Empirical Optimal Risk to Quantify Model Trustworthiness for Failure Detection

Shuang Ao[1,*], Stefan Rueger[2] and Advaith Siddharthan[3]

,2,3[1]Knowledge Media Institute, The Open University, Walton Hall, Kents Hill, Milton Keynes MK7 6AA, UK

Abstract

Failure detection (FD) in AI systems is a crucial safeguard for the deployment for safety-critical tasks. The common evaluation method of FD performance is the Risk-coverage (RC) curve, which reveals the trade-off between the data coverage rate and the performance on accepted data. One common way to quantify the RC curve by calculating the area under the RC curve. However, this metric does not inform on how suited any method is for FD, or what the optimal coverage rate should be. As FD aims to achieve higher performance with fewer data discarded, evaluating with partial coverage excluding the most uncertain samples is more intuitive and meaningful than full coverage. In addition, there is an optimal point in the coverage where the model could achieve ideal performance theoretically. We propose the Excess Area Under the Optimal RC Curve (E-AUoptRC), with the area in coverage from the optimal point to the full coverage. Further, the model performance at this optimal point can represent both model learning ability and calibration. We propose it as the Trust Index (TI), a complementary evaluation metric to the overall model accuracy. We report extensive experiments on three benchmark image datasets with ten variants of transformer and CNN models. Our results show that our proposed methods can better reflect the model trustworthiness than existing evaluation metrics. We further observe that the model with high overall accuracy does not always yield the high TI, which indicates the necessity of the proposed Trust Index as a complementary metric to the model overall accuracy. The code are available at https://github.com/AoShuang92/optimal_risk.

Keywords

Failure Detection, Evaluation, Trustworthiness, Risk-Coverage Curve, Model Calibration

1. Introduction

The deployment of deep neural networks (DNNs) in safety-critical applications such as autonomous driving [1] and medical diagnosing [2, 3] requires high trustworthiness and reliability, as mistakes can be expensive and raise serious concerns. To reduce mispredictions, a model should be equipped with a safeguard for automatic failure detection [4, 5, 6] or a reject option [7], where samples with high uncertainty or low confidence can be discarded or sent to an expert or the third system. Specifically, failure detection (FD) determines the portion of coverage over the entire dataset deemed to be safe predictions and discards data using a threshold on model confidence or uncertainty. If the confidence or uncertainty is below or above the threshold, the model rejects samples and defers them to human experts or third systems to re-evaluate. Otherwise, the model considers these samples in a coverage range for safe and trusted prediction. FD is beneficial for gaining higher trust from users and for time and cost savings by only requiring human interventions for a small percentage of data.

One of the criteria for FD is for the model to achieve better performance with fewer instances removed; hence the evaluation is about the trade-off between the coverage of data and model accuracy or risk (error). Popular visualisation methods of FD performance such as risk-coverage (RC) curve [8] and accuracy-rejection curves (ARCs) [9, 10] plot model risk or accuracy against coverage of data. However, the quantification of FD performance is a less explored domain. Recent studies attempt to quantify FD by using the area under the RC-curve (AURC) [11] and the area under the ARCs [10]. Nevertheless, both methods include the full coverage of data, ignoring the selection of thresholds and the FD performance under and above thresholds.

Theoretically, a perfectly calibrated model should achieve the ideal performance (i.e., accuracy of 1) after removing the most uncertain samples in numbers equal to the error percentage. In other words, the perfect performance takes place hypothetically by covering the portion of samples equivalent to model accuracy. Therefore, the model risk is supposed to be 0 at this very coverage point, which is denoted as the optimal point in work on uncertainty estimation [12] as shown in Figure 1. A perfectly calibrated model should not contain any risk before the optimal point, whereas the risk increases monotonically until the model error after the optimal point. This risk

AISafety-SafeRL 2023 Workshop (IJCAI), August 19–21, 2023, Macao, SAR, China

[*] Corresponding author.

shuang.ao@open.ac.uk (S. Ao); stefan.rueger@open.ac.uk (S. Rueger); advaith.siddharthan@open.ac.uk (A. Siddharthan)
https://github.com/AoShuang92 (S. Ao); https://kmi.open.ac.uk/people/member/stefan-rueger (S. Rueger); https://people.kmi.open.ac.uk/advaith-siddharthan/ (A. Siddharthan)
0000-0003-2648-3082 (S. Ao); 0000-0003-0796-8826 (A. Siddharthan)

is naturally inherited from the model as DNNs cannot obtain the perfect performance in practice, thus, should perhaps be discounted in FD evaluations. Based on this hypothesis, Geifman et.al [12] exclude the area under the optimal risk (grey part in Figure 1) for the AURC and propose the metric of Excess-AURC (E-AURC) (yellow part in Figure 1). However, this still evaluates FD based on the whole dataset even though some data are supposed to be safe and trusted predictions.

