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

Melanie Berhardt*, Fabio de Sousa Ribeiro*, Ben Glocker*

* Imperial College London, UK

Published in Transaction on Machine Learning Research (10/2022)

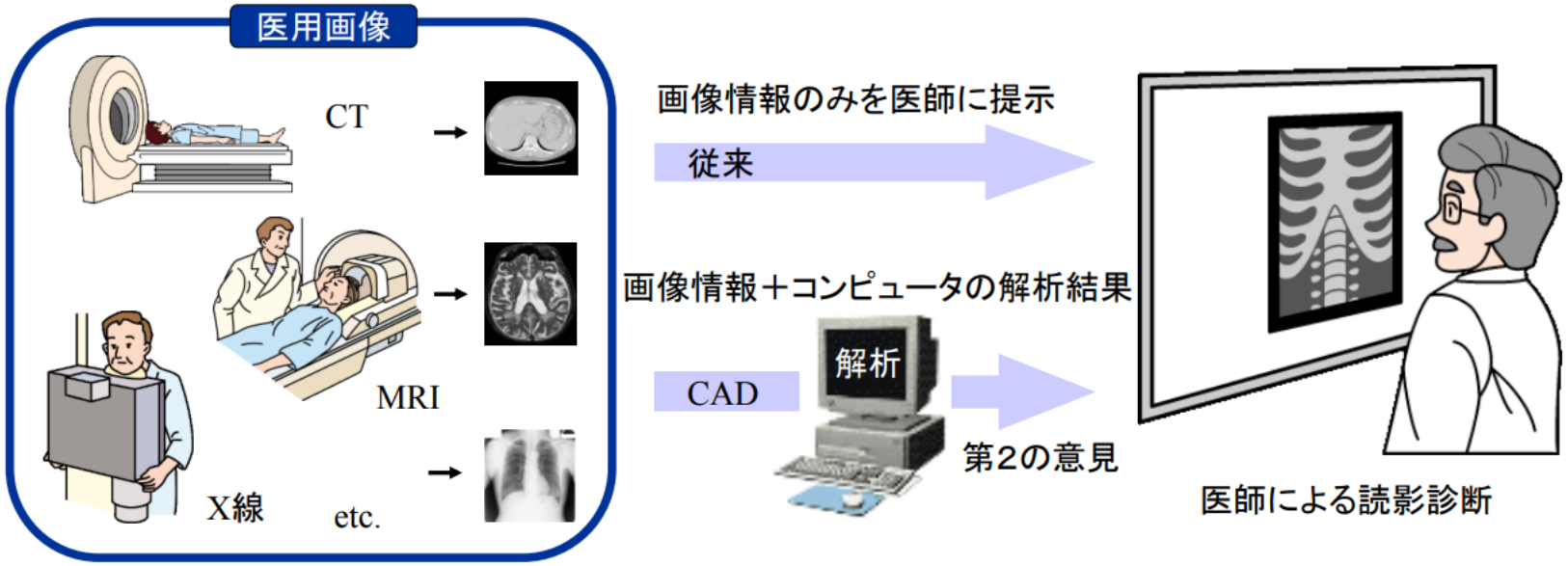
https://openreview.net/forum?id=VBHuLfnOMf

https://github.com/melanibe/failure_detection_benchmark

Index

Background

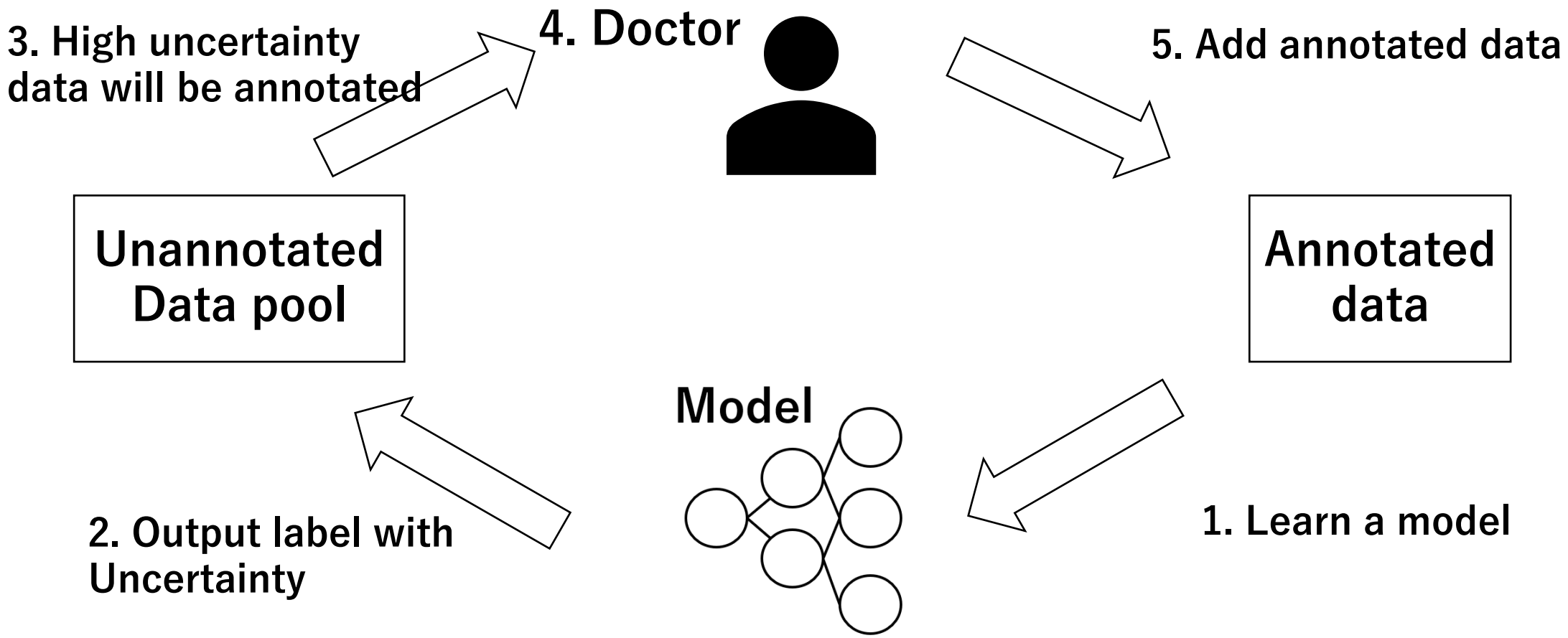
- Recently, many deep learning model has been developed and used in the various field, i.e., logistics, financing, retailing…
- Especially, in the medical field, computer aided diagnosis is well-known techonology.



