test}} : z + b < 0\}$, and $\mathcal{P} = \{x \in \mathcal{D}_{\text{test}} : z + b \geq 0\}$.

3.2.3 SPLIT PLATT SCALING

Platt scaling (Platt (2000)) is a simple algebraic transformation on model outputs to obtain calibrated probabilities. Inspired by this, we propose Split Platt Scaling (SPS), which transforms model outputs using a shift and a scaling parameter.

$$t_{\text{SPS}}(z; a_{\mathcal{N}}, a_{\mathcal{P}}, b) = \begin{cases} a_{\mathcal{N}}(z + b), & \text{if } z + b < 0 \\ a_{\mathcal{P}}(z + b), & \text{otherwise} \end{cases}$$

where $a_{\mathcal{N}}$, $a_{\mathcal{P}}$ and b are tunable parameters (tuning procedure described in the next section). We scale the logits of positive and negative predictions by different factors, which allows balancing their relative uncertainties. Because the relative uncertainties depend on the ratio of $a_{\mathcal{N}}$ and $a_{\mathcal{P}}$ and not their magnitudes, we consider $a_r = \frac{a_{\mathcal{P}}}{a_{\mathcal{N}}}$ and b as the two tunable parameters.

3.3 TUNING THE TRANSFORMATION PARAMETERS

For SSR and SPS we optimize the transformation parameters with an iterative shrinking grid search (Appendix C). We consider the following optimization objectives:

- **Risk-coverage**: Maximizing the area under a performance metric versus referral rate curve, using held out validation data from the OOD set; this is equivalent to minimizing the area under a risk-coverage curve (El-Yaniv & Wiener (2010)). The performance metric was either AUROC or Accuracy, depending on whichever metric was being evaluated for the OOD test set.
- **Mean-matching**: Minimizing the squared difference between the mean logits of ID and OOD, separately for each prediction label.

$$\min_t (\mathbb{E}[t \circ f(\mathcal{N}_{\text{OOD}})] - \mathbb{E}[f(\mathcal{N}_{\text{ID}})])^2 + (\mathbb{E}[t \circ f(\mathcal{P}_{\text{OOD}})] - \mathbb{E}[f(\mathcal{P}_{\text{ID}})])^2 + \beta(\text{Var}[t \circ f(\mathcal{N}_{\text{OOD}})] + \text{Var}[t \circ f(\mathcal{P}_{\text{OOD}})])$$

where $t \circ f$ represents the transformed logits and the last term represents a penalty for transformations that significantly increase the range of the transformed outputs ($\beta = 0.01$).

$$\mathcal{N}_{\text{ID}} = \{x \in \mathcal{D}_{\text{ID}} : f(x) < 0\};\ \mathcal{N}_{\text{OOD}} = \{x \in \mathcal{D}_{\text{OOD}} : t \circ f(x) < 0\}$$

$$\mathcal{P}_{\text{ID}} = \{x \in \mathcal{D}_{\text{ID}} : f(x) >= 0\};\ \mathcal{P}_{\text{OOD}} = \{x \in \mathcal{D}_{\text{OOD}} : t \circ f(x) >= 0\}$$

Note that risk-coverage objective requires an annotated validation set, which may not be available for OOD data in practical settings. On the other hand, the mean-matching objective does not require such an annotated OOD dataset. Moreover, for SPS with the mean-matching objective, we add the constraint that equalizes the range of the transformed logits across the two classes, $\max_{x \in \mathcal{N}_{\text{OOD}}} |t \circ f(x)| = \max_{x \in \mathcal{P}_{\text{OOD}}} |t \circ f(x)|$, to enable a balance in the relative predictive uncertainties across the two classes. We also explore the effect of varying the size of the OOD validation set for parameter tuning in Appendix G.

4 RESULTS

To illustrate the advantages with our referral strategies we examine the ROC curves for the Country shift DR test set (OOD/APTOS) with the MCD model at referral rates of 50% and 70% (Fig. 1a, Fig. 1b). We observe systematically poor AUROCs at each of these referral rates with the baseline MCD model (Fig. 1a & Fig. 1b, black curve); each of these is significantly alleviated by our methods (Figs. 1b and 1b, colored curves).

