KMi Knowledge Media Institute

The Open University

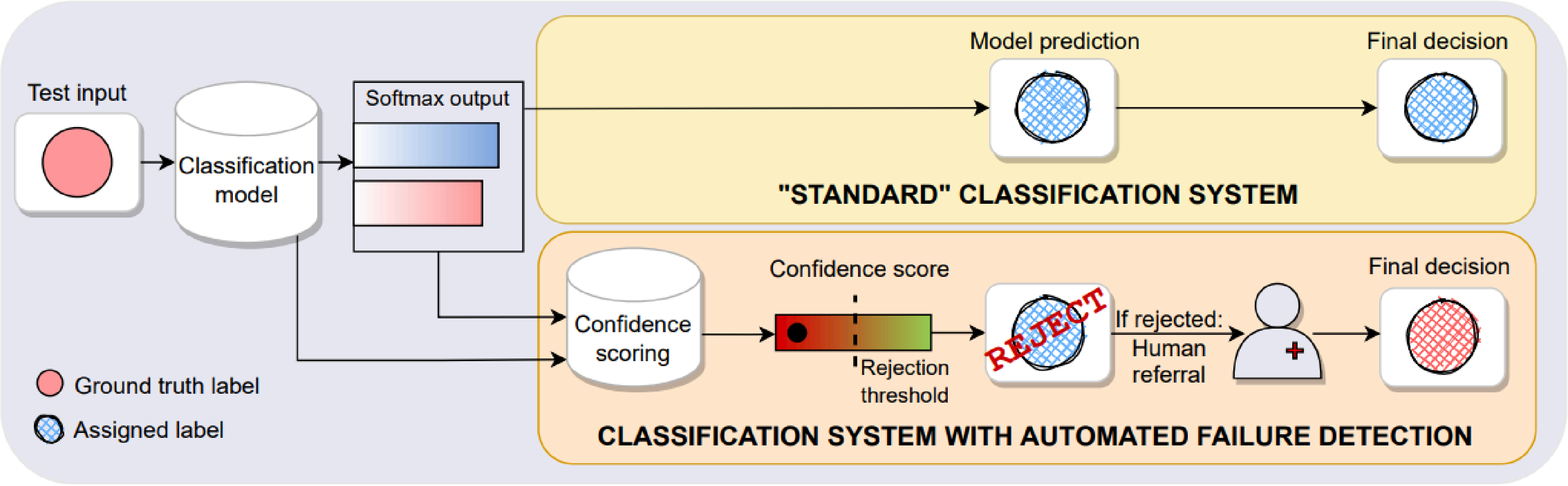
AISafety 2023

IJCAI/2023 MACAO

Empirical Optimal Risk to Quantify Failure Detection for Model Trustworthiness

Shuang Ao
PhD Candidate
Knowledge Media Institute (KMi)
The Open University, UK

Motivation

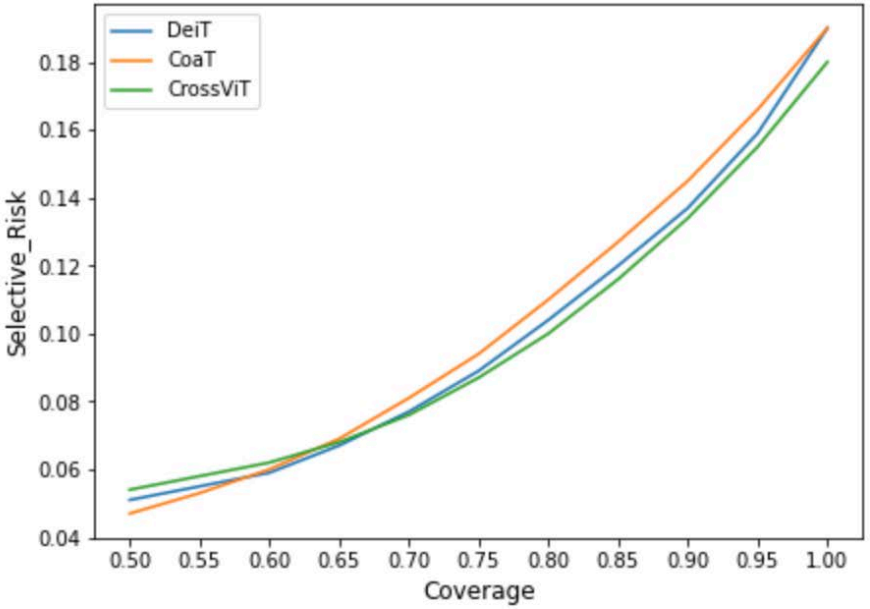


Automatic Failure Detection (FD)

Bernhardt, Mélanie, Fabio De Sousa Ribeiro, and Ben Glocker. "Failure detection in medical image classification: A reality check and benchmarking testbed." *arXiv preprint arXiv:2205.14094* (2022).