http://www.clg.niigata-u.ac.jp/~lee/jyugyou/info_system/medsys006_print.pdf

Background

- Active learning based on prediction uncertainty



Background

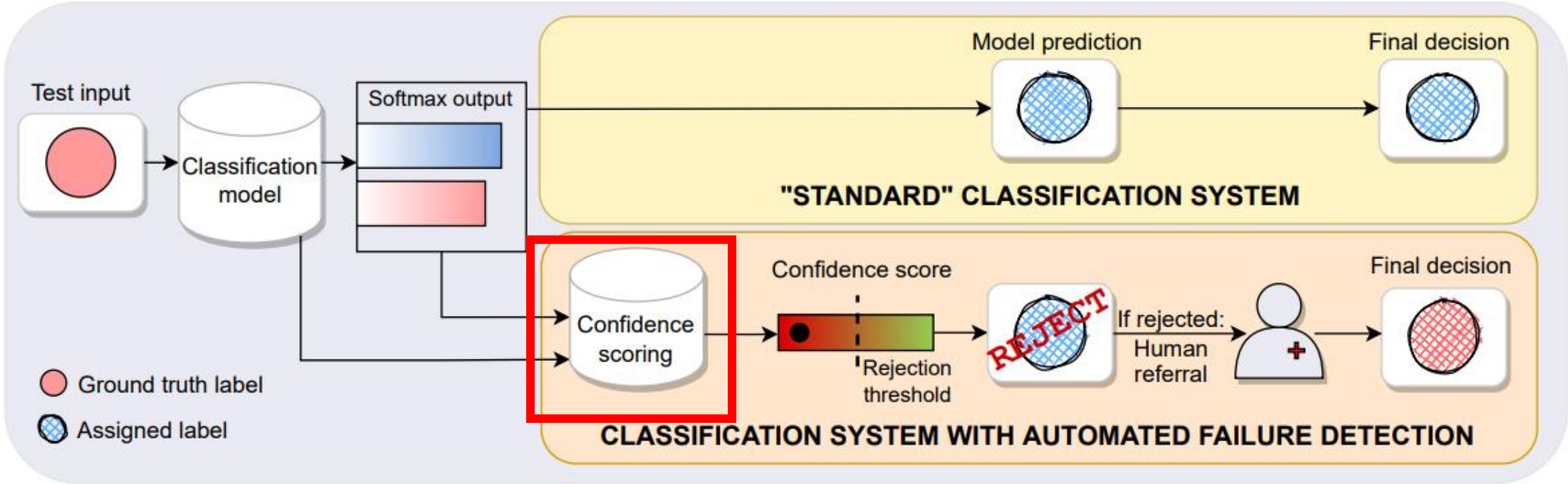


Figure 1: Standard classification system versus a system with automated failure detection. In the latter, a confidence score is computed for each test sample in addition to the model prediction to determine whether the prediction should be accepted or referred for human annotation, based on a chosen rejection threshold.

Purpose

(i) Confidence scores based on softmax outputs (softmax baseline Hendrycks & Gimpel (2016),
(ii) DOCTOR (Granese et al. (2021));
(iii) Bayesian uncertainties (MC-dropout Gal & Ghahramani (2016)),
(iv) Laplace (Laplace (1774));
(v) SWAG (Maddox et al. (2019));
(vi) Ensembles (Lakshminarayanan et al., 2016);
(vii) non-softmax based models (DUQ, Van Amersfoort et al. (2020));
(viii) confidence scores based on feature representations (**TrustScore** Jiang et al. (2018); **ConfidNet** Corbière et al. (2019))

Many studies have been evaluated in terms of robustness to out-of-distribution (OOD) inputs.

There are few studies on confidence estimation of within-distribution inputs.

The goal of this study is to compare methods for obtaining better confidence estimates of machine learning classifiers and to evaluate them for in-distribution inputs.