Next, we examine the AUROC-referral curve for progressively increasing values of referral rates. Although AUROCs are virtually identical with the baseline model (Fig. 1c, black curve) at lower referral rates ($<$0.2), with our strategies, the MCD model is rescued from catastrophic failure at all higher referral rates (Fig. 1c, colored curves). A qualitatively similar improvement is also apparent when we plot accuracy-referral curves (Fig. 1d). With our referral strategies, the long-tail in the logit distribution for the healthy class is also alleviated (Fig. 2c, dashed blue histogram).

We quantify overall performance improvements using the area under the performance-coverage curve (AUPCC); as mentioned earlier, this metric is inversely related to the area under the risk-coverage curve, measured conventionally (El-Yaniv & Wiener (2010)). Across the spectrum of models, incorporating our referral strategies yields systematic improvements in performance (Table 1, and Appendix D.4). Overall, shifted split referral (SSR), with one tunable parameter, yields the highest gains over baseline AUPCC as measured with *AUROC* (up to 13.3% for the MAP model, 19.9% for MCD and 4.6% for MFVI). Split Platt scaling (SPS) yields gains comparable to SSR in nearly all cases (up to 12.5% for the MAP model, 19.7% for MCD and 3.8% for MFVI), despite having one additional tunable parameter. Despite having no tunable parameters, vanilla split referral (SR) achieves significant improvements, albeit with more modest gains (up to 11.6% for the conventional model, 17.7% for MCD and 3.5% for MFVI) (Table 1). A qualitatively similar pattern of gains is also observed for AUPCC measured with *accuracy* (Table 1). We also evaluate the robustness of our strategies by swapping the ID and OOD dataset in the Country shift case (Appendix H).

In summary, we propose targeted, lightweight strategies to rescue catastrophic referral failures for DR images under covariate shift. Our solutions are relevant for developing robust and reliable automated diagnosis and referral pipelines in medical imaging applications. Future work will involve incorporating self-supervised representations tailored for medical imaging data (Azizi et al. (2022)), as well as extending these solutions to multiclass settings (Shui et al. (2022)).

Referral method	**MAP**		**MCD**		**MFVI**	
	AUROC	**Accuracy**	**AUROC**	**Accuracy**	**AUROC**	**Accuracy**
None (baseline)	0.833	0.904	0.773	0.894	0.899	0.901
Split Referral (SR)	0.949	0.931	0.950	0.933	0.934	0.913
SSR (Risk-Coverage)	0.966	0.941	**0.972**	0.946	0.945	0.919
SSR (Mean-Matching)	0.956	0.939	0.958	0.942	0.943	0.919
SPS (Risk-Coverage)	0.961	0.957	0.970	**0.965**	0.937	0.934
SPS (Mean-Matching)	0.962	0.944	0.967	0.951	0.920	0.911

Table 1: Area under the performance-coverage curve (AUPCC) for 3 different models: deterministic/maximum a posteriori (MAP), Monte-Carlo dropout (MCD) and mean field variation inference (MFVI), after applying each of our referral strategies. Blue: best performance for the respective model. **Bold**: best performance across all models.

REFERENCES

APTOS. Aptos 2019 blindness detection dataset, 2019.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=jyd4Lyjr2iB.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

EyePACS. Diabetic retinopathy detection dataset, 2015.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir Abdi. Stop overcomplicating selective classification: Use max-logit. *arXiv preprint arXiv:2206.09034*, 2022.

Adam Fisch, Tommi S. Jaakkola, and Regina Barzilay. Calibrated selective classification. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=zFhNBs8GaV.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 2179–2187. PMLR, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 33:19365–19376, 2020.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0092867418301545.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/0c4b1eeb45c90b52bfb9d07943d855ab-Paper.pdf.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Qixuan Feng, Angelos Filos, Marton Havasi, Rodolphe Jenatton, et al. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

Changjian Shui, Gezheng Xu, Qi CHEN, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YsRH6uVcx2l.

Dustin Tran, Jeremiah Zhe Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda E Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, E. Kelly Buchanan, Kevin Patrick Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=6x0gB9gOHFg.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. *arXiv preprint arXiv:1212.2512*, 2012.

