Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis

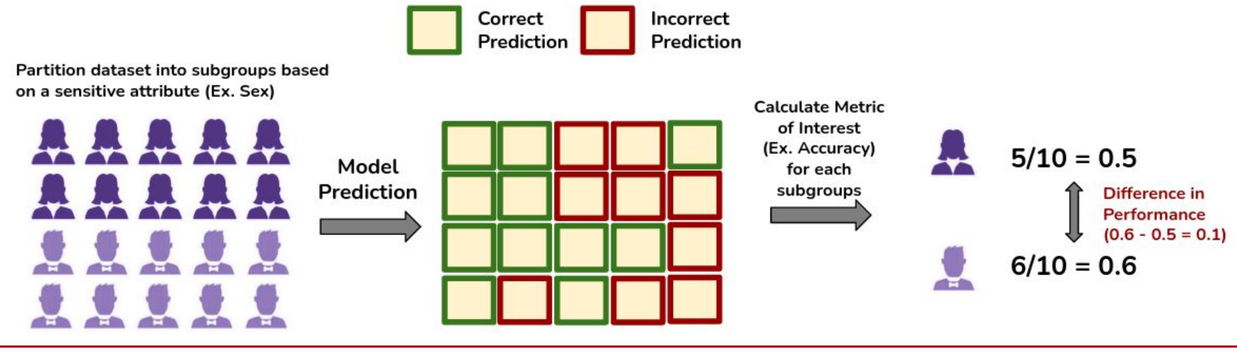
Raghav Mehta [1,2], Changjian Shui [1,2], and Tal Arbel [1,2]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada [2] MILA Quebec AI Institute, Montreal, Canada

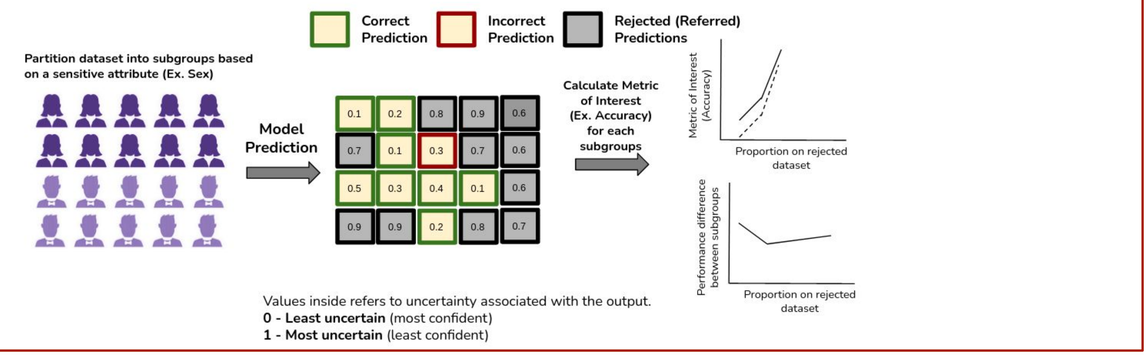
(1) Introduction

- **Deployment** of machine learning models for medical image analysis **requires**: (i) **fairness/robustness** across different subpopulation and (ii) **uncertainty quantification** to express model confidence
- **Goal**: **Analyze** effect of popular **fairness models** to overcome biases in terms of **both absolute performance and uncertainty quantification**