As the percentage of rejected samples is generally customised during deployment of a model, there is a lack of common ground for a fair comparison of failure detection among models with varying accuracies. In addition, most of the existing evaluation metrics (i.e., AURC, E-AURC) measure the entire area under the curve, which cannot reveal the FD performance for a specific coverage. For example, the performance of a model at very low coverage is not of interest to real applications. To address the above issues, we propose the Excess area under the optimal RC curve (E-AUoptRC) as an alternative metric for failure detection that considers the risk in the range from the optimal point to the full coverage (shown as pink area in Figure 2). We emphasise this area for reasons as follows: (1) with a perfectly calibrated model, samples falling into the coverage from 0 to optimal point (yellow area in Figure 2) are already highly trusted ones; (2) we argue that it is more important to compare models in the region that errors are made, for instance, samples in the E-AUoptRC include the high uncertainty ones, and the corresponding risk here should be primarily utilised to determine the trustworthiness of the model. (3) Furthermore, with our precise method of FD quantification, a model with lower accuracy may yield higher trustworthiness and vice versa, capturing the intuition that a model with higher accuracy may not be the most trusted one. Finally, we propose a Trust Index (TI) as a novel evaluation metric, which measures the accuracy of the model at the optimal point, mimics the behaviour of E-AUoptRC, and is easier to compute. The Trust Index combines the performance and calibration of the model into a single metric. A higher TI suggests better model performance and calibration and higher trust and reliability of the model predictions.

Our contributions and findings are summarized as below:

1. We propose the E-AUoptRC to quantify the RC curve with the coverage from the optimal point to the full coverage.
2. We propose Trust Index as an evaluation metric.
3. With extensive experiments and observations we find that: (i) a model with higher AURC or E-AURC can obtain lower E-AUoptRC ; (ii) A model with a high overall accuracy does not necessarily yield higher Trust Index; (iii) Our proposed methods can better evaluate failure detection for model trustworthiness.

2. Related Work

2.1. Failure Detection

In the deployment of safety-critical scenarios, DNNs tend to fail silently by providing high-confidence in woefully incorrect predictions, which makes the uncertainty estimation a great concern to AI safety [13, 14]. These high-confidence predictions are often produced by the softmax function as it is computed with a fast-growing exponential function. It is clearly necessary to identify potentially wrong predictions. Hendrycks et al. [4] proposed to detect misclassified samples by enlarging the softmax probabilities between correct and incorrect samples. Meanwhile, utilizing true class probability instead of maximum class probability has been shown to be more reliable in the context of failure detection [5]. In addition, training the model with data that can reflect the complexity of real-world scenario can improve the reliability in prediction, such as curating diabetic retinopathy for training Bayesian DNNs [6].

To make the model more cautious when it is uncertain, a rejection option allows it to abstain from making a prediction when it is likely to be a mistake. Geifman and El-Yaniv [15] designed a selective classifier that allows users to set a desired risk level. They further proposed a selective network with a shared classifier of dedicated prediction and ambiguity rejection layer [16]. What's more, Geifman et.al [12] developed a selective mechanism by using early snapshots for samples with high confidence in model training.

Besides training classifiers with a rejection option, studies also shed light on post-hoc approaches for failure detection. Setting thresholds based on confidence or uncertainty ranking of samples is widely used to distinguish correct and incorrect predictions, such as AI for breast cancer screening [17] and decision-making models for low-power Internet of Things (IoT) devices [18]. The threshold needs to be tuned as its value trades off the predictor's coverage rate and the performance on accepted examples [8, 7]. In our work, we will provide an insightful reference for such threshold selection.

2.2. Evaluation Metrics

The quantification of failure detection (FD) performance shares the same characteristic as selective prediction (SP). FD focuses on the model performance after rejecting worst predicted samples under coverage, while SP highlights the model accuracy or error with partial input.