APPENDICES: AVOIDING CATASTROPHIC REFERRAL FAILURES IN DOMAIN-SHIFTED MEDICAL IMAGES

A MODEL ARCHITECTURE AND TRAINING

We use the model architectures and trained checkpoints from the Retina Benchmark to reproduce their results for the deterministic, Monte Carlo dropout (MCD) and mean-field variation inference (MFVI) models. The deterministic model uses the maximum a posteriori estimate for classification and is also referred to as the MAP model. Each model has a ResNet-50 (He et al. (2016)) backbone with a fully connected layer on top. Expectations for the MCD and MFVI models in equation (2) (see section 2.1) are estimated using Monte Carlo sampling over the parameter distributions; the number of samples per test input is set at 5. When running the Benchmark code we noticed (and corrected) a logical error that affects entropy computation due to sub-optimal floating point truncation; our results include this correction.

More details, including hyperparameter settings, are available at the Retina Benchmark repository.

B DIABETIC RETINOPATHY DATASETS AND PREPROCESSING

In the diabetic retinopathy Country shift benchmark, the in-domain data is from the EyePACS dataset (EyePACS (2015)), and the out-of-domain data is from the APTOS dataset (APTOS (2019)). The EyePACS dataset is part of the Kaggle DR Detection Challenge (2015), and consists of 35,126 training, 10,906 validation, and 42,670 test RGB images of human retinas, collected in USA. Each image is annotated with a severity level from 0 to 4. The annotations are interpreted as follows: 0 - no DR, 1 - mild DR, 2 - moderate DR, 3 - severe DR, and 4 - proliferative DR. APTOS consists of retina images with the same annotation scheme, but was collected in India using different medical equipment. 80% of the images (2,929) are used as the test set and the other 20% (733) as a secondary validation set (used for selecting training checkpoints); the validation set was not used in our study.

The severity labels $\{0, 1\}$ are considered to be healthy (negative class), and the labels $\{2, 3, 4\}$ are considered retinopathic (positive class). EyePACS has $\sim$20% positive samples, whereas APTOS has $\sim$40% positive samples. The images are preprocessed following the Kaggle competition winner (EyePACS (2015)): resizing such that retinas have a radius of 300 pixels, local Gaussian blur smoothing, and clipping to 90% size to avoid imaging artifacts at the retina boundary.

C OPTIMIZING TRANSFORMATION PARAMETERS

We optimize shifting/scaling parameters with an iterative shrinking grid search. These are first initialized such that the transformation is the identity function: $b^0 = 0$ (and $a_r^0 = 1$ if using SPS).

In iteration i, we search the set of parameters B^i (for SPS we search $A^i \times B^i$), given by:

$$B^i = \{b^i + j s_b^i : -w_b \leq j \leq w_b\}$$
$$A^i = \{a_r^i \exp(j s_a^i) : -w_a \leq j \leq w_a\}$$

where w_* (search width / number of steps) and s_*^0 (step size) are hyperparameters. The terms are updated as:

$$b^{i+1} = \text{argmin}_{b \in B^i} \mathcal{L}(b)$$

for SPS:

$$(a_r^{i+1}, b^{i+1}) = \text{argmin}_{(a_r, b) \in A^i \times B^i} \mathcal{L}(a_r, b)$$

where $\mathcal{L}(\cdot)$ is the optimization objective (area under risk-coverage curve or mean-matching) as a function of the parameters. The step size is halved:

