Failure Detection in Medical Image Classification: a Reality Check and Benchmarking Testbed

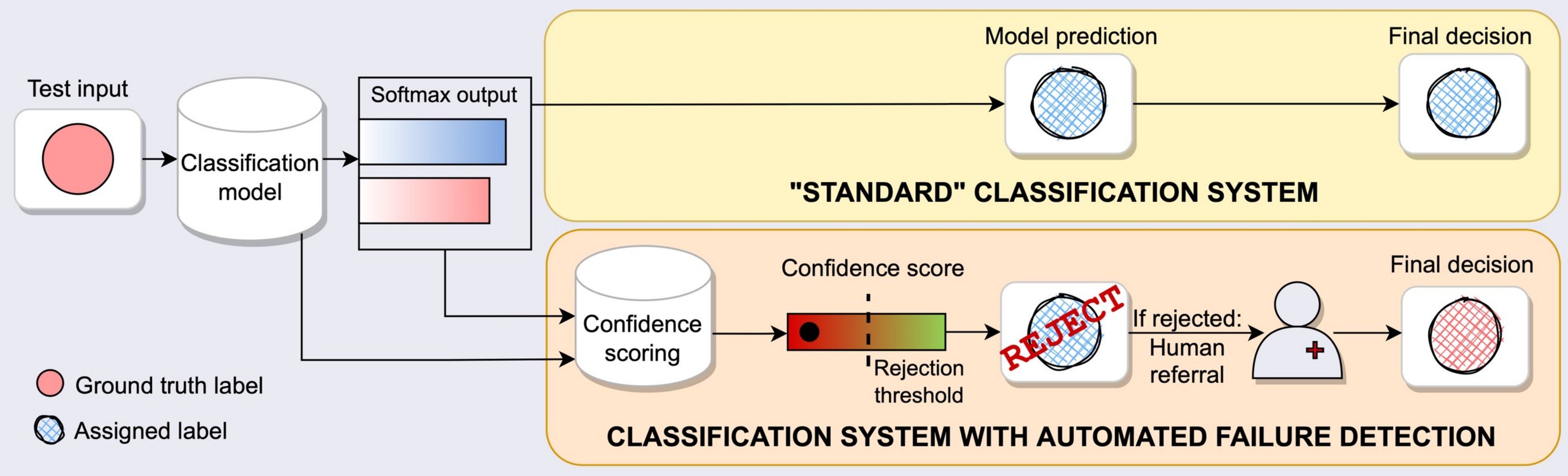
Mélanie Bernhardt – Fabio De Sousa Ribeiro – Ben Glocker
mb121@ic.ac.uk

Paper Code

How useful are commonly used uncertainty estimates for in-domain failure detection?

- We evaluate 9 widely used confidence scores on 6 different medical datasets for in-domain failure detection.
- None of these confidence scores consistently outperform a simple softmax baseline for misclassification detection.
- Results show that improved OOD performance does not necessarily translate to improved in-domain failure detection.
- In-domain failure detection needs to be studied separately.

Motivation



Automated failure detection is a crucial component of safe AI deployment in health-related scenario.