Purpose

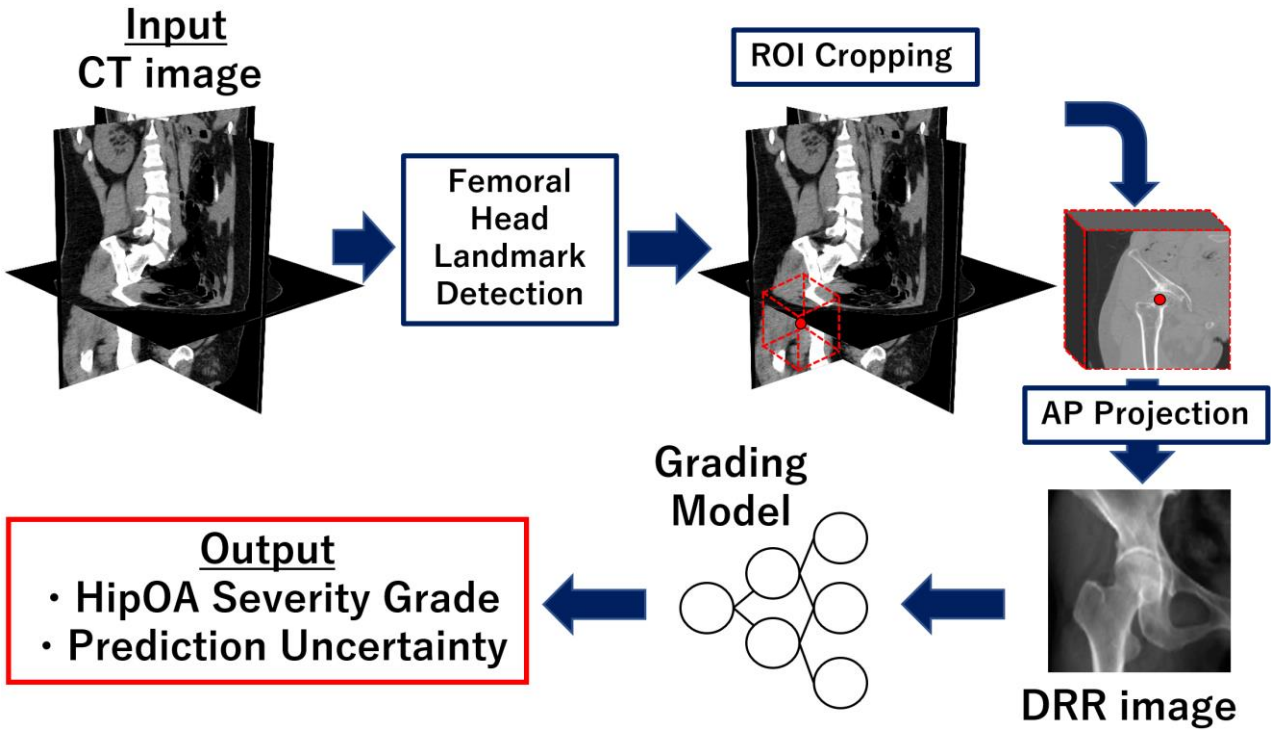
Many studies have been evaluated in terms of robustness to out-of-distribution (OOD) inputs.

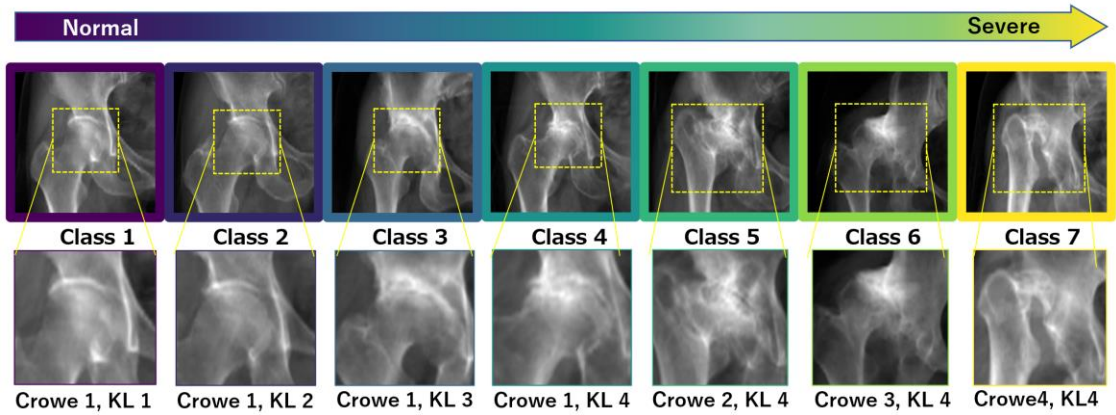
There are few studies on confidence estimation of within-distribution inputs.

The goal of this study is to compare methods for obtaining better confidence estimates of machine learning classifiers and to evaluate them for in-distribution inputs.

Motivation

- My research
  “Automatic hip osteoarthritis grading with uncertainty estimation from CT”





Motivation

$$Variance = \frac{1}{T}\sum_{i=1}^{T}(\mathrm{Softmax}(f(x;\ \theta_i)) - \bar{y})^2 \tag{1}$$

where x is the input image, T is the number of dropout samples, $\bar{y}$ is the average of the outputs obtained by dropout sampling, and θ_i is the parameter set corresponding with the sample i. In this study, T was set to 50.

Motivation

- Main point

MCdropout and other uncertainty estimation methods are inferior to normal softmax probability.

- Reason for selecting

It may be necessary to revise our approach to the evaluation of uncertainty in the future.

Methods

Base : Softmax confidence score baseline
MC : Monte-Carlo dropout (MC-S, MC-E)
Dα : Doctor
TS : TrustScore
L : Laplace
SWAG : SWAG
DUQ : DUQ
CN : ConfidNet
Ens : Ensembles

Methods

- MCdropout

Gal & Ghahramani (2016) showed that training a neural network with dropout regularization (Srivastava et al., 2014) produces a Bayesian approximation of the posterior, where the approximation is obtained by Monte-Carlo sampling of the network's parameters
i.e. by applying dropout at test-time and averaging the outputs over several inference passes.
The confidence in the prediction can be approximated by the negative entropy of the output.

https://www.ai-shift.co.jp/techblog/2518
https://docs.aws.amazon.com/ja_jp/prescriptive-guidance/latest/ml-quantifying-uncertainty/mc-dropout.html
https://st1990.hatenablog.com/entry/2019/07/31/010010

Methods

https://www.iwanttobeacat.com/entry/2019/09/22/230333
https://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/pdf/2133-07.pdf
https://masamunetogetoge.com/laplace-bayeslogistic
https://proceedings.mlr.press/v139/daxberger21a.html

- *Laplace approximation* (L) (Laplace, 1774): where the posterior is locally approximated with a Gaussian distribution centered at a local maximum (in practice around the maximum-a-posteriori estimate), with covariance matrix corresponding to the local curvature (obtained by an approximation of the Hessian). Recent work from Daxberger et al. (2021) proposed a simple-to-use, lightweight, Python implementation of this approximation, enabling benchmarking the approach easily on pre-trained neural networks.
- *SWAG*: Maddox et al. (2019) define a Gaussian distribution whose mean is parameterized by the stochastic weight averaging (Izmailov et al., 2018) solution, and whose covariance matrix is a low rank matrix plus diagonal covariance derived from the stochastic gradient descent iterates. They then sample several times from this distribution to form the approximate posterior solution.

https://arxiv.org/abs/1902.02476
https://github.com/wjmaddox/swa_gaussian

Methods

- *Deep Ensembles* (Ens): Lakshminarayanan et al. (2016) have shown that ensembling predictions from several trained models can yield better calibrated uncertainty estimates than the ones obtain from single deterministic models and even from Bayesian neural networks, in particular due to their increased capacity to capture multi-modal solutions.

