A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification

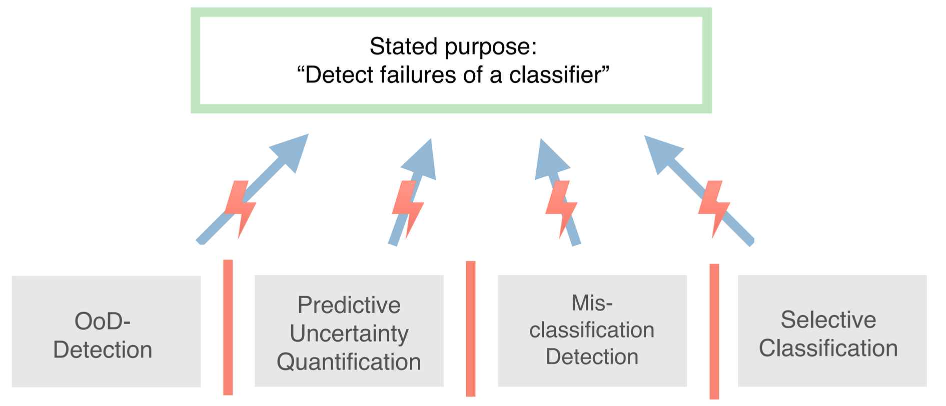
Paul F. Jaeger[1,2], Carsten T. Lueth[1,2], Lukas Klein[1,2,3], Till Bungert[1,2]

[1] Interactive Machine Learning Group, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2] Helmholtz Imaging, DKFZ, Heidelberg, Germany
[3] Institute for Machine Learning, ETH Zürich, Zürich, Switzerland

Paper: Benchmark:

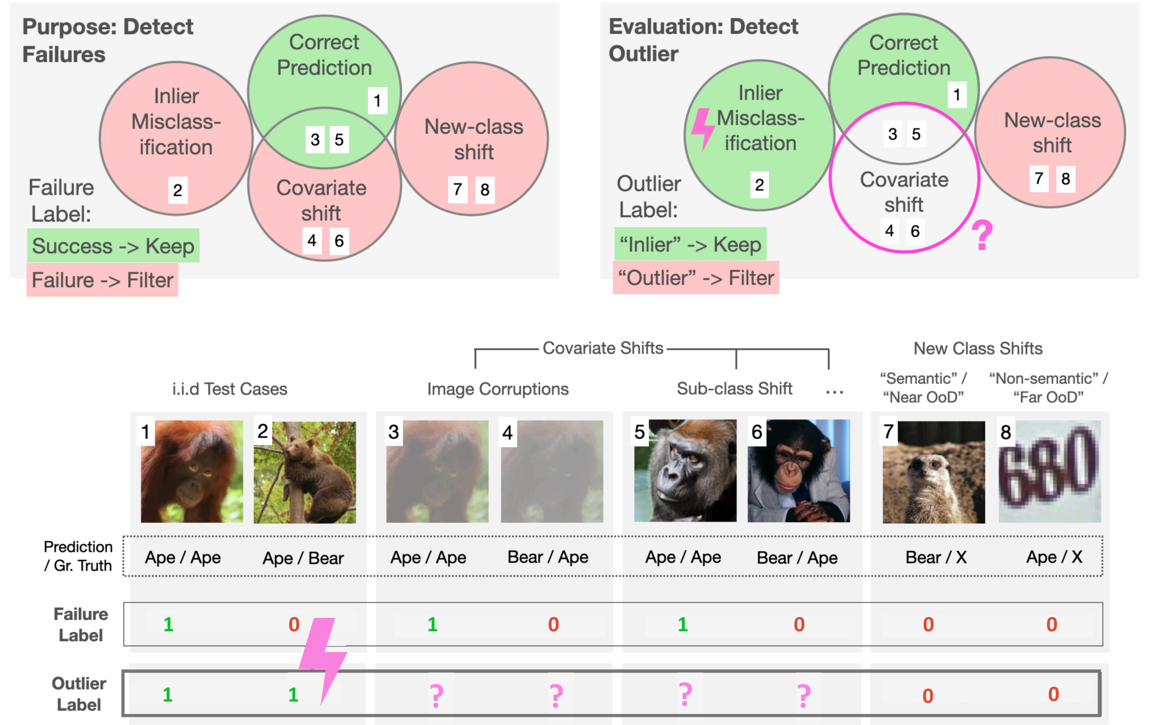
HELMHOLTZ IMAGING · IML Interactive Machine Learning Group at Helmholtz Imaging and the DKFZ · dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Status Quo: Inconsistent landscape of confidence scoring