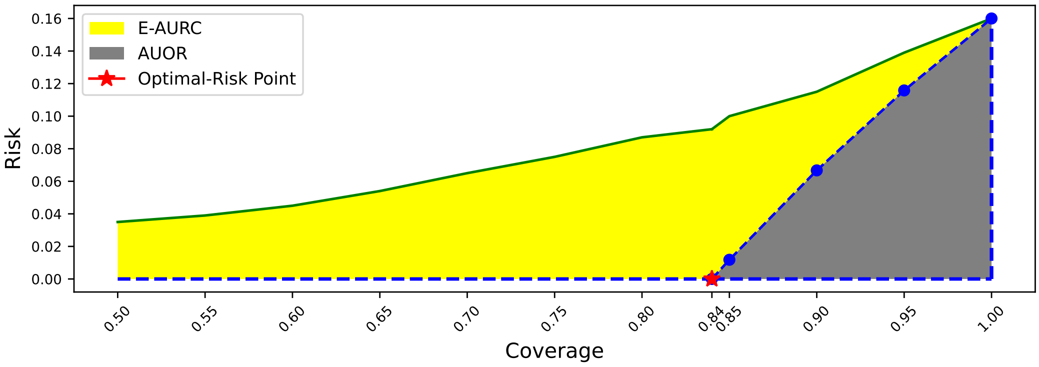

Figure 1: Risk-coverage curve for the ImageNet dataset with SwinTran model. The entire AURC is the yellow plus grey area, with the E-AURC shown as yellow area and the area under the optimal risk (AUOR) as the optimal-risk area. The optimal-risk point is at the coverage of model accuracy (in this case, 0.84).

More broadly, they are techniques for uncertainty estimation [11]. Therefore, the evaluation metrics for SP should also be applicable for FD, such as Area Under the Receiver Operating Characteristic curve (AUROC) [19] and Area Under the Precision-Recall Curve (AUPR) [20]. Despite the wide use of these metrics for such threshold-independent performance evaluation [21, 22, 17], [11] point out that AUROC and AUPR can cause misleading and meaningless results for classification tasks with softmax function. The main reason lies in the assumption that the numbers of correct and wrong predictions are the same. To mitigate this issue, Risk-Coverage (RC) curve is applied for SP in terms of the multi-class classification tasks[12, 11, 15, 23]. Hence, this paper utilises the RC curve for the following experiments and analysis.

2.3. Model Calibration

To measure the performance of calibration methods, the Expected Calibration Error (ECE) [24] was proposed and is widely applied in various tasks, such as image classification [12, 23] and sentiment analysis [25, 26]. ECE splits the data into bins , calculates for each bin the average confidence and average accuracy, and averages over all bins. To alleviate the miscalibration issue for DNNs, calibration techniques have been proposed and then widely applied. Label Smoothing (LS) [27] reduces over-confidence by computing the cross-entropy loss with uniformly squeezed labels instead of one-hot labels. Extensions of LS such as Margin-based Label Smoothing (MBLS) [28] further provides a unifying constrained-optimization perspective of calibration losses. Focal Loss (FL) [29] adds a focusing factor to the standard cross-entropy loss to deal with an imbalanced dataset. Recent work on sample-dependent focal loss (FLSD) [30] investigated the effect of the loss on the training data and achieved impressive performance in calibration. However, it is arguable to what extent calibration techniques can improve the model trustworthiness [23]. Our work will provide a more comprehensive evaluation method regarding this issue.

3. Methodology

The issue we address in this paper is the quantification of the failure detection performance for supervised classification models with the utilization of softmax function. Let X be the input space and $Y = \{1, 2, 3, \ldots, k\}$ be the set of class labels. Given $D(X, Y)$ as the data distribution over $X \times Y$, a classifier is the function f where the error (true risk) err and accuracy acc is obtained by $f : X \rightarrow Y$. For each input $x \in X$ and its corresponding true label y, the probability distribution of the model prediction is $P(y \mid x)$, and the predicted label is $\hat{y} = \text{argmax}_{y \in Y} P(y \mid x)$.

3.1. Problem Setting

In the Risk-Coverage (RC) curve, the coverage c is the percentage of covered set over the entire data, which is written as $c = \frac{|X_c|}{|X|}$. For each coverage, the risk is the corresponding error in model prediction. A model with better FD performance should obtain less risk/ higher accuracy with fewer samples rejected.