- *Deterministic Uncertainty Quantification* (DUQ): Van Amersfoort et al. (2020) estimate predictive confidence based on distances between points and class centroids in the embedding space and demonstrated improved OOD performance.

- *DOCTOR*: Granese et al. (2021) proposed to use $D_\alpha(x) = \frac{1-g(x)}{g(x)}$ where $g(x) = \sum_c \hat{p}_c^2(x)$ as a score quantifying the likelihood of being misclassified (i.e. negative confidence score) as an alternative to the classic predicted softmax confidence score.

Other Methods

- Trust Score (TS)

Jiang et al. (2018) construct a neighbour-graph in the embedding space from the penultimate layer (on the training set) and use distances in this space to derive TrustScore (TS). In practice, this confidence score is defined as the ratio between: (i) the distance between the test point and the closest point that does not belong to the predicted class; (ii) the distance between the test point and closest point that belongs to the predicted class.

- ConfidNet (CN)

Corbière et al. (2019) proposed ConfidNet (CN) a regression network that is placed on top of the classification model to predict the “true class probability” i.e. the probability predicted by the main model for the true class (as opposed to the probability of the predicted class). In other words, the image is fed through the trained classification model to extract its embedding (penultimate layer), which in turn is fed into ConfidNet which predicts the softmax output for the true class, as originally predicted by the main model

Experiments

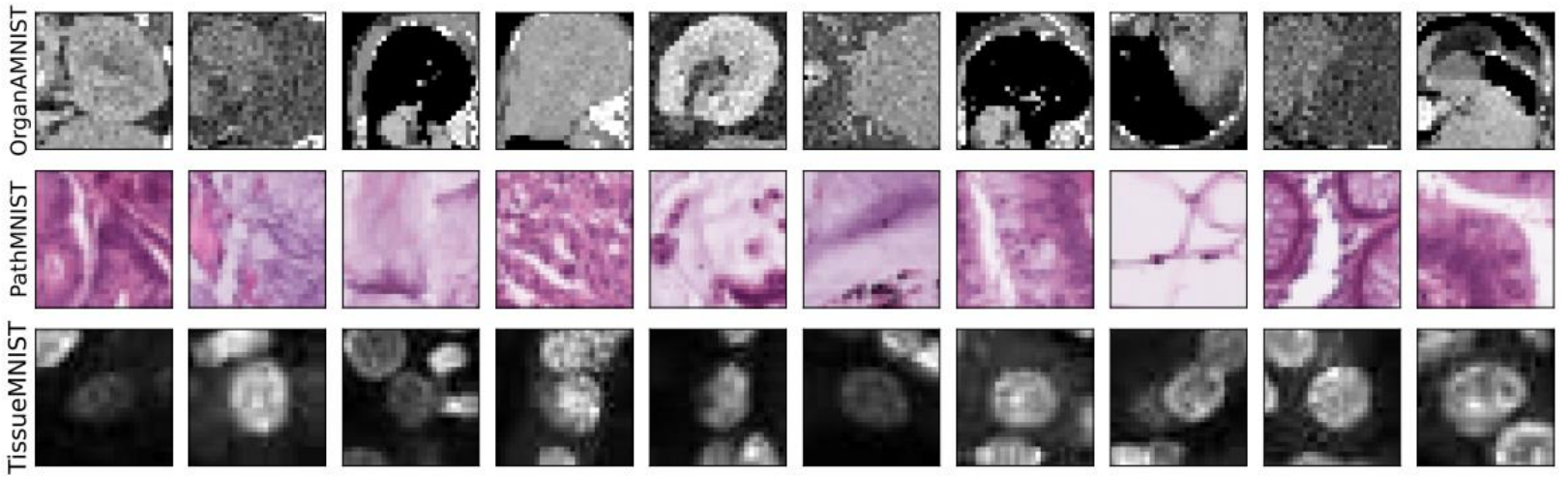


Figure 3: Examples images of each dataset used in the failure detection testbed. For the EyePACS dataset we depict the preprocessed images.

- Dataset1

MedMNIST-v2

https://medmnist.com/

3 tasks (5000images testset)

1. PathMNIST

   9 colon disease class classification

2. TissueMNIST

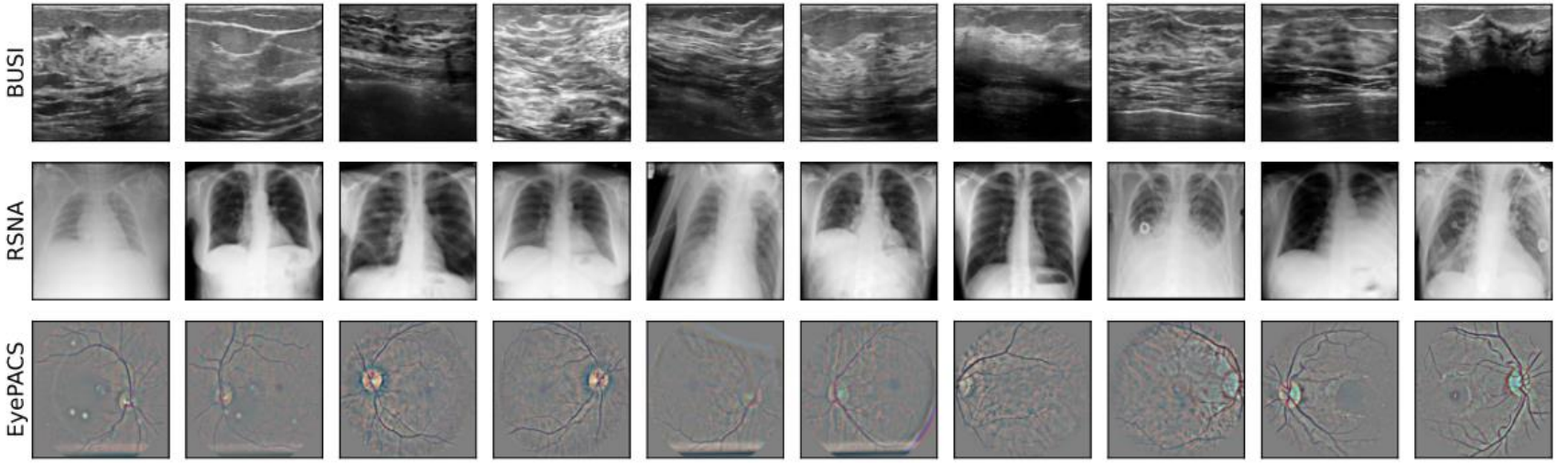
   kidney cortex cells microscope images 8 class classification