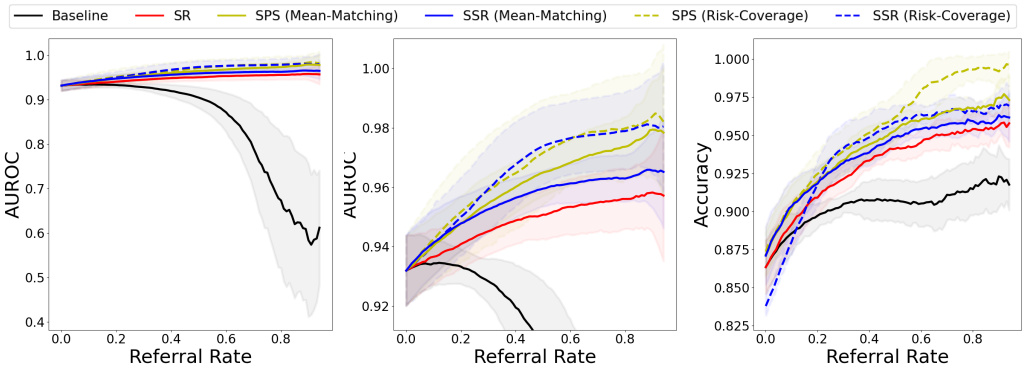
$$s_*^{i+1} = \frac{1}{2} s_*^i$$

We set the search width ($w_a = w_b = 2$) and initial step size ($s_a^0 = 3.0, s_b^0 = 32.0$) such that the algorithm searches a very large space, but the iteratively shrinking search space ensures a local optimum is precisely achieved. The total number of iterations is fixed at 7. Increasing this enables higher precision of the local optima.

D DR DATASET: ADDITIONAL RESULTS

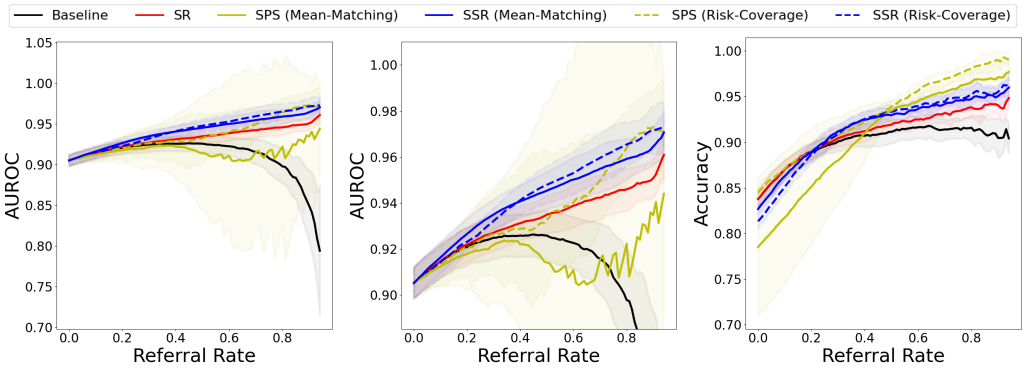
D.1 CATASTROPHIC REFERRAL FAILURES IN OUT-OF-DOMAIN DATA

We show here that catastrophic referral failures, similar to those reported in the main text, also occur with another Bayesian network (MFVI, Fig. 4) as well as the deterministic model (MAP, Fig. 3)) on out-of-domain (OOD/APTOS) data.



(a) AUROC-Coverage (b) AUROC-Coverage (zoomed in) (c) Accuracy-Coverage

Figure 3: Referral failures in the deterministic (MAP) model. Other conventions are as in Fig. 1c.



(a) AUROC-Coverage (b) AUROC-Coverage (zoomed in) (c) Accuracy-Coverage

Figure 4: Referral failures in the Bayesian (MFVI) model. Other conventions are as in Fig. 1c.

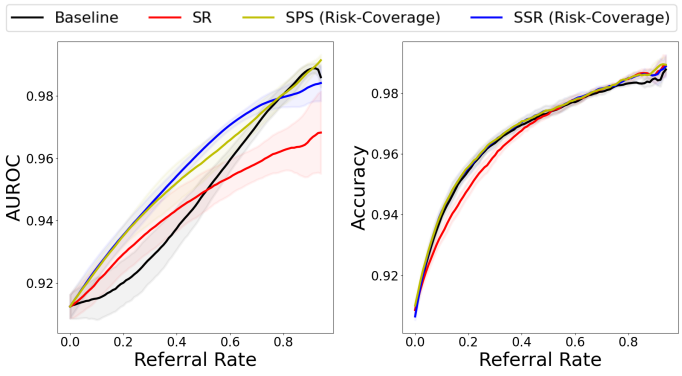
D.2 CATASTROPHIC FAILURES ARE ABSENT IN-DOMAIN DATA

By contrast, catastrophic referral failures do not occur for in-domain (ID/EyePACS) data (Fig. 5).