To efficiently quantify the FD performance of a model, we first need to construct the reject function $\mathcal{R}$ to decide whether to reject samples or not under different thresholds. By adopting settings in [31, 5, 12], we utilize the predictive uncertainty u to rank samples. A sample with low uncertainty indicates high confidence and better reliability of the model prediction; whereas a sample with

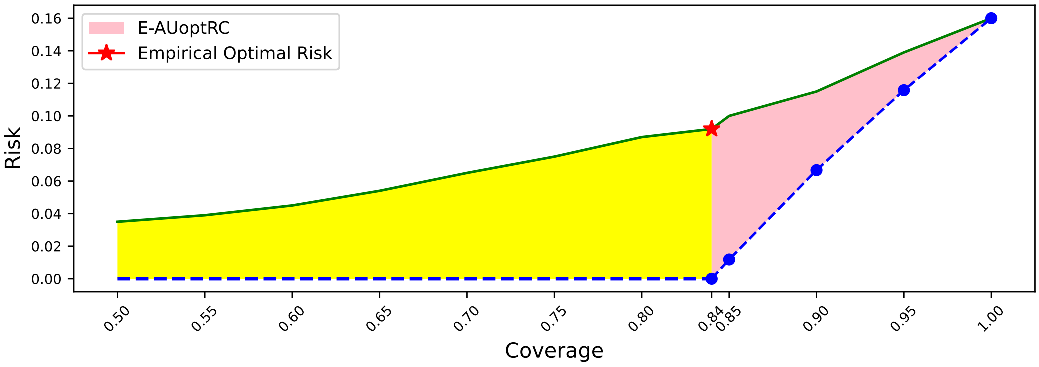

Figure 2: Our proposed method in the RC-curve for ImageNet with SwinTran model. Our proposed E-AUoptRC is shown as the pink area while the E-AURC is the yellow plus pink area. The empirical optimal risk shows the real performance at the optimal point.

high u is more likely to be rejected when narrowing the coverage. Given a fixed or adaptive threshold t, the reject function $\mathcal{R}$ is written as follows:

$$\mathcal{R}(x) = \begin{cases} \text{cover}, x \in X_c, & \text{if } u <= t \\ \text{reject}, x \in X_r, & \text{if } u > t \end{cases} \tag{1}$$

where X_c is the covered input set and X_r is the reject set.

There are two types of risks namely empirical risk and optimal risk [12]. The empirical risk $erisk$ is the predicted error of the model under different coverage, as shown in the solid green line in Figure 1. As the aleatory uncertainty inherits from the data, some risks inevitably exist in certain coverage regardless of the model performance. For a model with perfect uncertainty estimation, if we discard the error percentage of high uncertainty samples, the risk in the remaining coverage input should be zero. This specific coverage point of $1 - err$ (or acc) was proposed by [12] as the optimal point op and shown as the red star in Figure 1. Specifically, the risk between coverage of op to 1 monotonically increases until the error of the model. For optimal calibration, the above risks are called optimal risk $optrisk$ illustrated as the blue dotted line in the figure. For example, the model error in the figure is 0.16 and the op is 0.84. Therefore, the optimal risk $optrisk$ under coverage 0 to op is supposed to be 0; while it increases from 0 to 0.16 under op to full coverage. It is worth-noticing that the monotonic increment of $optrisk$ is not exactly in the linear way.

Both $erisk$ and $optrisk$ can be calculated by Area Under the RC-curve (AURC) [12, 11], named $empAURC$ (yellow plus grey area in Figure 1) and $AUOR$ (grey area in Figure 1) respectively. The difference between $empAURC$ and $AUOR$ is the real FD area, shown as the yellow area in Figure 1. [12] propose this specific area as the Excess-AURC (E-AURC), where E-AURC $=$ $empAURC - AUOR$.

3.2. E-AUoptRC

The E-AURC reveals the total risk in coverage range from 0 to 1. However, in real-world applications, the coverage is mainly customised due to specific deployment requirements, making it challenging to compare the failure detection (FD) performance for various models. In addition, the E-AURC cannot reveal the failure detection (FD) performance in a specific coverage range. To mitigate the above issues, we propose E-AUoptRC with the coverage from op to 1 (E-AUoptRC, shown as pink in Figure 2). We emphasise the E-AUoptRC for the following reasons: (1) it is more practical for deployment, as it is unlikely to discard more than half of data in applications; (2) 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.