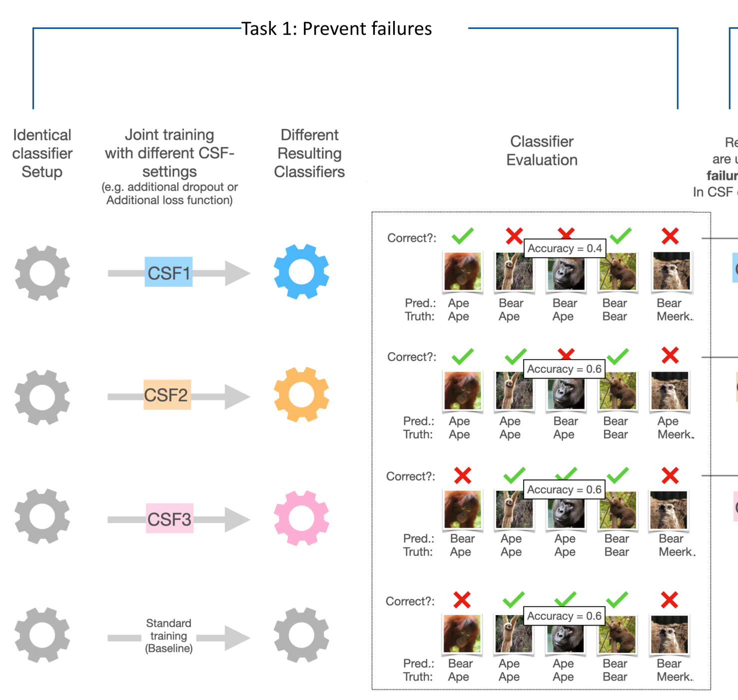
- Motivational statements in many publications of these four research areas (OoD-D, PUQ, MisD, SC) indicate that **all respective studies address the same goal of using confidence scoring functions (CSFs) to detect failures of a classifier**.
- In our work, we make the case that the evaluation in these fields is substantially flawed for two reasons:
  1. There is a **discrepancy between this stated purpose and the way methods are evaluated**.
  2. All these fields are currently **mostly siloed** meaning that there is **no cross-comparison of methods**, although they address the same goal.

Pitfall 1: OoD-Detection often deviates from its stated purpose



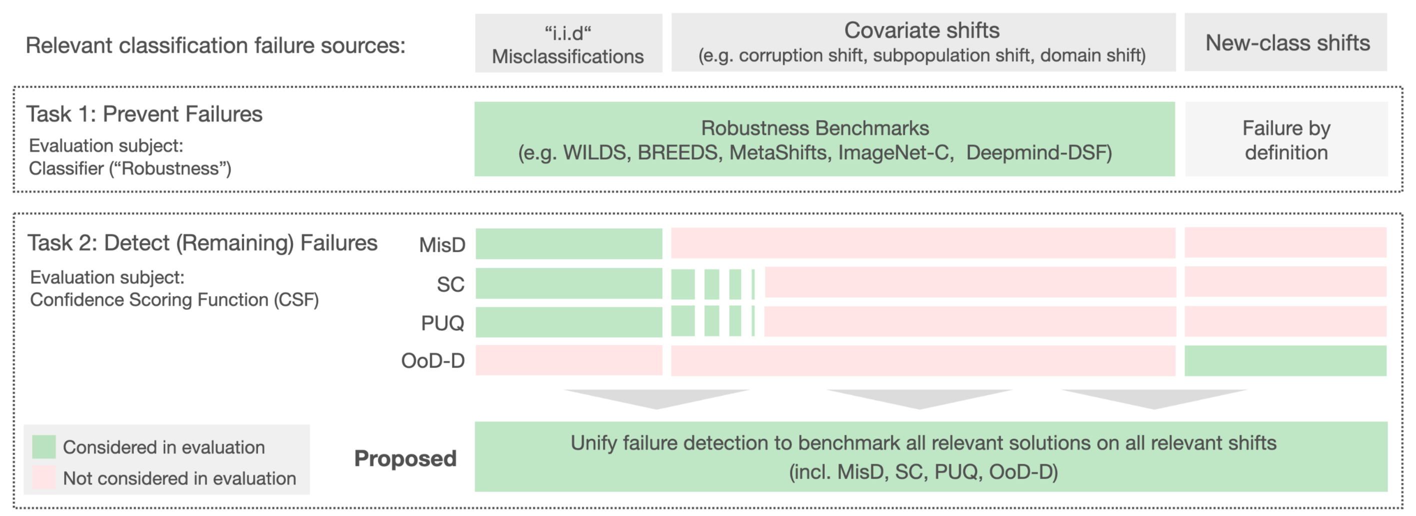
- In Out-of-distribution Detection, indeed, the majority of publications states failure detection in a well-defined classification task as their purpose.
- In the OoD-D task protocol, however, an outlier label is employed instead of the failure status of the classifier, only aiming to determine whether cases are subject to a new-class shift or not.
- Thus, for one, a CSF is **rewarded for giving high confidence to all inliers (purple lightning)**, including failures, but perhaps even more concerning, the subjective outlier label is **not clearly defined on the covariate shifts (purple question marks)**.
- It could be argued that these more subtle shifts where the image label is preserved are the more realistic and thus more relevant ones.