D.3 VISUALIZING THE MODEL OUTPUTS

In Figure 2b, in the main text, the contours are obtained using Gaussian kernel density estimation over the projections of the penultimate ResNet-50 layer features (2048 dimensional), on the following two components: the x-axis represents the projection on the last layer classifier weights for DR detection, and the y-axis represents a dimension that maximally discriminates ID from OOD data that is orthogonal to the last-layer classifier projection. The y-axis projection is calculated as follows:

Let $h(x)$ be the penultimate layer features for a given image x. Fitting a logistic regression on $h(x)$ to discriminate ID and OOD on the combined test set with a cross-entropy minimization objective

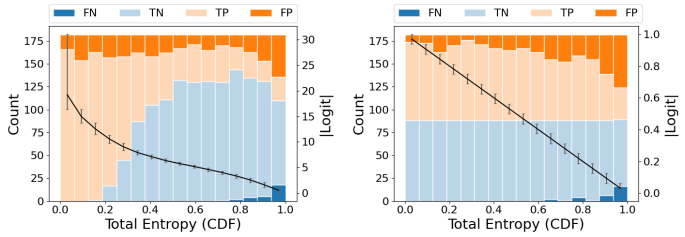


(a) AUROC-Coverage (b) Accuracy-Coverage

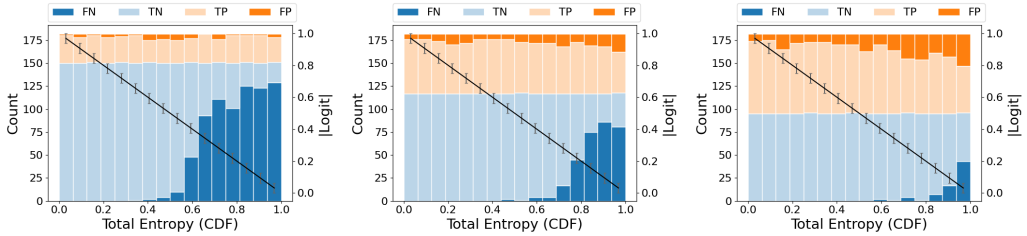
Figure 5: Referral failures are absent for in-domain data (MCD model). Other conventions are as in Fig. 1c.

gives us a weight vector w_D and bias b_D ($\hat{y}(x) = (w_D \cdot h(x) + b_D)$). Now let the retinopathy classifier be $(w_C \cdot h(x) + b_C)$, and define $w'_D = w_D - w_D \cdot \frac{w_C}{||w_C||_2}$.

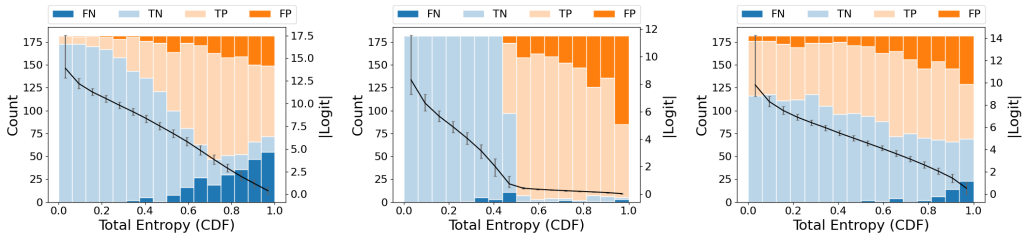
In Figure 2b the y-axis shows the (unnormalized) value of $(w'_D \cdot h(x))$. Thus OOD data assumes higher values on the y-axis, on average, than ID data. Moreover, OOD points that are lower on the y-axis (closer to the ID distribution) are reasonably well classified, whereas the tail of false positives in the OOD data lies higher up on y-axis and, thus, those points are less similar to the ID data.