3.3. Trust Index

Model accuracy acc should track the confidence of the model prediction. For example, a model with 80% accuracy suggests 80% confidence in its own predictions, which also defines the perfect confidence score in calibration. As the risk at the optimal point (op) is supposed to be 0, the accuracy at op should be 1, indicating the prediction's highest model confidence and trustworthiness. In other words, after removing $err\%$ data with high uncertainty, the correctly predicted samples in the remaining data are most trusted. The accuracy at op also reveals the model calibration, as the discarded $err\%$ data can be misclassified. To represent the model performance in terms of accuracy and calibration, we propose the ac-

Table 1
Main Results of AURC, E-AURC, E-AUoptRC, accuracy(ACC) and trust index (TI) on the ImageNet (IN) and Cifar100 (CF100) dataset with CNNs and variants of transformers models. AURC, E-AURC, fE-AURC and lE-AURC are shown as multiply with 10^3 for clarity.

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

curacy at the *op* as a Trust Index (TI), a complementary evaluation to the accuracy metric to indicate the model's trustworthiness. For example, in Figure 2, with the model accuracy of 84%, the model is 0.84 trust of the prediction. After removing 16% samples with high uncertainty (the *op* is 0.84), the risk is approximately 0.08. The TI , the accuracy over the most confident 84% of samples is 0.92. The higher TI suggests the better trustworthiness of the model predictions, and we next present empirical data to substantiate this.

4. Experimental Setup

4.1. Datasets and Baselines

We validate the proposed method with three benchmark image datasets: ImageNet 2012 (IN) [32], CIFAR100 (C100) [33] and Tiny-ImageNet [34]. For baselines, we use state-of-the-art (SOTA) Vision Transformer (ViT) [35] and its variants such as Swin-Transformer (SwinT) [36], Class-Attention in Image Transformers (CaiT) [37], Cross-Attention Multi-Scale Vision Transformer (CrossViT) [38], ConvNext [39] with the ImageNet pretrained weights from TIMM [1] library. To report comprehensive results on various models architectures, we also use the Convolutional neural networks (CNNs) in our experiments, namely DenseNet121 [40], ResNet56 [41], variants of VGG [42] and MobileNetV2 [43]. All models are with pretrained weights of ImageNet dataset. For recent SOTA calibration techniques label smoothing (LS) [27], focal loss (FL) [29], MBLS [28] and FLSD [30], we utilize the pre-trained model and official implementation from the repository [2].

As the evaluation of failure detection is a post-processing approach, we primarily utilize each dataset's test set. For the ImageNet dataset, we equally divide its original test set of 50,000 images into validation and test sets for a fair comparison. For Tiny-ImageNet and CIFAR100 dataset, an 80/10/10 for training/validation/test split is applied.

4.2. Implementation Details

For a fair comparison and replicability of experimentation, we utilized publicly available existing pre-trained weights for our investigation and experimentation. The GPU of the Nvidia Tesla P40 was used for all experiments. The bins number for ECE was set as $M = 15$.

5. Results

We conducted extensive experiments on benchmark datasets ImageNet and Cifar100 with various CNNs and variants of transformers to compare the AURC, E-AURC and our proposed E-AUoptRC. We further observed the limitation of the conventional overall model accuracy and how our proposed Trust Index (TI) mitigates it. Finally, to validate the efficacy of our method, we applied it to SOTA calibration techniques with Tiny_ImageNet on ResNet50 dataset. All the experiments and results are shown in Tables 1 and 2, and Figure 3.

Table 1 shows the results for image classification with the benchmark datasets. For AURC, E-AURC and E-AUoptRC in the ImageNet dataset, the variants of transformers outperform CNNs model. The E-AURC for ViT is about half of the E-AURC of SwinTran, CaiT and ConvNext, indicating that ViT greatly outperforms the other three models in failure detection. However, regarding the E-AUoptRC, the difference is almost ignorable and the ConvNext is slightly better than the other three models. The risk-coverage (RC) curve (Left in Figure 3) also shows that at the coverage of 0.84 (near the optimal point) to 1,

[1] https://github.com/rwightman/pytorch-image-models
[2] https://github.com/by-liu/MbLS

Table 2

Results for SOTA calibration techniques on failure detection with Tiny_ImageNet dataset with ResNet50 model. AURC, E-AURC, fE-AURC and lE-AURC are shown as multiply with 10^3 for clarity. ECE_OP denots the ECE at the optimal point. ECE and ECE_OP are shown in percentage.