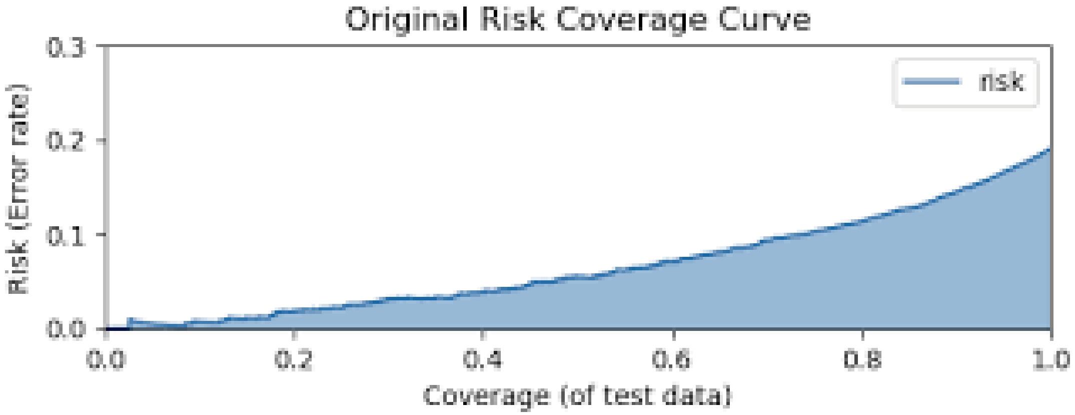
Motivation

Qualitative



Risk-Coverage Curve

Quantitative

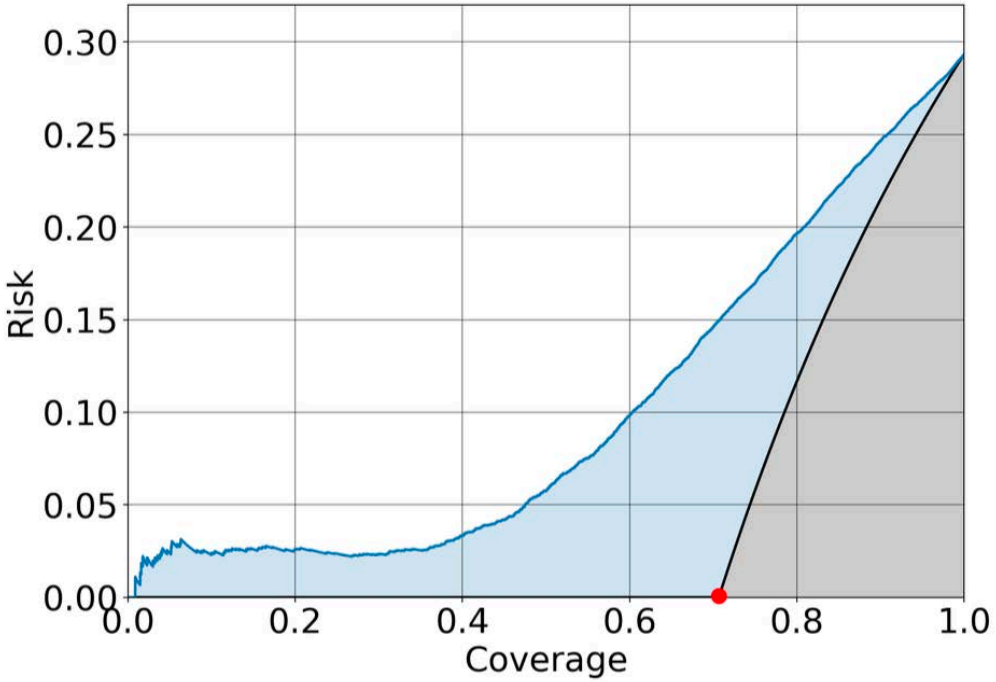


Risk-Coverage Curve

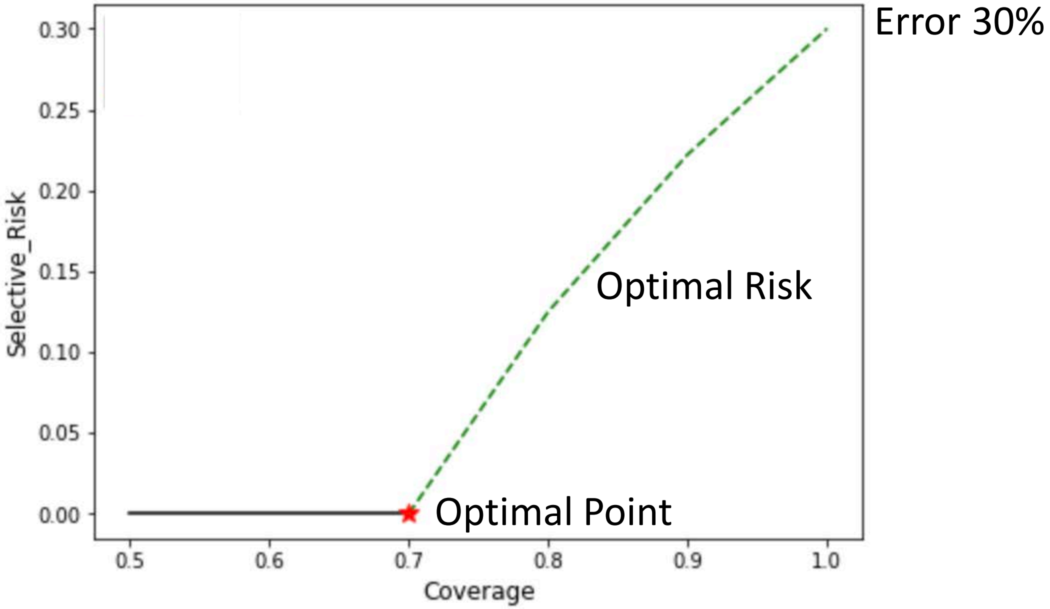
Ao, Shuang. "Building Safe and Reliable AI Systems for Safety Critical Tasks with Vision-Language Processing." European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2023.

