Understanding Silent Failures in Medical Image Classification

Till J. Bungert[1,2(✉)], Levin Kobelke[1,2], and Paul F. Jäger[1,2]

[1] Interactive Machine Learning Group, German Cancer Research Center (DKFZ), Heidelberg, Germany
till.bungert@dkfz-heidelberg.de
[2] DKFZ, Helmholtz Imaging, Heidelberg, Germany

Abstract. To ensure the reliable use of classification systems in medical applications, it is crucial to prevent silent failures. This can be achieved by either designing classifiers that are robust enough to avoid failures in the first place, or by detecting remaining failures using confidence scoring functions (CSFs). A predominant source of failures in image classification is distribution shifts between training data and deployment data. To understand the current state of silent failure prevention in medical imaging, we conduct the first comprehensive analysis comparing various CSFs in four biomedical tasks and a diverse range of distribution shifts. Based on the result that none of the benchmarked CSFs can reliably prevent silent failures, we conclude that a deeper understanding of the root causes of failures in the data is required. To facilitate this, we introduce SF-Visuals, an interactive analysis tool that uses latent space clustering to visualize shifts and failures. On the basis of various examples, we demonstrate how this tool can help researchers gain insight into the requirements for safe application of classification systems in the medical domain. The open-source benchmark and tool are at: https://github.com/IML-DKFZ/sf-visuals.

Keywords: Failure detection · Distribution shifts · Benchmark

1 Introduction

Although machine learning-based classification systems have achieved significant breakthroughs in various research and practical areas, their clinical application is still lacking. A primary reason is the lack of reliability, i.e. failure cases produced by the system, which predominantly occur when deployment data differs from the data it was trained on, a phenomenon known as *distribution shifts*. In medical applications, these shifts can be caused by image corruption ("corruption shift"), unseen variants of pathologies ("manifestation shift"), or deployment in new clinical sites with different scanners and protocols ("acquisition shift") [4]. The *robustness* of a classifier, i.e. its ability to generalize across these shifts, is

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_39.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 400–410, 2023.
https://doi.org/10.1007/978-3-031-43898-1_39

