Understanding Silent Failures in Medical Image Classification

Till J. Bungert [1 2], Levin Kobelke [1 2], Paul F. Jäger [1 2]
[1] Helmholtz Imaging, [2] Interactive Machine Learning Group, DKFZ

An ever-growing number of ML systems is set up for clinical usage and uncertainty methods are heavily studied to make them reliable. But what if both, classifier and failure detection fail, creating a silent failure?

Contribution 1: Failure Detection Benchmark

Comprehensive benchmark of silent failure prevention in the biomedical field, comparing various confidence scoring functions (CSF) under a **wide range of distribution shifts** on four biomedical datasets. (Fig. 1)

Dataset	Chest X-ray			Dermoscopy				FC-Microscopy			Lung Nodule CT		
Study	iid	cor	acq	iid	cor	acq	man	iid	cor	acq	iid	cor	man
MSR	15.3	18.6	23.1	0.544	0.913	0.799	49.3	13.3	55.6	32.4	6.69	8.18	12.1
PE	15.5	18.9	23.6	0.544	0.913	0.799	49.3	14.1	56.3	32.7	6.69	8.18	12.1
MCD-MSR	14.9	17.9	22.1	0.544	0.913	0.799	49.3	12.6	56.5	31.8	5.80	7.13	11.5
MCD-PE	15.1	18.2	22.7	0.544	0.913	0.799	49.3	13.2	57.2	32.1	5.80	7.13	11.5
MCD-EE	15.1	18.2	22.7	0.544	0.913	0.799	49.3	13.3	57.2	32.1	5.68	7.16	11.9
ConfidNet	15.1	18.5	22.8	0.581	0.979	0.806	51.1	21.9	63.7	61.9	5.77	7.50	15.7
DG-MCD-MSR	14.4	19.0	24.4	0.611	0.893	0.787	50.1	7.46	54.3	33.2	3.97	9.04	12.9
DG-RES	19.4	26.5	32.8	0.814	1.46	1.32	46.8	10.6	55.0	38.1	4.94	8.95	15.0
Devries et al.	14.7	18.4	23.5	0.801	1.08	0.882	45.5	12.9	62.3	51.4	4.99	9.41	20.2

Fig. 1: Benchmark Results (AURC [%], lower is better). "cor": average over all corruption types and intensities "acq"/ "man": averages over all acquisition/manifestation shifts per dataset. "iid": without distribution shifts.

Insights based on Benchmark

- None of the evaluated methods from the literature beats the Maximum Softmax Response (MSR) baseline across a realistic range of failure sources.
- MCD and loss attenuation are able to improve the MSR.
- Effects of particular shifts on the reliability of a CSF might be interdependent.
- **Current systems are not generally reliable enough for clinical application.**

Silent failures are a **significant bottleneck** in the clinical translation of ML systems and require further attention in the medical community.

Contribution 2: SF-Visuals, a tool for visual analysis of silent failures

SF-Visuals, a visualization tool that facilitates **identifying silent failures** in a dataset and **investigating their causes**, generating a deeper understanding of the root causes in the data itself. (Fig. 2)

Insights based on Visual Analysis

- **SF-Visuals enables comprehensive analysis of silent failures.**
  Fig. 2a shows a silent failure. Fig. 2b shows how the target data distribution differs from the inlier distribution, moving malignant target lesions into the benign cluster, causing silent failures. Fig. 2c visually confirms that some of the malignant target lesions are similar to benign inlier lesions, comparison with Fig. 2a shows that this explains the silent failure.
- **SF-Visuals generates insights across tasks and distribution shifts. (Fig. 3)**
  Fig. 3b and 3d show for the Lung CT data and the dermoscopy data, how corruptions can lead to silent failures in low-confident predictions. The brightening of the image leads to a malignant lesion taking on benign characteristics (brighter and smoother skin on the dermoscopy data, decreased contrast between lesion and background on the Lung CT data).