Motivation

E-AURC

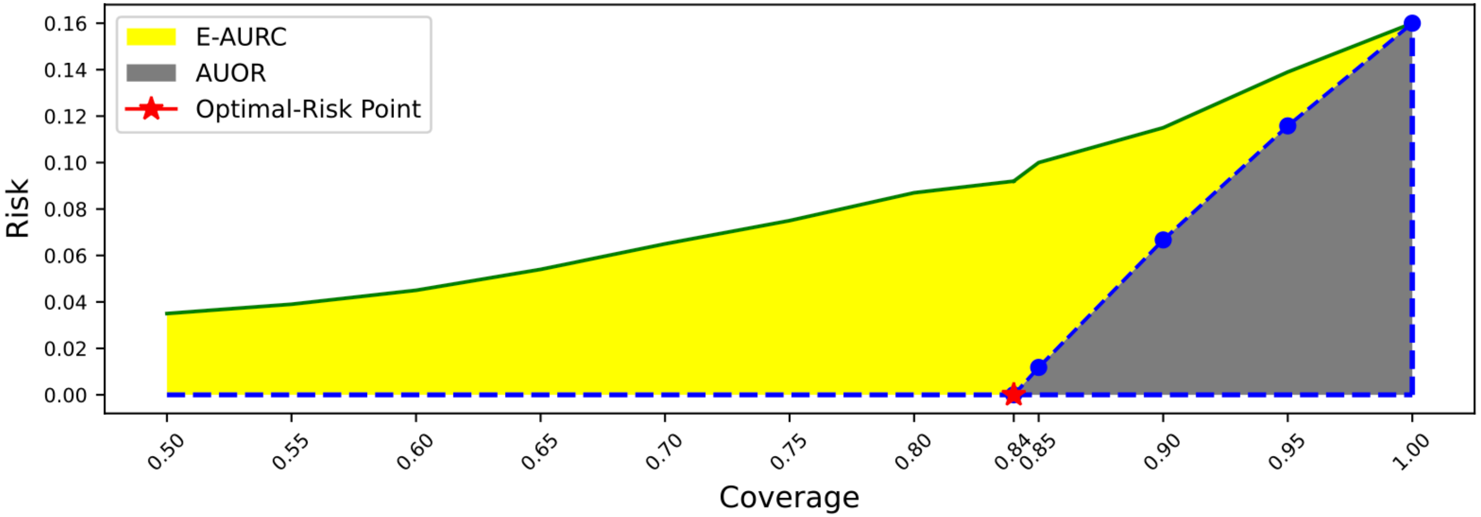


Ideal Risk - Coverage Curve



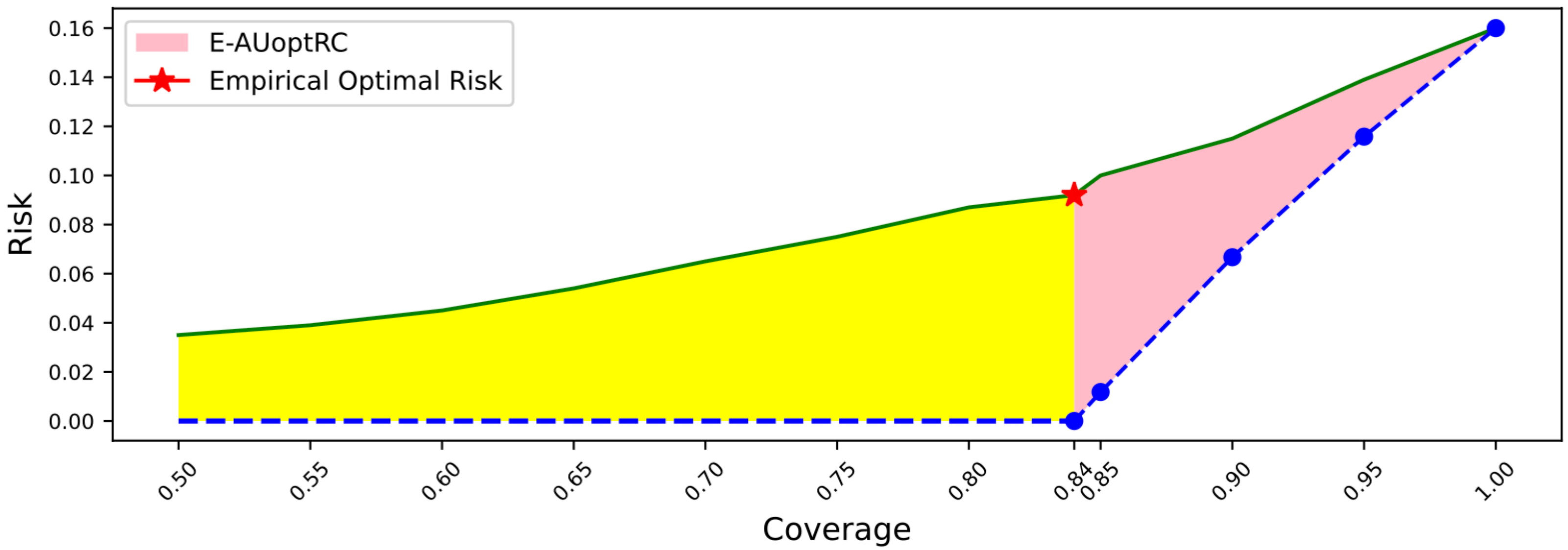
Geifman, Yonatan, Guy Uziel, and Ran El-Yaniv. "Bias-reduced uncertainty estimation for deep neural classifiers." *arXiv preprint arXiv:1805.08206* (2018).

Limitations of E-AURC



1. With a perfectly calibrated model, samples falling into the coverage from 0 to optimal point are already highly trusted ones.
2. Samples as high uncertainty ones, and the corresponding risk here should be primarily utilised to determine the trustworthiness of the model.