3. OrganAMNIST

   center slices from abdominal CT images in axial view, classified by organ type (11 classes).

PathMNIST	Colon Pathology	Multi-Class (9)	107,180	89,996 / 10,004 / 7,180
TissueMNIST	Kidney Cortex Microscope	Multi-Class (8)	236,386	165,466 / 23,640 / 47,280
OrganAMNIST	Abdominal CT	Multi-Class (11)	58,850	34,581 / 6,491 / 17,778

Experiments



- Dataset2

RSNA Pneumonia Detection Challenge (RSNA) https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pneumonia-detection-challenge-2018

- Dataset3

Breast Ultrasaound Image Dataset (BUSI)

- Dataset4

https://academictorrents.com/details/d0b7b7ae40610bbeaea385aeb51658f527c86a16

EyePACS2 Diabetic Retinopathy Detechtion Challenge (EyePACS)

https://www.kaggle.com/c/diabetic-retinopathy-detection

(images resized to 224×224 for RNSA and BUSI, 512×512 for EyePACS)

Experiments

Evaluation

•It aims to predict whether a given sample has been classified correctly and uses the confidence score as the prediction score.

•Specific assessments were evaluated with standard binary classification metrics such as ROC-AUC or false positive rate (FPR) at a given true positive rate (TPR).

Experiments

Model Architecture

- ResNet-18 for MedMNIST, BUSI tasks
- ResNet-50 for RSNA Pneumonia, EyePACS tasks

All models are trained with an additional dropout layer after each weights layer to be able to run the MC-dropout comparison (with dropout probability p=0.1 for all experiments, based on validation performance).

https://github.com/melanibe/failure_detection_benchmark

Results

Base : Softmax confidence score baseline
MC : Monte-Carlo dropout (MC-S, MC-E)
Dα : Doctor
TS : TrustScore
L : Laplace
SWAG : SWAG
DUQ : DUQ
CN : ConfidNet
Ens : Ensembles