Pitfall 2: Current evaluation metrics lead to biased and incomplete comparison



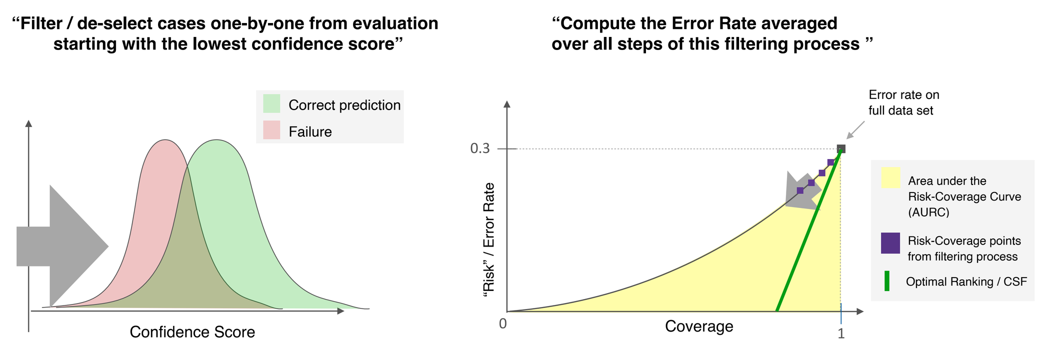
- The protocol in **Misclassification Detection (MisD)**, for instance, employs ranking metrics like AUROC and **excludes a substantial part of relevant CSFs from comparison**
- This is because any CSF that affects the underlying classifier (e.g. by introducing dropout or alternative loss functions) alters the set of classifier failures, i.e. creates their individual set of reference labels, and thus **leads to baised comparison**.
- This indicates that **CSFs should be evaluated as part of a symbiotic system with the associated classifier.**
- On the other hand, predominant metrics in **Predictive Uncertainty Quantification (PUQ)**, like NLL or Brier Score, **exclude all CSFs beyond the classifier output** (e.g. softmax response) by design.
- Further, such proper scoring rules **conflate the task of failure detection with the task of calibration**, which is an orthogonal interest that is not necessarily required.

Pitfall 3: Current benchmarks ignore the major part of relevant failure sources



- For Task 1, we see that there are **many benchmarks to test the "robustness" of a classifier featuring a diverse and nuanced range of shifts**
- In contrast, for Task 2, the current landscape of benchmarks is not only inconsistent and siloed, but also there is a **severe lack of testing CSFs under different distribution shifts**. This discrepancy begs the question: If simulating realistic classification failures is such a delicate effort, why are there no analogous benchmarking efforts in the research on *detecting failures?*
- In our study, we fill this gap and **propose a benchmark that overcomes the stated pitfalls, unifies all previously separated fields, and allows to compare arbitraty CSFs on the entire range of realistic failure sources.**

The Area under the Risk-Coverage Curve overcomes previous evaluation pitfalls



- To overcome the stated evaluation pitfalls **we employ the Area under the Risk-Coverage Curve (AURC)**. This metric has been introduced for selective classification before, but we argue it is able to overcome the described pitfalls in all previously separated fields and **unify them under their stated purpose of failure detection.**
- **AURC takes into account potential effects of a CSF on the classifier**, to isolate differences across classifiers as a nuisance factor from CSF evaluation, it is crucial to ensure identical classifiers before training.
- It can be interpreted as the **error rate averaged over the process of filtering cases ranked from low to high confidence score**.