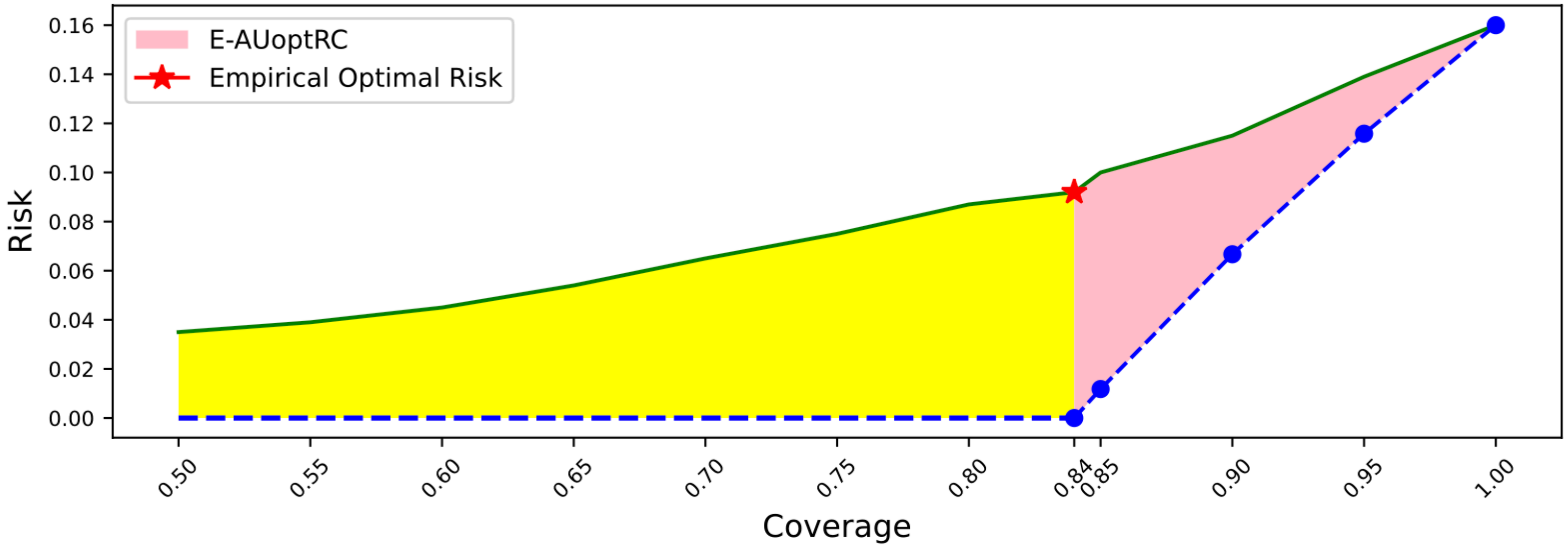
Proposed Method: E-AUoptRC



E-AUoptRC = E-AURC – AURC^{op}

1. It is more important to compare models in the region that errors are made.
2. It is more practical for deployment, as it is unlikely to discard more than half of data in applications;
3. The smaller E-AUoptRC indicates more samples with high uncertainty are successfully removed so that the model prediction on the remaining data will be more reliable.