(a) Baseline (b) SR



(c) SSR - Risk Cov. (AUROC) (d) SSR - Risk Cov. (Accuracy) (e) SSR - Mean matching

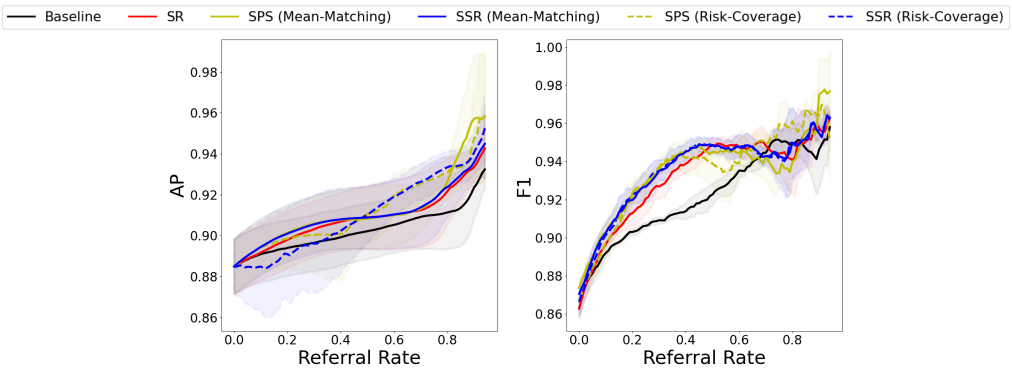


(f) SPS - Risk Cov. (AUROC) (g) SPS - Risk Cov. (Accuracy) (h) SPS - Mean matching

Figure 6: Entropy partitions across outcomes (TP, FP, TN, FN) following the different referral strategies proposed here. Other conventions are the same as in Fig. 2a.

D.4 Average Precision and F1 scores

We show that our approaches marginally benefit, but certainly do not impair, precision and recall on OOD (APTOS) data. For this, we plot the area-under-the-precision-recall curve (AUPRC or average precision, Fig. 7a), as well as the F1 scores (Fig. 7b), for different referral rates (MCD model). We observe a marginal increase in F1 scores with our approaches and no detriment in average precision.



(a) AUPRC-Coverage (b) F1-Coverage

Figure 7: Average precision and F1 scores for the MCD model

E Related Work

We build upon the Retina Benchmark, a part of the Uncertainty Baselines repository (Nado et al. (2021)); the benchmark comprising various Bayesian deep networks is described in Section 2. PLEX (Tran et al. (2022)) extends the benchmark by applying state-of-the-art pretrained vision transformers with its last layers modified for reliable uncertainty quantification. Yet, PLEX suffers from many of the same shortcomings as the Retina Benchmark.

Multiple studies on selective classification have proposed training separate networks to estimate predictive uncertainty for abstention (e.g. SelectiveNet Geifman & El-Yaniv (2019), Deep Gamblers Liu et al. (2019) & Self-Adaptive Training Huang et al. (2020)). Yet, Feng et al. (2022) show that simple softmax response (Geifman & El-Yaniv (2017)) based abstention outperforms these approaches. Similarly, Gangrade et al. (2021) formulate selective classification as an optimization problem of maximizing coverage with an upper bound on model error. Yet, none of these studies, to our knowledge, have addressed the challenge of selective classification in the domain generalization setting.

Also relevant to our objective are studies on model calibration. Yet, while well calibrated models can give reliable uncertainty estimates, calibration on ID data does not ensure that uncertainty estimates generalize for OOD data. (Fisch et al. (2022)) show state-of-the-art performance on a related problem: calibrated selective classification under domain shift. Yet, this approach requires retraining a base model to improve selective calibration error, unlike our approaches which can be applied *post hoc* on pretrained models.

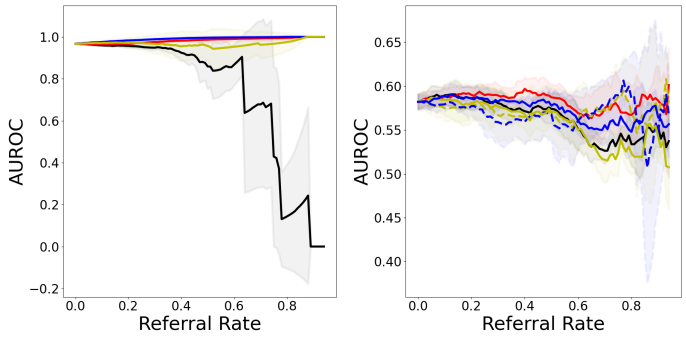
F Covariate shift in the Pneumonia dataset

We show that the pattern of catastrophic referral failures happens in a different medical domain: detecting pneumonia from chest X-ray images.