Median of Baseline

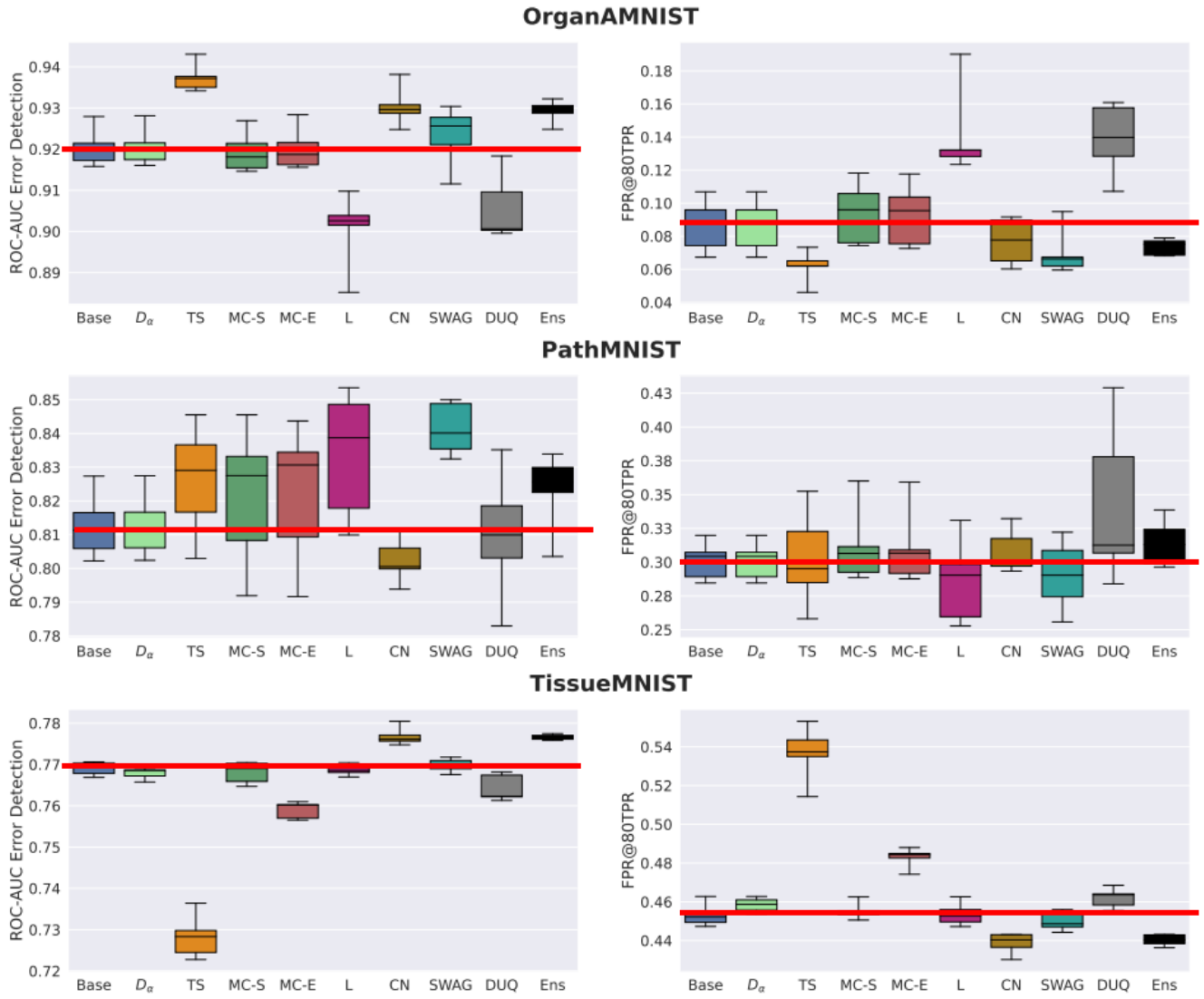


Figure 4: Failure detection benchmark for smaller resolutation datasets: OrganAMNIST, PathMNIST, TissueMNIST. Comparison of Baseline (Base), DOCTOR D_α, TrustScore (TS), MC-dropout with softmax score (MC-S), MC-dropout with entropy score (MC-E), Laplace (L), ConfidNet (CN), SWAG, DUQ and Ensemble (Ens). Except for Ensemble, boxplots are constructed with results of repeated training over 5 