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

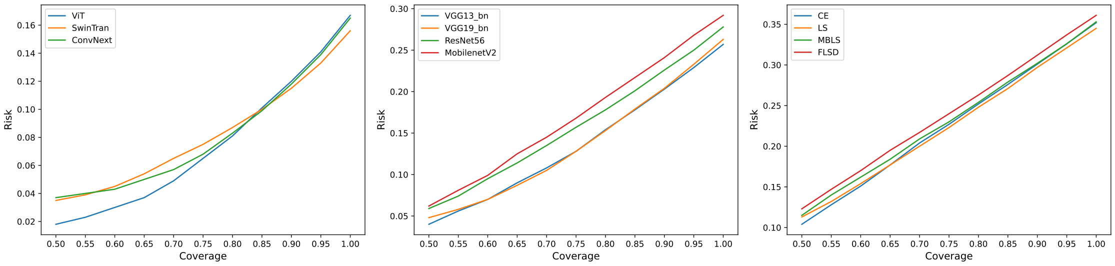

Figure 3: Risk-coverage curve for the visualization of failure detection performance. Left: ImageNet dataset with transformer models of ViT, SwinTran and ConvNext; Middle: Cifar100 dataset with CNNs models of VGG13_bn, VGG19_bn, ResNet56 and MobileNetV2. Right: Tiny_ImageNet with ResNet50 model on SOTA calibration techniques. CE, LS, MBLS, FLSD denots baseline, lable smoothing, margin-based label smoothing and sample-dependent focal loss respectively. The coverage starts from 0.5 instead of 0 for the clarity of visualization.

the risk curve of ViT and ConvNext is nearly overlapping. The lower risk for VIT occurs at very low coverage levels, which are not of interest for most real world applications. For CF100 dataset with CNNs, VGG13_bn substantially outperforms other models in terms of AURC and E-AURC. However, the difference in E-AUoptRC between VGG13_bn and VGG19_bn is much smaller. This can be understood from the Middle plot in Figure 3, where the curve for VGG13_bn and VGG19_bn overlaps at coverage between 0.74(near the optimal point) to 0.9. These differences in the metrics provide empirical evidence that our proposed E-AUoptRC more accurately reflects real differences in failure detection performance than other methods.

Similar to the results of AURC-related evaluation, the variants of transformer models also outperform CNNs in terms of overall model accuracy and trust index (TI). The SwinTran obtains the highest overall model accuracy for the ImageNet dataset, but it does not yield the highest TI. For the Cifar100 dataset, the VGG13_bn achieves the highest overall model accuracy, whereas the VGG19_bn obtains the best TI. It indicates that the model with the highest overall accuracy does not guarantee the highest TI, which shows that our proposed TI is necessary for model trustworthiness evaluation.

In Table 2, the baseline (CE) obtains better AURC and E-AURC, but label smoothing outperforms other methods and CE in terms of overall accuracy (improves by 0.6%) and TI. MBLS nearly halves the overall ECE of baseline and achieves the best ECE at the optimal point. In the Right RC curve in Figure 3, LS is with the lowest risk at the coverage of 0.65 to 1 (the likely operating range when the model is deployed), and our proposed E-AUoptRC and TI metrics are the only ones that capture this. Failure detection performance should be a significant evaluation for calibration techniques, and our methods provide a more insightful view of the model trustworthiness.

6. Discussion & Conclusion

In this paper, we proposed the E-AUoptRC to more precisely quantify the failure detection performance in the key region of interest, and the Trust Index (TI) that measures model accuracy at its optimal point. The empirical results show that our methods can better reveal the model trustworthiness under a fair comparison. In the real-world deployment, a fixed threshold is often used due to specific task requirements and simplicity of implementation. Our proposed TI can be utilized as the reference for the threshold selection with following reasons: (1) the accuracy should indicate the model confidence in its prediction, suggesting the TI can interpret the confidence;

(2) TI is obtained at the optimal point, where the model is supposed to achieve the ideal performance. This is an objective method for the fair comparison of models with different accuracy and calibration (as shown in Table 1 and 2); (3) TI is easy to calculate, which is a time and computational cost saving. We have shown several benefits of our proposed metrics over existing ones and in our future work, we will further investigate the role of TI in improving failure detection.