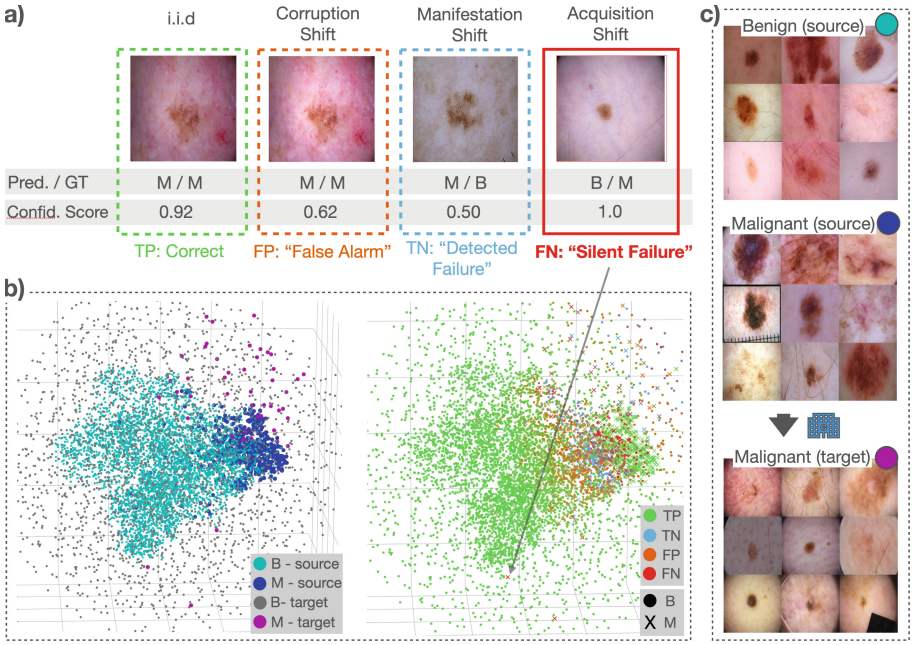

Fig. 1. a) Exemplary predictions of the classifier and the accompanying confidence scoring function (CSF, here: ConfidNet) on the dermoscopy dataset across several distribution shifts. **Note that True/False Positives/Negatives (T/F P/N) do not refer to the classifier decision, but to the failure detection outcome, i.e. the assessment of the CSF.** In this context, FN, i.e. cases with incorrect predictions ("failure") and a high confidence score ("failure not detected") are referred to as *silent failures*. **b)** SF-Visuals allows to identify and analyze silent failures in a dataset based on an Interactive Scatter Plot in the classifier's latent space (each dot represents one image, which is displayed when selecting the dot). **c)** SF-Visuals further features Concept Cluster Plots to gain an intuition of how the model perceives distinct classes or distribution shifts. More details on the displayed example are in Sect. 4.2. Abbreviations: B: Benign, M: Malignant, Pred.: Prediction, GT: Ground truth, Confid.: Confidence, Source: Source domain, Target: Target domain.

extensively studied in the computer vision community with a variety of recent benchmarks covering nuanced realistic distribution shifts [13,15,19,27], and is also studied in isolated cases in the biomedical community [2,3,33]. Despite these efforts, perfect classifiers are not to be expected, thus a second mitigation strategy is to detect and defer the remaining failures, thus *preventing failures to be silent*. This is done by means of confidence scoring functions (CSF) of different types as studied in the fields of misclassification detection (MisD) [5,11,23], Out-of-Distribution detection (OoD-D) [6,7,11,20,21,32], selective classification (SC) [9,10,22], and predictive uncertainty quantification (PUQ) [18,25].

We argue, that silent failures, which occur when test cases break both the classifier and the CSF, are a significant bottleneck in the clinical translation of ML systems and require further attention in the medical community.

Note that the task of silent failure prevention is orthogonal to calibration, as, for example, a perfectly calibrated classifier can still yield substantial amounts of silent failures and vice versa [15].

Bernhardt et al. [3] studied failure detection on several biomedical datasets, but only assessed the performance of CSFs in isolation without considering the classifier's ability to prevent failures. Moreover, their study did not include distribution shifts thus lacking a wide range of realistic failure sources. Jaeger et al. [15], on the other hand, recently discussed various shortcomings in current research on silent failures including the common lack of distribution shifts and the lack of assessing the classifier and CSF as a joint system. However, their study did not cover tasks from the biomedical domain.

In this work, our contribution is twofold: **1)** Building on the work of Jaeger et al. [15], we present the first comprehensive study of silent failure prevention in the biomedical field. We compare various CSFs under a wide range of distribution shifts on four biomedical datasets. Our study provides valuable insights and the underlying framework is made openly available to catalyze future research in the community. **2)** Since the benchmark reveals that none of the predominant CSFs can reliably prevent silent failures in biomedical tasks, we argue that a deeper understanding of the root causes in the data itself is required. To this end, we present SF-Visuals, a visualization tool that facilitates identifying silent failures in a dataset and investigating their causes (see Fig. 1). Our approach contributes to recent research on visual analysis of failures [13], which has not focused on silent failures and distribution shifts before.