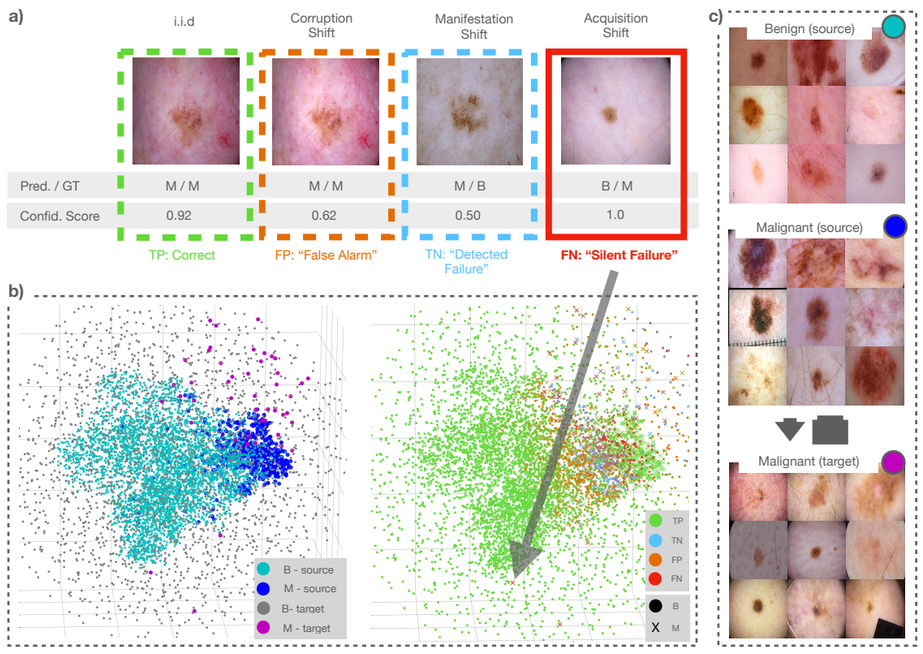

Fig. 2: a) Exemplary Predictions of the Classifier and Confidence Scores **b)** Classifier Latent Space **c)** Concept Cluster Plots. B: Benign, M: Malignant, Pred.: Prediction, GT: Ground truth, Confid.: Confidence Score, Source: Source domain, Target: Target domain.

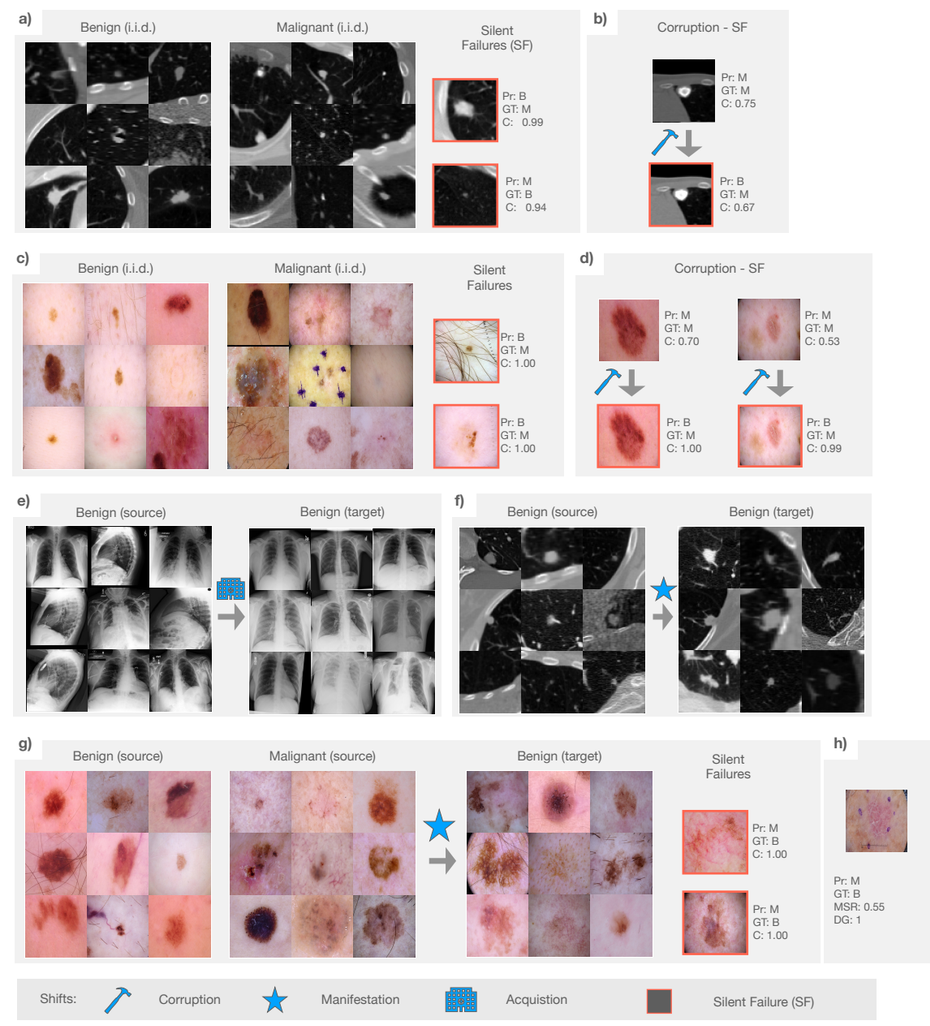

Fig. 3: Various Examples of how the SF-Visuals tool fosters a deeper understanding of root causes of silent failures. i.i.d: Independent and identically distributed, Pr.: Prediction. GT: Ground Truth, C: Confidence Score, Source: Source domain, Target: Target domain.

HELMHOLTZ IMAGING

GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Paper https://tbung.de/sfpaper

Code https://tbung.de/sfcode