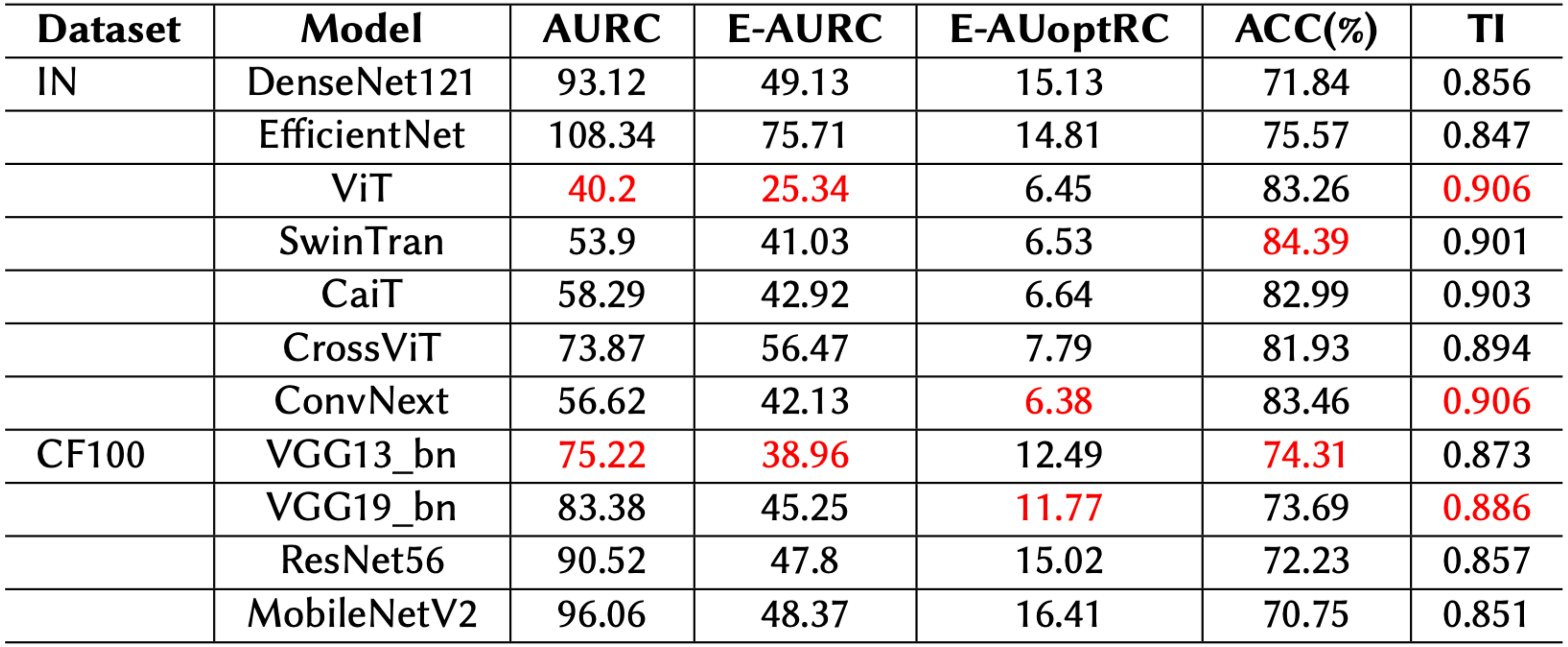
Proposed Method: Trust Index (TI)



Trust Index (TI) = 1 - Risk^{OP}

1. Model accuracy is not enough to show the model performance; higher accuracy does not mean a well-calibrated/ trusted model.
2. One way to show both the performance and generalization of the model.
3. TI: complimentary of the model accuracy; showing how much of the model prediction is trusted