We trained the MAP and MCD models on the pneumonia dataset from (Kermany et al. (2018)), which consists of frontal chest X-ray scans from children, having 5,232 train and 624 test images. These are labeled as one of healthy, bacterial pneumonia, or viral pneumonia. We consider only the binary classification task with the negative class being healthy and postive class being pathology (either bacterial or viral pneumonia). We evaluated these models on the ChestX-ray14 dataset (Wang et al. (2017)), consisting of a total of 112k images from 33k unique patients, annotated with

14 predetermined pathologies (including pneumonia), which are extracted with NLP from textual reports.

In this case, we observe catastrophic referral failures, even for in-domain data (Fig. 8a, black curves). Our methods alleviate this catastrophic referral failure (Fig. 8a, colored curves)

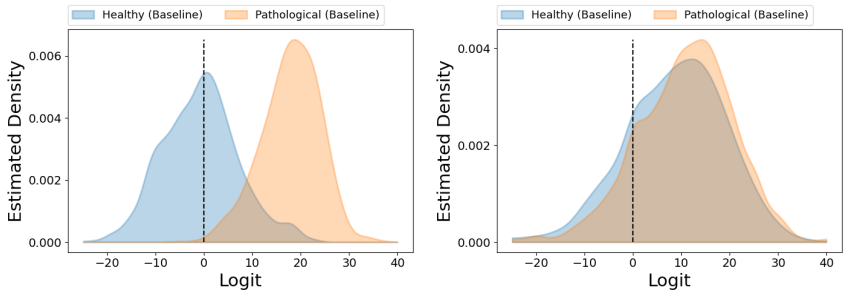


(a) AUROC-Coverage, ID (b) AUROC-Coverage, OOD

Figure 8: MC-Dropout model for Pneumonia detection. Other conventions are as in Fig. 1c.

Interestingly, we observed that with the OOD data, AUROCs were not much greater than 0.5, even for zero referral rates (Fig. 8b, black curve). Moreover, none of our referral methods could rescue this failure (Fig. 8b, colored curves).

Examining the OOD logits revealed that the model did not generalize at all to the OOD data (Fig. 9b), even at zero referral rates. This failure case suggests that the *post hoc* solutions we have proposed here are relevant for correcting referral failures arising from an imbalance in entropy when the model generalizes reasonably well. Alternative domain adaptation approaches – for example, retraining the model in an unsupervised or semi-supervised setting – may be necessary for rescuing more extreme cases of OOD generalization failures that occur even prior to referral.

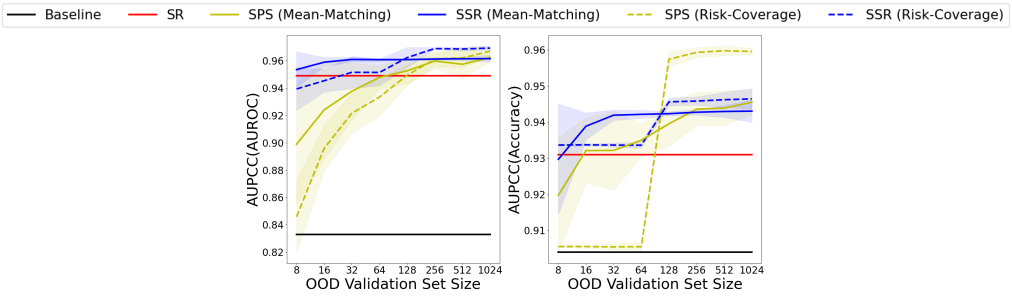


(a) Logit distribution, ID (b) Logit distribution, OOD

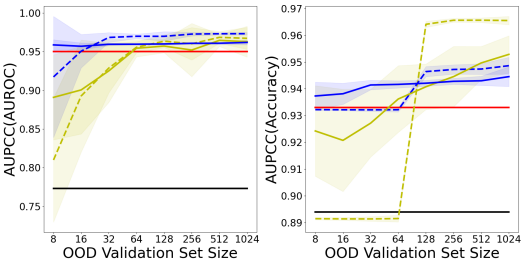
Figure 9: Logit distributions for Pneumonia detection (MC-Dropout model).