seeds, whiskers denote minimum and maximum value observed. For Ensemble, we formed 5 different ensembles by taking 5 different combinations of 3 out of the 5 trained models.

Results

Base : Softmax confidence score baseline
MC : Monte-Carlo dropout (MC-S, MC-E)
Dα : Doctor
TS : TrustScore
L : Laplace
SWAG : SWAG
DUQ : DUQ
CN : ConfidNet
Ens : Ensembles

Median of Baseline

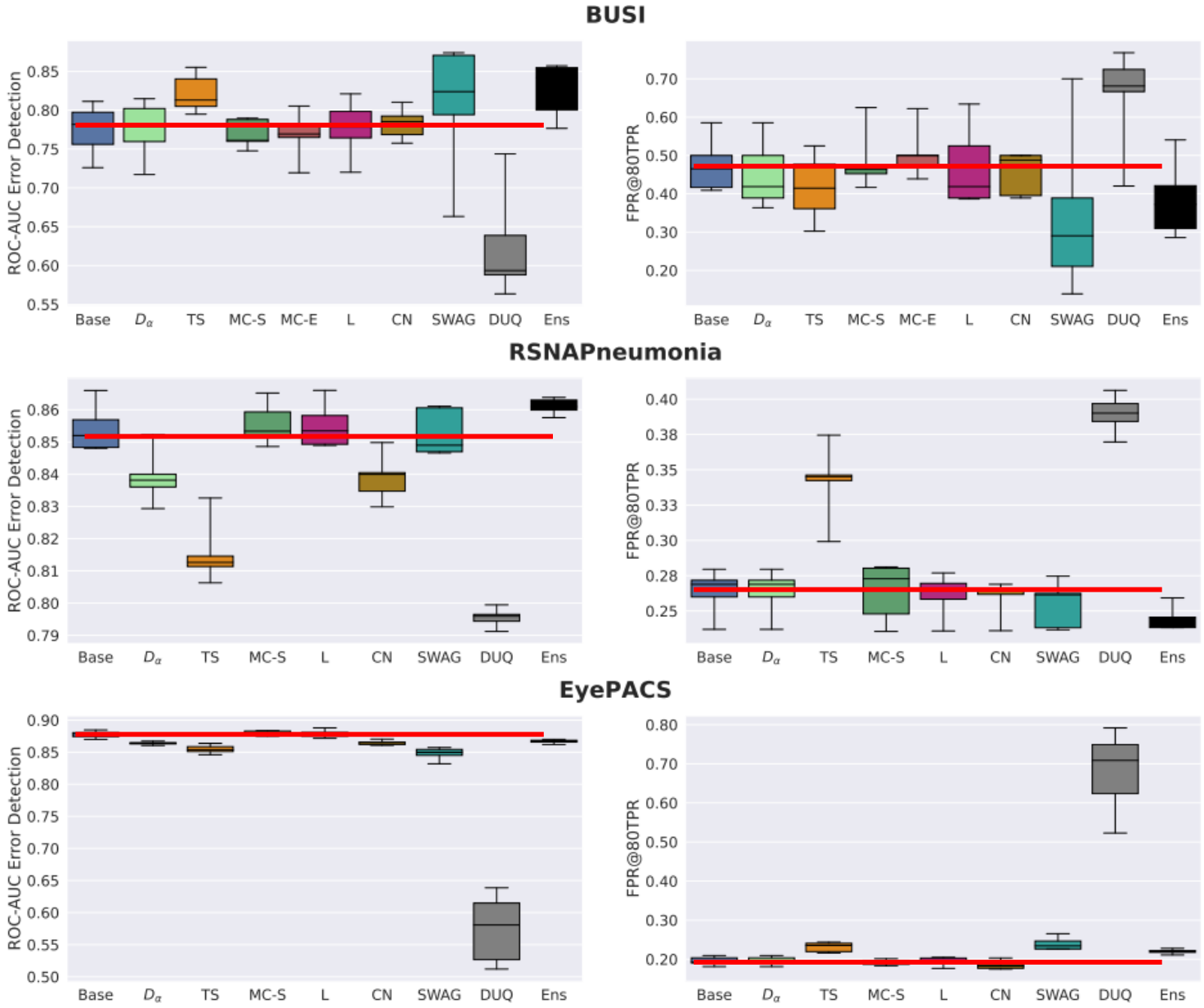
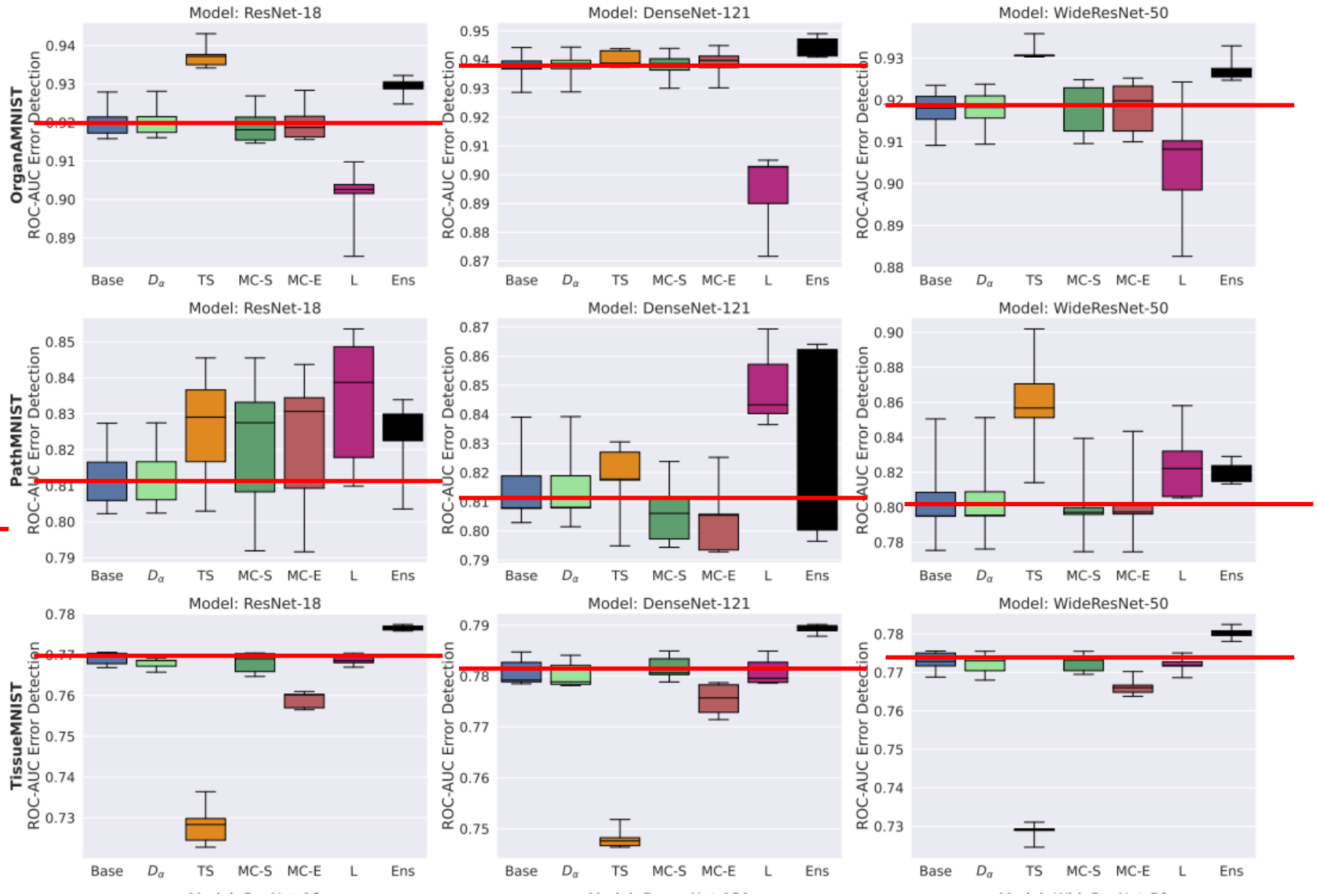