FD-Shifts: A large-scale benchmark for confidence scoring functions

	study	iWildCam iid	iWildCam sub	iWildCam s-ncs	BREEDS iid	BREEDS sub	CAMELYON iid	CAMELYON sub	CIFAR-100 iid	CIFAR-100 sub	CIFAR-100 cor	CIFAR-100 s-ncs c10	CIFAR-100 ns-ncs svhn	CIFAR-100 ns-ncs ti	CIFAR-10 iid	CIFAR-10 cor	CIFAR-10 s-ncs c100	CIFAR-10 ns-ncs svhn	CIFAR-10 ns-ncs ti	SVHN iid	SVHN s-ncs	SVHN ns-ncs c10	SVHN ns-ncs c100	SVHN ns-ncs ti
	ncs-data set																							
MSR	CNN	62.3	69.5	217	8.19	175	10.1	143	70.2	230	289	312	533	321	5.62	93.0	226	429	198	4.85	132	55.1	55.8	55.7
MLS	CNN	90.2	87.8	240	11.7	188	10.1	143	87.6	233	312	318	544	270	6.38	94.0	221	421	191	5.49	131	52.2	53.0	52.9
PE	CNN	62.7	69.9	215	8.21	174	10.1	143	77.0	234	308	319	587	289	5.59	92.1	225	427	196	4.85	132	54.3	55.0	54.9
MCD-MSR	CNN	52.6	65.9	180	7.79	173	6.93	151	67.0	217	269	311	546	318	5.23	81.4	226	438	203	4.75	164	54.1	54.9	54.0
MCD-MLS	CNN	84.1	78.6	189	11.5	186	7.22	156	79.0	220	279	315	528	298	6.25	84.0	221	427	193	5.37	161	51.7	52.4	52.1
MCD-PE	CNN	53.5	67.1	175	7.89	173	6.93	151	72.4	217	273	312	540	306	5.43	81.3	224	434	199	4.82	163	52.8	53.5	52.7
MCD-EE	CNN	54.1	67.6	175	7.94	173	6.94	151	72.5	218	274	313	544	308	5.49	80.6	223	429	196	4.85	162	52.5	53.3	52.8
MCD-MI	CNN	53.5	71.0	218	8.19	176	7.40	164	76.9	221	282	313	535	302	5.71	88.2	227	453	212	4.91	165	54.5	55.3	53.7
ConfidNet	CNN	143	144	214	8.30	176	5.04	132	72.7	232	290	321	552	285	5.32	88.8	224	427	197	4.81	132	55.4	56.0	55.8
DG-MCD-MSR	CNN	54.8	72.7	224	7.05	167	4.30	273	66.3	216	268	311	547	327	5.29	82.3	230	443	209	4.63	118	54.4	55.1	54.4
DG-Res	CNN	101	88.8	246	11.0	185	4.08	218	89.8	376	311	325	520	291	5.39	88.8	243	422	195	5.68	124	52.4	53.1	53.4
Devries et al.	CNN	95.8	100	234	9.71	179	33.8	282	91.3	241	327	332	617	338	5.22	97.5	226	420	195	7.27	154	55.7	57.7	55.4
MSR	ViT	86.1	70.6	267	1.89	105	0.09	62.5	14.2	70.0	69.0	238	456	255	0.95	9.77	192	404	208	4.02	147	52.7	53.2	52.0
MLS	ViT	111	88.4	271	3.49	120	0.13	61.0	23.5	86.1	79.0	226	452	253	1.27	10.4	192	401	205	4.91	146	51.2	52.0	50.2
PE	ViT	85.8	70.5	266	1.89	105	0.09	62.5	14.2	70.0	68.6	236	455	254	0.95	9.75	192	404	208	4.02	147	52.5	53.1	51.8
MCD-MSR	ViT	113	145	163	1.54	108	0.05	304	13.7	76.3	91.1	240	490	232	0.86	18.2	196	410	198	3.70	136	52.4	52.7	51.3
MCD-MLS	ViT	143	156	173	2.61	123	0.10	350	24.1	86.2	101	227	468	225	1.01	17.3	190	400	190	4.47	136	50.4	50.8	49.9
MCD-PE	ViT	112	143	161	1.55	108	0.05	304	13.8	76.0	90.4	238	487	230	0.86	18.1	195	409	198	3.69	136	52.0	52.4	50.9
MCD-EE	ViT	113	144	161	1.55	108	0.05	304	13.8	77.2	89.9	236	485	228	0.86	17.7	194	407	197	3.69	135	51.7	51.9	50.6
MCD-MI	ViT	112	145	161	1.55	108	0.06	316	13.9	76.0	93.3	244	494	234	0.87	18.9	197	411	200	3.72	136	52.5	53.0	51.4
MAHA	ViT	188	170	360	4.71	159	1.23	95.1	21.4	158	83.4	219	470	222	1.48	16.0	185	410	185	5.21	124	48.8	49.2	50.0

FD-Shifts overcomes the stated pitfalls, unifies all previously separated fields, and for the first time, compares arbitraty CSFs on the entire range of realistic failure sources.