G Effect of OOD validation set size used for tuning transformation parameters

We studied the effect of varying the OOD validation set size used for tuning the transformation parameters for the SSR and SPS approaches (Figure 10). With smaller validation size the test AUPCCs were lower, especially for SPS(Risk-Coverage). Though SPS(Risk-Coverage) exhibited high scores overall, it also needed larger validation set sizes for optimization, compared to the other methods. On the other hand, SSR(Mean-Matching) was more robust to validation set size variation – only a marginal change in AUPCCs occurred across diverse validation set sizes (8 to 1024).



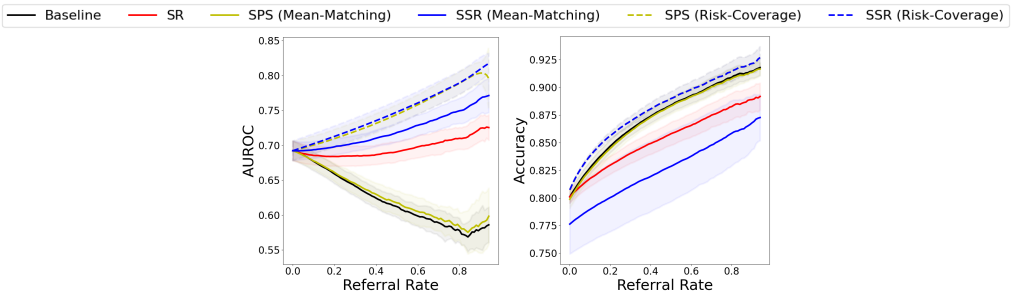
(a) MAP model



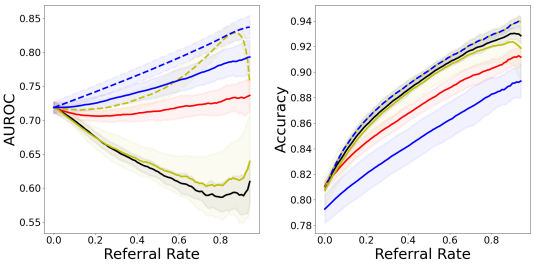
(b) MCD model

Figure 10: Variation of Area Under the Performance Coverage Curves (AUPCCs) with varying OOD validation set sizes. Other conventions are as in Fig. 1c.

H Robustness check by swapping ID and OOD Datasets: Country Shift case



(a) MAP AUROC (b) MAP Accuracy



(c) MCD AUROC (d) MCD Accuracy

Figure 11: Swapping ID and OOD datasets in the Country shift case. Models were trained on the APTOS dataset, and tested on the EyePACS dataset. Other conventions are as in Fig. 1c.

To test the robustness of our methods, we also evaluated them by swapping the ID and OOD datasets in the Country Shift DR setting. In this case, the ID dataset is APTOS and the OOD dataset is EyePACS. We follow the same procedure as Country shift – train the model (MAP or MCD) on

APTOS, perform model output transformations using validation sets from both datasets, and report the results on the EyePACS test set.

In this case also, we observe dramatic referral failures at baseline: the AUROC decreases systemtaically with increasing referral rates (Fig 11a,c, black). Nearly all of our methods, except SPS(Mean-matching), were able to rescue this failure (Fig 11a,c, all curved except solid green). On the other hand accuracies show a marginal decrease with SR and SSR(Mean-matching) (Fig 11b, d); notably, both of these strategies lack outlier label exposure.

We suspect that this failure occurred due to the smaller size of APTOS (2,929 train images) compared to EyePACS (35,126 train images), resulting in potential overfitting and poor generalization. This hypothesis is supported by the fact that zero referral rate AUROC was considerably higher for the original case ($>$0.9 for EyePACS/ID and APTOS/OOD, Fig. 1c) as compared to the swapped case ($\sim$0.7 for APTOS/ID and EyePACS/OOD, Fig. 11c).