Figure 5: Failure detection benchmark for higher resolution datasets: BUSI, RNSA Pneumonia Detection and EyePACS datasets. Note that we excluded MC-E on the binary tasks as the chosen classification threshold was different from 0.5.

Results

Median of Baseline



Results

Median of Baseline

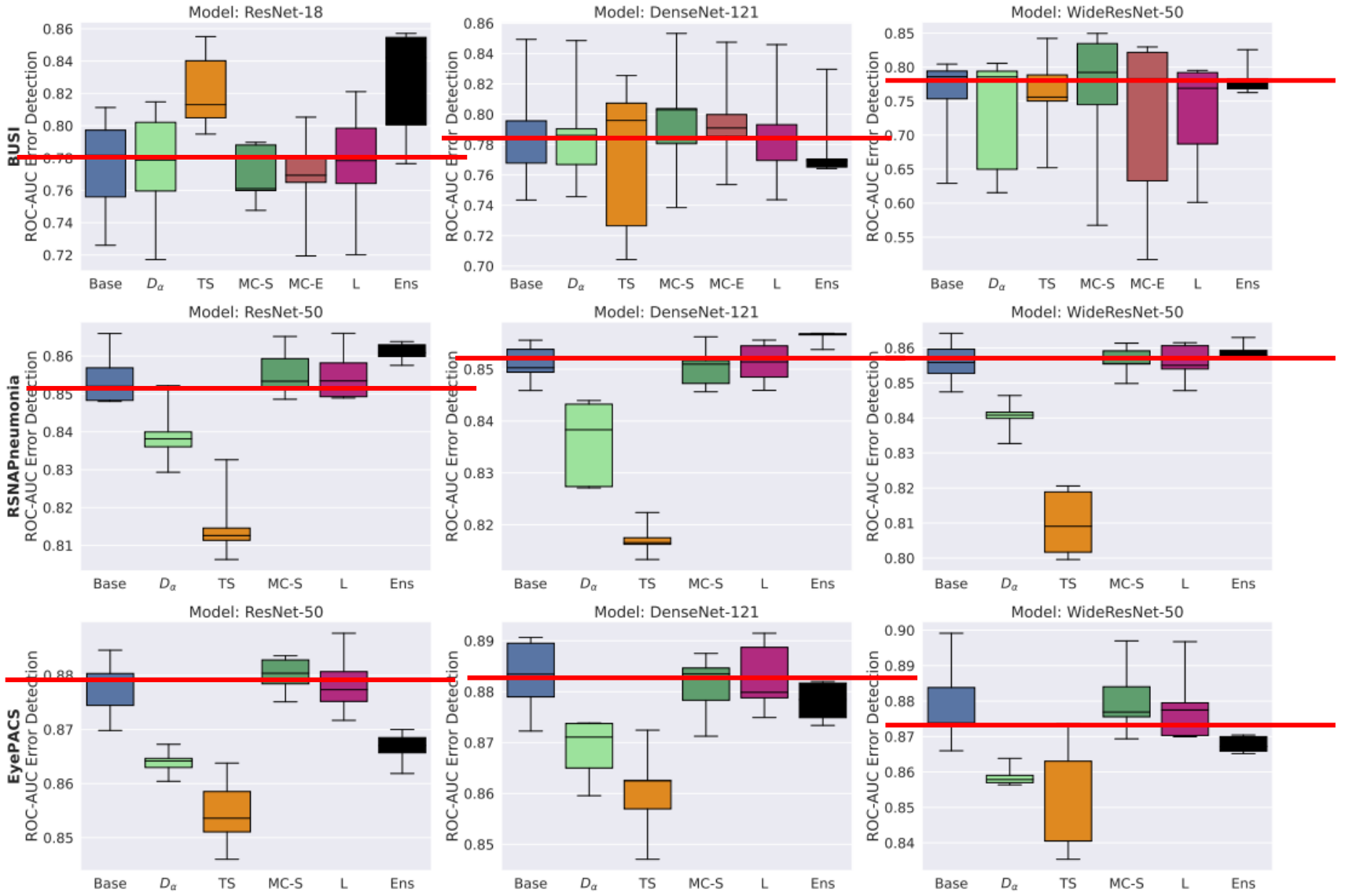


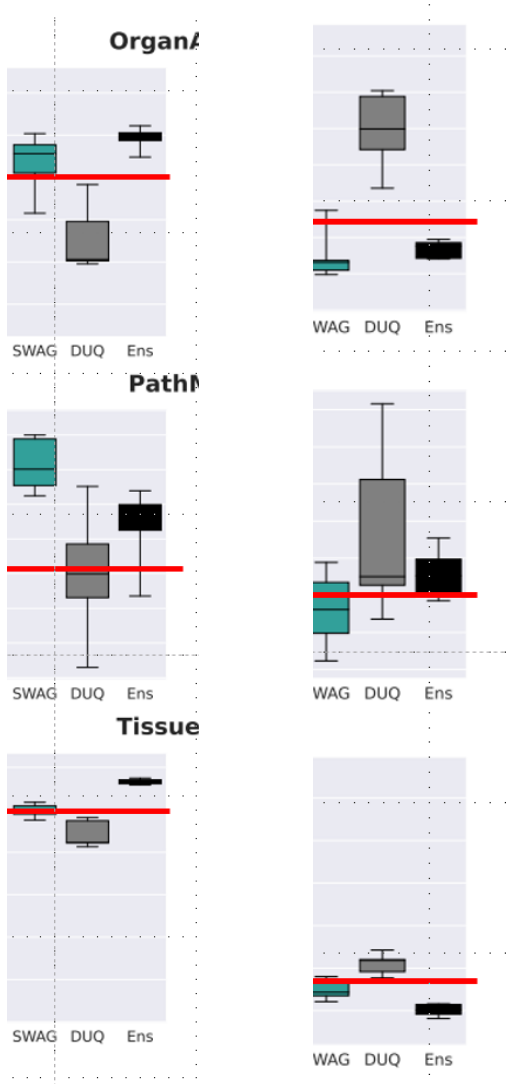
Figure 6: Model architecture effect ablation study. SWAG, ConfidNet and DUQ not included in this ablatior because of computational cost (as they all require separate training).

Results

Table 2: Classification performance for ResNet models: accuracy for multiclass tasks and ROC-AUC for binary tasks. For all methods except Ensemble, we report the average over 5 seeds, standard deviation in brackets. For Ensemble, we formed 5 different ensembles by taking 5 different combinations of 3 out of the 5 trained models, we report the average over of the 5 ensembles, standard deviation in brackets.

Dataset	Baseline	MC	Laplace	SWAG	DUQ	Ensemble
OrganAMNIST	.902 (.005)	.902 (.005)	.901 (.006)	.910 (.004)	.917 (.004)	**.921 (.002)**
PathMNIST	.838 (.009)	.832 (.010)	.835 (.010)	.830 (.010)	.828 (.025)	**.853 (.002)**
TissueMNIST	.663 (.004)	.664 (.003)	.663 (.004)	.670 (.001)	.655 (.007)	**.682 (.001)**
BUSI	.740 (.020)	.740 (.021)	.740 (.020)	**.792 (.017)**	.561 (.000)	.742 (.017)
RSNA	.871 (.007)	.873 (.006)	.871 (.006)	.873 (.003)	.865 (.003)	**.877 (.003)**
EyePACS	.899 (.006)	.899 (.007)	.899 (.007)	.913 (.004)	.730 (.007)	**.918(.002)**

Discussion



ROCAUC **TPR 80%, FPR 20%**

Discussion

Thier investigation showed that the softmax confidence score baseline is difficult to beat: none of the benchmarked advanced confidence scoring methods were able to consistently outperform the baseline across datasets.

Importantly, their experiments show that previously demonstrated improved robustness for out-of-distribution detection or model does not necessarily translate to improvements in error detection for in-domain inputs.

Conclusion

Results of this study indicate that current uncertainty scoring, including a simple softmax baseline, are able to detect some misclassified cases at a substantially better than chance level (ROCAUC > .75 for all benchmarked datasets), demonstrating the potential to improve practical performance of AI systems by deferring suspicious cases to humans.