Results

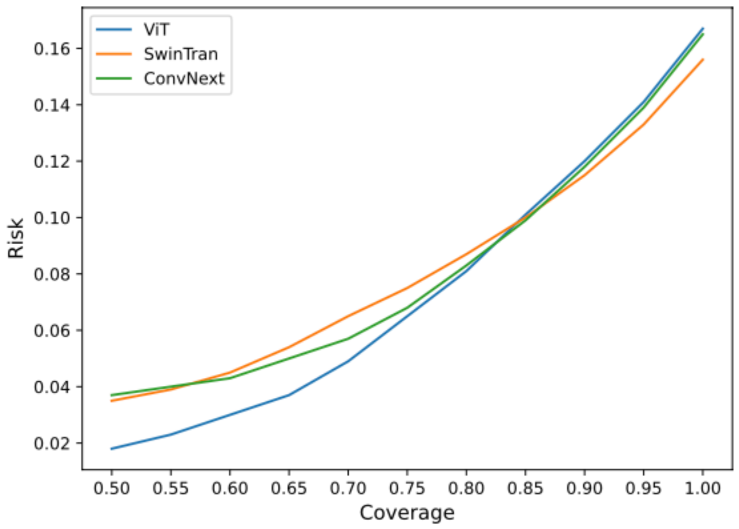


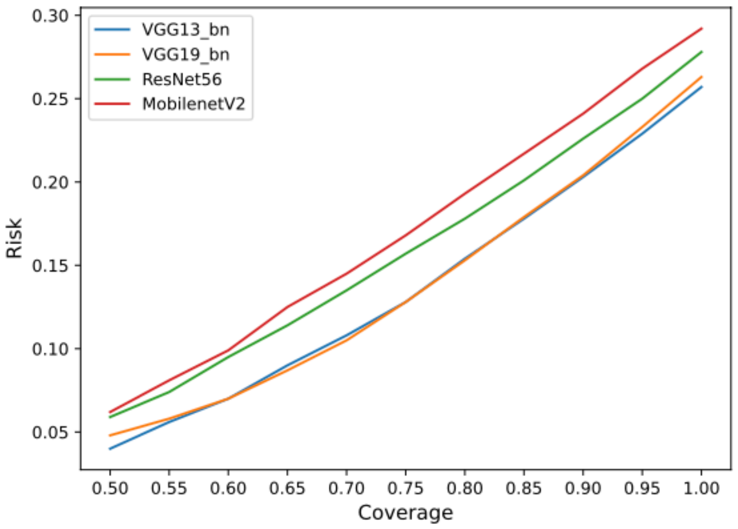
Dataset	Model	AURC	E-AURC	E-AUoptRC	ACC(%)	TI
IN	DenseNet121	93.12	49.13	15.13	71.84	0.856
	EfficientNet	108.34	75.71	14.81	75.57	0.847
	ViT	40.2	25.34	6.45	83.26	0.906
	SwinTran	53.9	41.03	6.53	84.39	0.901
	CaiT	58.29	42.92	6.64	82.99	0.903
	CrossViT	73.87	56.47	7.79	81.93	0.894
	ConvNext	56.62	42.13	6.38	83.46	0.906
CF100	VGG13_bn	75.22	38.96	12.49	74.31	0.873
	VGG19_bn	83.38	45.25	11.77	73.69	0.886
	ResNet56	90.52	47.8	15.02	72.23	0.857
	MobileNetV2	96.06	48.37	16.41	70.75	0.851

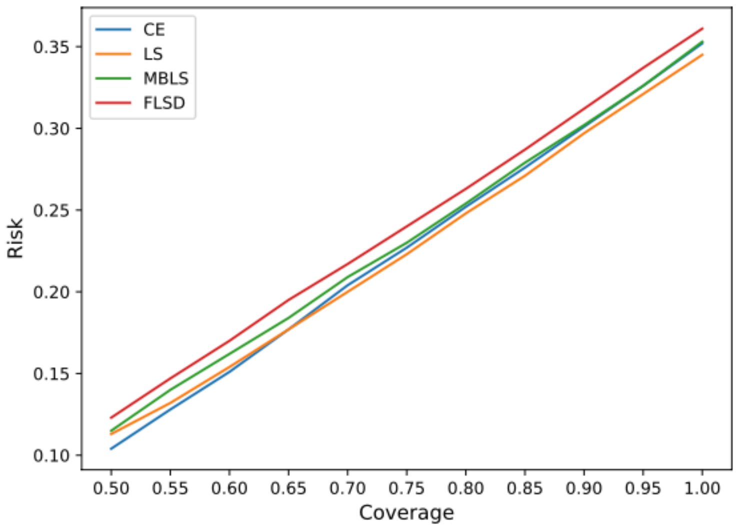
1. E-AUoptRC reveals the real failure detection performance;