(2) Background: Fairness evaluation



(3) Proposed Evaluation: Fairness and Uncertainty

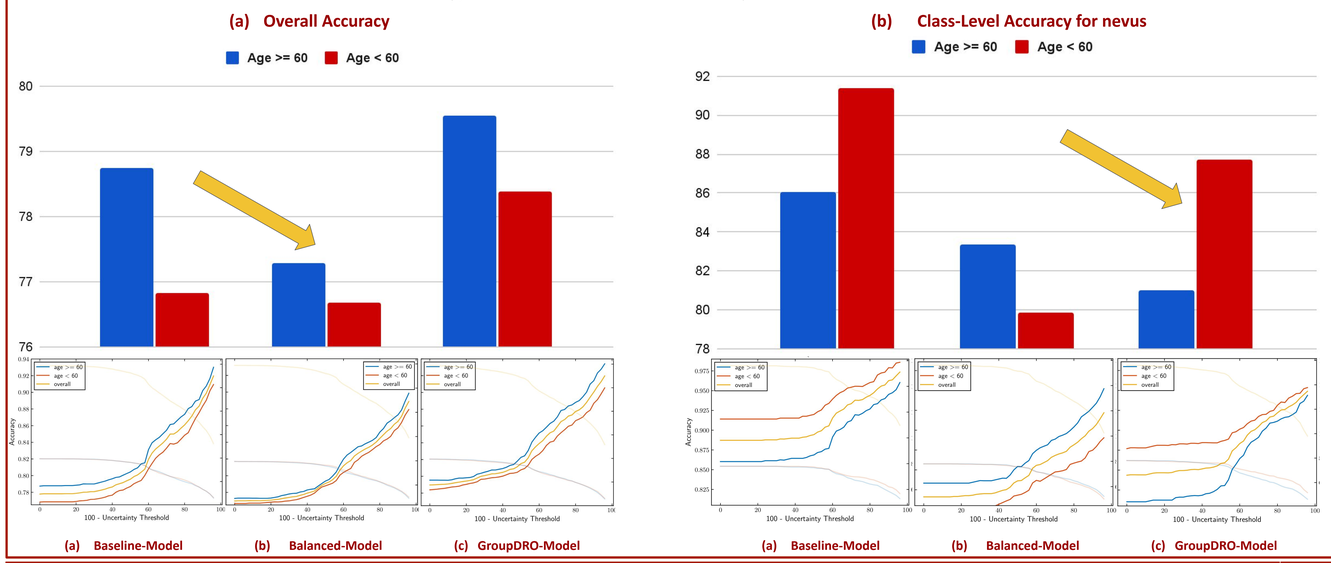


(4) Experiments and Results

- **Evaluated bias mitigation (fairness) models**: (a) **Baseline-Model**: trained on a dataset without consideration of any sub-group information.
  (b) **Balanced-Model**: trained on a balanced dataset across sub-groups [1].
  (c) **GroupDRO-Model**: trained with GroupDRO loss [2] to re-weigh the loss for each subgroup.
- **Uncertainty quantification method**: **Ensemble Dropout** [3]
- **Uncertainty Measure**: **Entropy** (for classification and segmentation); **total variance** (for regression)

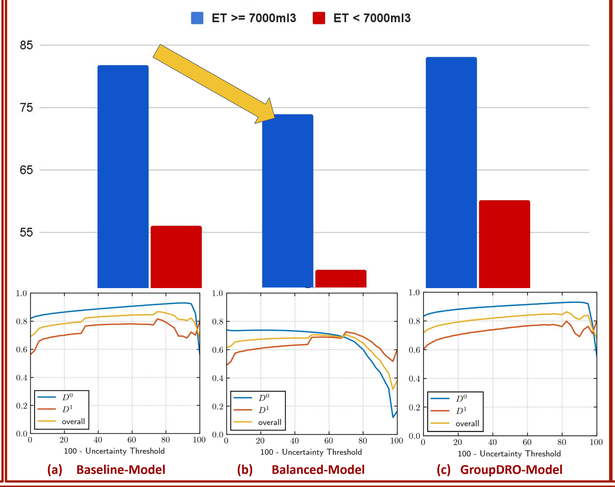
Skin Lesion Classification

- **Dataset**: ISIC 2019 [4] **Sensitive Attribute.**: Age (>= 60 and <60)
- **Evaluation Metric**: overall accuracy, and class-level accuracy for nevus.



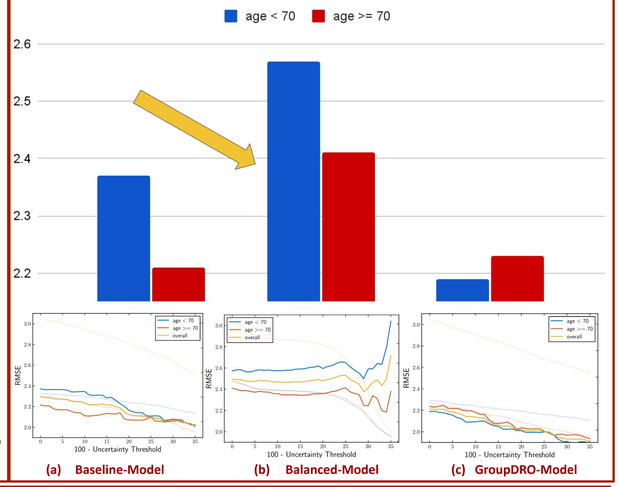
Brain Tumour Segmentation

- **Dataset**: BraTS 2019 [5]
- **Sensitive Attribute**: Enhancing Tumour Size (>= 7000ml^3 and <7000ml^3)
- **Evaluation Metric**: Dice for Enhancing Tumour



Clinical Score Regression

- **Dataset**: ADNI [6]
- **Sensitive Attribute**: Age (<70 and >=70)
- **Evaluation Metric**: Root Mean Squared Error (RMSE) for Mini Mental State Examination



(5) Conclusion

- Experiments indicate that **popular bias mitigation methods do not work well** for all medical image analysis tasks.
- Mitigating **fairness** in terms of performance **can come at the cost of poor uncertainty estimates.**
- **Future work** should develop methods that tries to **mitigate both absolute performance and uncertainty differences** across subgroups.

(6) Reference

1) B. Idrissi et al.,"Simple data balancing achieves competitive worst-group-accuracy.", In Conference on Causal Learning and Reasoning (CLeaR) 2022.
2) S. Sagawa et al., "Distributionally Robust Neural Networks." In International Conference on Learning Representations (ICLR) 2019.
3) L . Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection." In Conference on Uncertainty in Artificial Intelligence (UAI) 2018.
4) N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)." arXiv preprint a, 2019.
5) S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge.", arXiv preprint arXiv:1811.02629, 2018.
6) C. R. Jack Jr et al., "The alzheimer's disease neuroimaging initiative (adni): Mri methods.", Journal of Magnetic Resonance Imaging, 2008

Acknowledgment: This investigation was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program.