References

[1] S. Atakishiyev, M. Salameh, H. Yao, R. Goebel, Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions, arXiv preprint arXiv:2112.11561 (2021).

[2] M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, J. Kleinberg, Direct uncertainty prediction for medical second opinions, in: International Conference on Machine Learning, PMLR, 2019, pp. 5281–5290.

[3] M. W. Dusenberry, D. Tran, E. Choi, J. Kemp, J. Nixon, G. Jerfel, K. Heller, A. M. Dai, Analyzing the role of model uncertainty for electronic health records, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 204–213.

[4] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, ICLR (2017).

[5] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez, Addressing failure prediction by learning model confidence, Advances in Neural Information Processing Systems 32 (2019).

[6] N. Band, T. G. Rudner, Q. Feng, A. Filos, Z. Nado, M. W. Dusenberry, G. Jerfel, D. Tran, Y. Gal, Benchmarking bayesian deep learning on diabetic retinopathy detection tasks, in: NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, 2021.

[7] K. Hendrickx, L. Perini, D. Van der Plas, W. Meert, J. Davis, Machine learning with a reject option: A survey, arXiv preprint arXiv:2107.11277 (2021).

[8] R. El-Yaniv, et al., On the foundations of noise-free selective classification., Journal of Machine Learning Research 11 (2010).

[9] C. Ferri, J. Hernández-Orallo, Cautious classifiers., ROCAI 4 (2004) 27–36.

[10] M. S. A. Nadeem, J.-D. Zucker, B. Hanczar, Accuracy-rejection curves (arcs) for comparing classification methods with a reject option, in: Machine Learning in Systems Biology, PMLR, 2009, pp. 65–81.

[11] Y. Ding, J. Liu, J. Xiong, Y. Shi, Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 4–5.

[12] Y. Geifman, G. Uziel, R. El-Yaniv, Bias-reduced uncertainty estimation for deep neural classifiers, in: International Conference on Learning Representations, 2019.

[13] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).

[14] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in ai safety, arXiv preprint arXiv:1606.06565 (2016).

[15] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, Advances in neural information processing systems 30 (2017).

[16] Y. Geifman, R. El-Yaniv, Selectivenet: A deep neural network with an integrated reject option, in: International conference on machine learning, PMLR, 2019, pp. 2151–2159.

[17] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis, The Lancet Digital Health 4 (2022) e507–e519.

[18] C. Cho, W. Choi, T. Kim, Leveraging uncertainties in softmax decision-making models for low-power iot devices, Sensors 20 (2020) 4603.

[19] T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (2006) 861–874.

[20] C. Manning, H. Schutze, Foundations of statistical natural language processing, MIT press, 1999.

[21] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv preprint arXiv:1610.02136 (2016).

[22] A. Malinin, M. Gales, Predictive uncertainty estimation via prior networks, Advances in neural information processing systems 31 (2018).

[23] F. Zhu, Z. Cheng, X.-Y. Zhang, C.-L. Liu, Rethinking confidence calibration for failure prediction, in: European Conference on Computer Vision, Springer, 2022, pp. 518–536.

[24] M. P. Naeini, G. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using bayesian binning, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[25] R. Müller, S. Kornblith, G. E. Hinton, When does label smoothing help?, Advances in neural information processing systems 32 (2019).

[26] S. Obadinma, H. Guo, X. Zhu, Class-wise calibration: A case study on covid-19 hate speech., in: Canadian Conference on AI, 2021.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[28] B. Liu, I. Ben Ayed, A. Galdran, J. Dolz, The devil is in the margin: Margin-based label smoothing for network calibration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 80–88.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[30] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, P. Dokania, Calibrating deep neural networks using focal loss, Advances in Neural Information Processing Systems 33 (2020) 15288–15299.

[31] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems 30 (2017).

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.

[33] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, Technical Report, University of Toronto, 2009.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[37] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, H. Jégou, Going deeper with image transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 32–42.

[38] C.-F. R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 357–366.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.

[40] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.