2. AURC and E-AURC are unable to show the true FD by calculating full coverage.

Results

Method	AURC	E-AURC	E-AUoptRC	ACC(%)	TI	ECE(%)	ECE_OP(%)
CE	128.71	57.94	22.13	64.82	0.821	3.76	4.25
LS	131.54	63.51	21.98	65.46	0.824	2.8	2.04
MBLS	135.39	64.27	22.78	64.74	0.817	1.87	0.92
FL	146.42	68.61	25.05	63.24	0.807	3.1	3.53
FLSD	139.72	64.85	23.91	63.89	0.812	2.8	2.49







Take Home Message

- E-AUoptRC is able to **reveal** the real capability of failure detection in a model.
- Trust Index indicates both model accuracy and **calibration**, which is a complementary of conventional accuracy metric.
- These metrics will further help to investigate the **threshold selection** for failure detection.

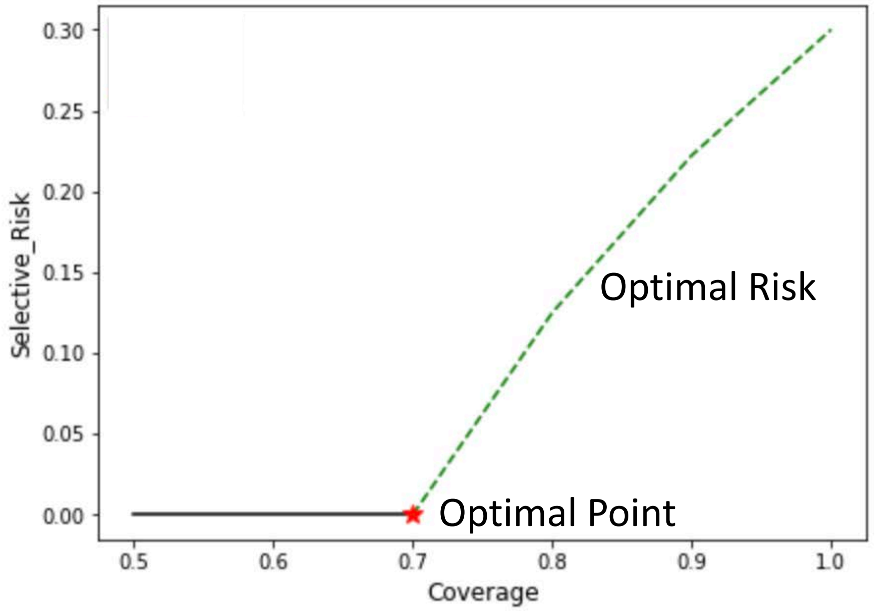
Thank You!