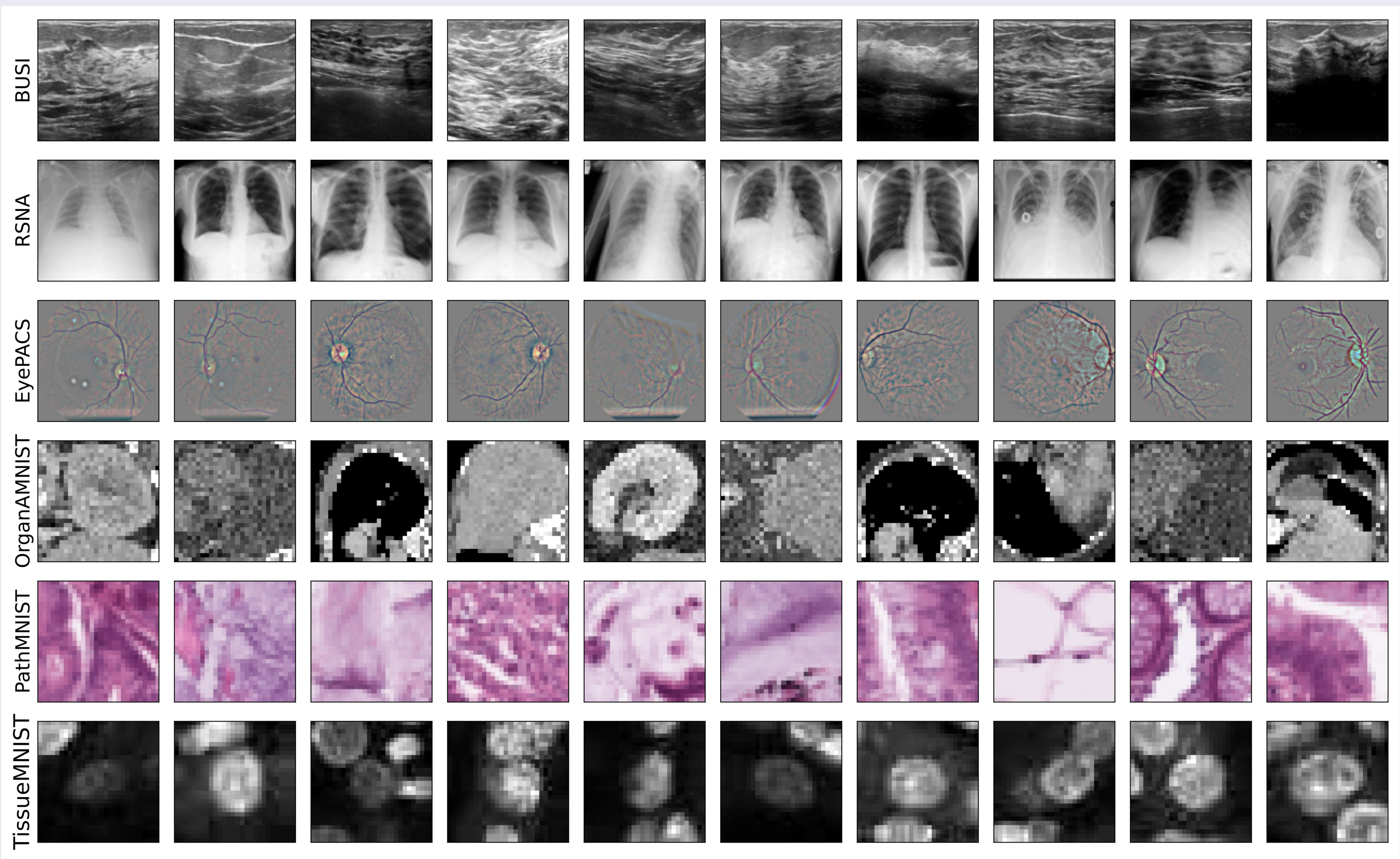
The community has proposed many confidence scoring schemes, however most of them only evaluate their performance for out-of-distribution or model calibration.

However, little is known about how good common uncertainty estimates are for misclassification detection across tasks.

→ There is a need for a comprehensive study focusing on misclassification detection comparing various types of confidences scores across different datasets.

Methods and datasets

Created a **testbed comprising 6 different datasets** and imaging modalities with resolution ranging from 28x28 to 512x512.



Compared **various commonly used confidence scores**:

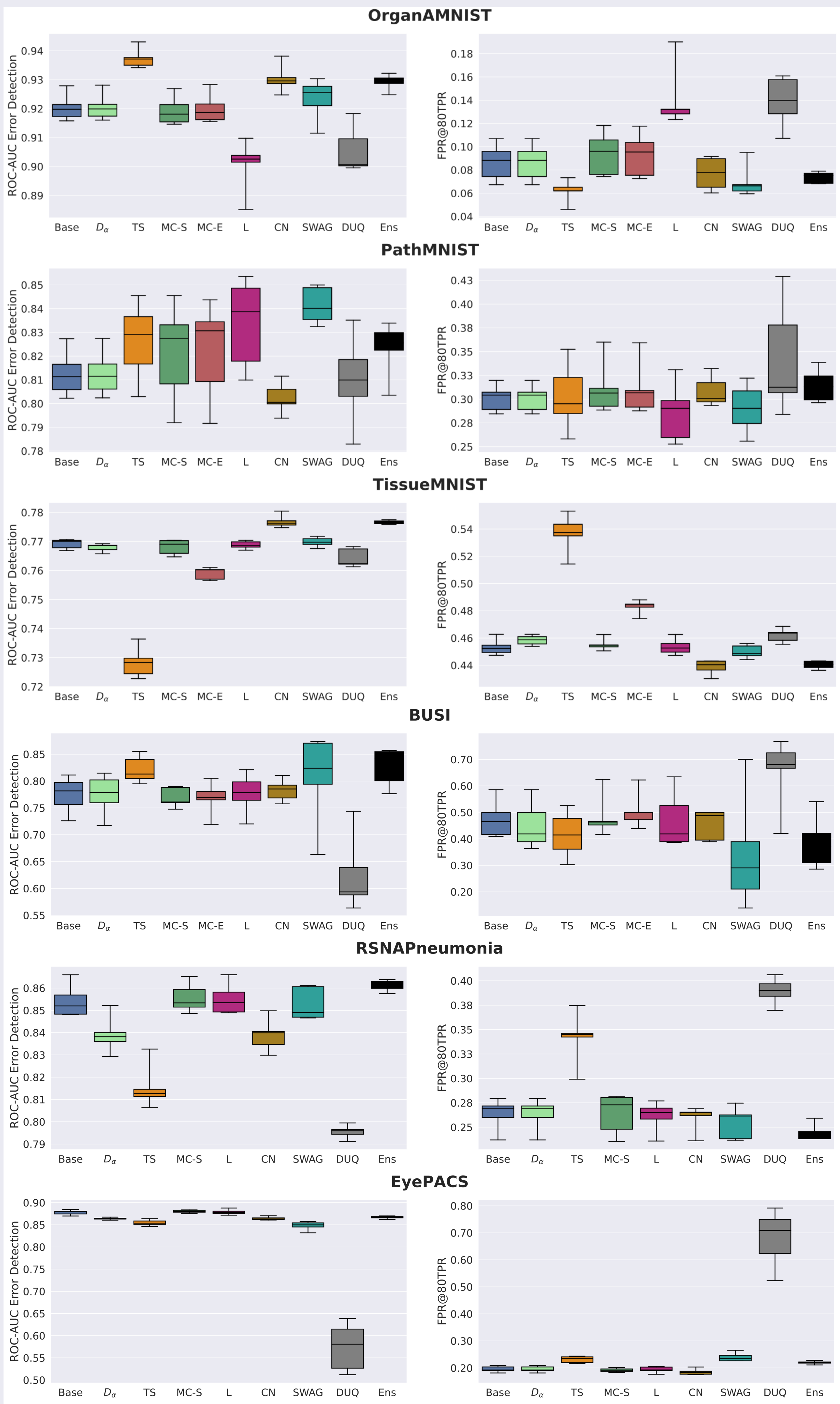
- Softmax-based confidence for predicted class[1]
- Bayesian uncertainty estimates (MC-dropout[2], Laplace[3], SWAG[4])
- Non-Bayesian uncertainty estimates (DUQ[5], ensembles)
- Embeddings-based confidence (TrustScore[6], ConfidNet[7])

Metrics:

- ROC-AUC for failure detection (where positive class = correctly classified)
- FPR@80: percentage of errors missed at 20% false alarms.

[1] Hendrycks et al. A baseline for detecting misclassified and out-of-distribution examples in neural networks.
[2] Gal et al. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In ICML, pp. 1050–1059. PMLR, 2016
[3] Daxberger et al. Laplace redux-effortless bayesian deep learning. Advances in Neural Information Processing Systems, 34, 2021.
[4] Izmailov et al. Averaging weights leads to wider optima and better generalization.
[5] Van Amersfoort et al. Uncertainty estimation using a single deep deterministic neural network. In ICML, pp. 9690–9700. PMLR, 2020.
[6] Jiang et al. To trust or not to trust a classifier.
[7] Corbière et al. Addressing failure prediction by learning model confidence. arXiv preprint arXiv:1910.04851, 2019.

Results



Benchmark results on ResNet models.

Conclusion

- None of the benchmarked confidence scores are able to **consistently outperform a simple softmax baseline** for misclassification detection.
- Results show that **improved OOD detection do not necessarily imply better misclassification detection**, calling for more research in this field and for more systematic evaluations of uncertainty estimates for the task of misclassification detection.
- Our **testbed is publicly available** to encourage more comprehensive and standardised evaluation of future confidence scores for failure detection.

This work has been published in *Failure Detection in Medical Image Classification: A Reality Check and Benchmarking Testbed. Bernhardt et al. Transactions on Machine Learning Research (2022)*