The table shows results measured as AURC $*$ 1000 (score range:[0, 1000], lower is better ↓). The color heatmap is normalized per column and classifier (separately for CNN and ViT), while whiter colors depict better scores. "cor" is the average over 5 intensity levels of image corruption shifts. AURC scores are averaged over 5 runs on all data sets with exceptions for the CNN: 10 runs on CAMELYON-17-Wilds (due to high volatility in results) and 2 runs on BREEDS. Abbreviations: ncs: new-class shift (s for semantic, ns for non-semantic), iid: independent and identically distributed, sub: sub-class shift, cor: image corruptions, c10/100: CIFAR-10/100, ti: TinyImagenet

FD-Shifts insights open up new research directions

1. "None of the evaluated methods from literature beats the simple Maximum Softmax Response baseline across a realistic range of failure sources."
2. "Prevalent OoD-D methods are only relevant in a narrow range of distribution shifts."
3. "AURC is able to resolve previous obscurities between classifier robustness and CSF performance."
4. "CNN beats ViT on the iWildCam benchmark, indicating interesting transfer-learning issues."
5. "Different types of uncertainty are empirically not distinguishable."
6. "CSFs beyond Maximum Softmax Response yield well-calibrated scores."
7. "The Maximum Softmax Response baseline is disadvantaged by numerical errors in the standard setting."

Great demand for next generation of robust CSFs!

Deeper understanding of uncertainty modeling in practice required!

Research perspective: Calibrated confidence beyond Softmax Response

Easter Egg finding: The Softmax baseline is often disadvantaged by numerical errors in ranking metrics like AUROC

		Round-to-one error rate $*100$ ↓ 16bit	32bit	64bit	AURC $*10^3$ ↓ 16bit	32bit	64bit	$\mathrm{AUROC}_f * 100$ ↑ 16bit	32bit	64bit	Accuracy $*100$ ↑
CNN	iWildCam	47.30	1.802	0.000	82.22	69.20	69.00	85.80	87.50	87.50	76.01
	BREEDS	35.25	4.268	0.003	18.81	12.89	12.84	89.89	92.22	92.22	90.72
	CAMELYON	5.365	0.001	0.000	10.25	10.12	10.12	89.18	89.21	89.21	93.99
	CIFAR-100	22.75	2.264	0.001	87.22	77.43	77.39	86.38	87.29	87.29	73.26
	CIFAR-10	41.54	1.465	0.000	8.346	5.620	5.617	91.98	93.73	93.73	94.35
	SVHN	41.76	17.29	0.001	8.074	4.902	4.850	89.59	92.81	92.87	96.09
ViT	iWildCam	44.41	14.91	0.000	221.6	177.8	177.0	75.97	80.35	80.38	62.12
	BREEDS	80.19	52.59	0.423	11.43	4.559	1.893	72.65	88.88	94.35	97.92
	CAMELYON	82.03	14.52	0.000	4.661	1.007	1.007	88.59	96.42	96.42	97.95
	CIFAR-100	68.65	30.27	0.000	36.27	14.95	14.23	79.29	90.10	90.29	91.62
	CIFAR-10	92.16	81.79	1.883	7.614	3.480	0.950	69.30	85.85	94.90	98.76
	SVHN	69.02	47.17	0.305	16.94	8.757	5.475	68.75	83.55	88.14	97.30

- Depending on floating point precision, rounding errors occur during the softmax operation thereby losing the ranking information between rounded scores.
- Especially on the ViT classiifer, these errors occur at astounding rates leading to substantial ranking performance drops as measured e.g. by AUROC.
- **Even at default 32-bit precision, this effect leads to a substantial disadvantage of softmax baselines in all ranking tasks including OoD-Detection.**

Hands-on recommendations for evaluating confidence scoring

1. **State a clear purpose of the confidence scoring function (CSF) and design an evaluation protocol that reflects this purpose.**
2. **If the purpose is failure detection, we recommend AURC as primary metric for method comparison.**
3. **Analogously to classifier robustness, CSFs need to be tested on a wide range of data sets and distribution shifts.**
4. **Compare against all viable solutions addressing the same goal, even if from seemingly separated fields.**
5. **Logits should be cast to 64-bit precision or temperature-scaled prior to the softmax operation for any ranking-related tasks to avoid subpar softmax baselines.**

HELMHOLTZ IMAGING

p.jaeger@dkfz.de @pfjaeger pfjaeger.github.io