Preliminary – Optimal Risk

Sample size: 100
70 correct predicted
30 wrong predicted
Accuracy: 70%
Error: 30%
In ideal situation:

Coverage	Error
0.7	0/70
0.8	10/80
0.9	20/90
1	30/100

Risk - Coverage Curve

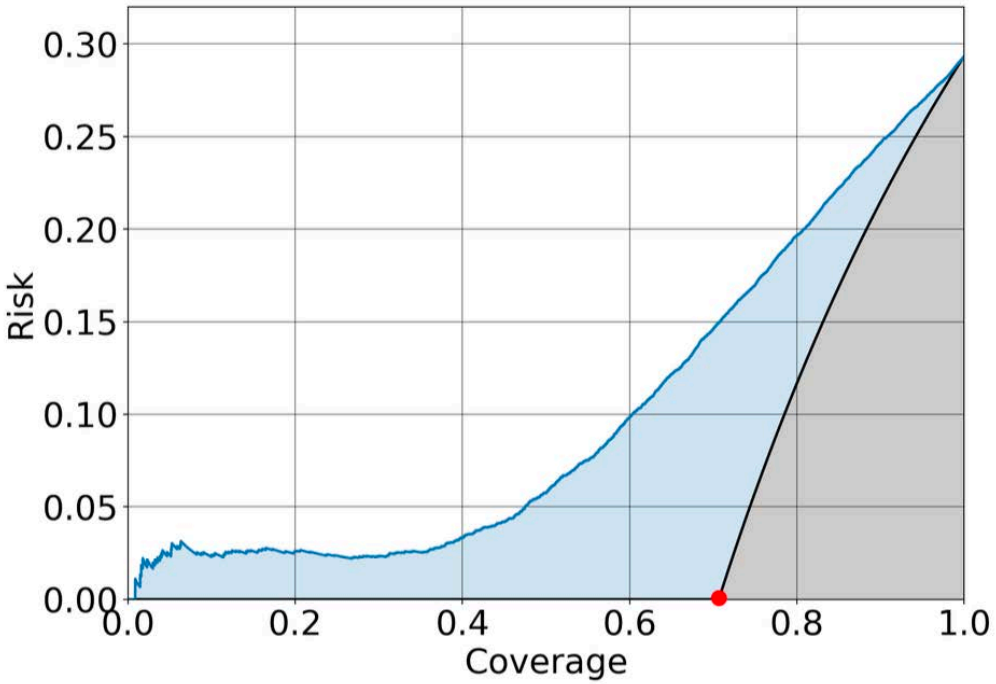


Error 30%

Optimal Risk

Optimal Point

Excess AURC (E-AURC)



AURC: area under risk-coverage curve (blue+grey area)
Normalized AURC: Unitless performance measure (grey area)
Excess-AURC (E-AURC): AURC - Normalized AURC
Optimal point (op) : red dot

Geifman, Yonatan, Guy Uziel, and Ran El-Yaniv. "Bias-reduced uncertainty estimation for deep neural classifiers." *arXiv preprint arXiv:1805.08206* (2018).

Trust Index (TI)

Motivation:

- Model accuracy is not enough to show the model performance; higher accuracy does not means a well-calibrated/ trusted model.
- One way to show both the performance and generalization of the model.
- TI: complimentary of the model accuracy; showing how much of the model prediction is trusted

Example:

Sample size: 1000

Model accuracy: 80%

After removing err% of data:

Accuracy at op: 0.9

TI for model: 0.9

Indication 0.9 of model prediction is trusted