2 Methods

Benchmark for Silent Failure Prevention under Distribution Shifts. We follow the spirit of recent robustness benchmarks, where existing datasets have been enhanced by various distribution shifts to evaluate methods under a wide range of failure sources and thus simulate real-world application [19,27]. To our knowledge, no such comprehensive benchmark currently exists in the biomedical domain. Specifically, we introduce corruptions of various intensity levels to the images in four datasets in the form of brightness, motion blur, elastic transformations and Gaussian noise. We further simulate acquisition shifts and manifestation shifts by splitting the data into "source domain" (development data) and "target domain" (deployment data) according to sub-class information from the meta-data such as lesion subtypes or clinical sites. **Dermoscopy dataset:** We combine data from ISIC 2020 [26], derma 7 point [17], PH2 [24] and HAM10000 [30] and map all lesion sub-types to the super-classes "benign" or "malignant". We emulate two acquisition shifts by defining either images from the Memorial Sloan Kettering Cancer Center (MSKCC) or Hospital Clinic Barcelona (HCB) as the target domain and the remaining images as the source

domain. Further, a manifestation shift is designed by defining the lesion subtypes "keratosis-like" (benign) and "actinic keratosis" (malignant) as the target domain. **Chest X-ray dataset:** We pool the data from CheXpert [14], NIH14 [31] and MIMIC [16], while only retaining the classes common to all three. Next, we emulate two acquisition shifts by defining either the NIH14 or the CheXpert data as the target domain. **FC-Microscopy dataset:** The RxRx1 dataset [28] represents the fluorescence cell microscopy domain. Since the images were acquired in 51 deviating acquisition steps, we define 10 of these batches as target-domain to emulate an acquisition shift. **Lung Nodule CT dataset:** We create a simple 2D binary nodule classification task based on the 3D LIDC-IDRI data [1] by selecting the slice with the largest annotation per nodule ($\pm$two slices resulting in 5 slices per nodule). Average malignancy ratings (four raters per nodule, scores between 1 and 5) > 2 are considered malignant and all others as benign. We emulate two manifestation shifts by defining nodules with high spiculation (rating > 2), and low texture (rating < 3) as target domains.

The datasets consist only of publicly available data, our benchmark provides scripts to automatically generate the combined datasets and distribution shifts.

The SF-Visuals Tool: Visualizing Silent Failures. The proposed tool is based on three simple operations, that enable effective and intuitive analysis of silent failures in datasets across various CSFs: *1) Interactive Scatter Plots:* See example in Fig. 1b. We first reduce the dimensionality of the classifier's latent space to 50 using principal component analysis and use t-SNE to obtain the final 3-dimensional embedding. Interactive functionality includes coloring dots via pre-defined schemes such as classes, distribution shifts, classifier confusion matrix, or CSF confusion matrix. The associated images are displayed upon selection of a dot to establish a direct visual link between input space and embedding. *2) Concept Cluster Plots:* See examples in Fig. 1c. To abstract away from individual points in the scatter plot, concepts of interest, such as classes or distribution shifts can be defined and visualized to identify conceptual commonalities and differences in the data as perceived by the model. Therefore, k-means clustering is applied to the 3-dimensional embedding. Nine clusters are identified per concept and the resulting plots show the closest-to-center image per cluster as a visual representation of the concept. *3) Silent Failure Visualization:* See examples in Fig. 2. We sort all failures by the classifier confidence and by default show the images associated with the top-two most confident failures. For corruption shifts, we further allow investigating the predictions on a fixed input image over varying intensity levels.

Based on these visualizations, the functionality of SF-Visuals is three-fold: 1) Visual analysis of the dataset including distribution shifts. 2) Visual analysis of the general behavior of various CSFs on a given task 3) Visual analysis of individual silent failures in the dataset for various CSFs.

3 Experimental Setup

Evaluating Silent Failure Prevention: We follow Jaeger et al. [15] in evaluating silent failure prevention as a joint task of the classifier and the CSF. The area

under the risk-coverage curve AURC reflects this task, since it considers both the classifier's accuracy as well as the CSF's ability to detect failures by assigning low confidence scores. Thus, it can be interpreted as a *silent failure rate* or the error rate averaged over steps of filtering cases one by one according to their rank of confidence score (low to high). Exemplary risk-coverage curves are shown in Appendix Fig. 3. **Compared Confidence Scoring Functions:** We compare the following CSFs: The maximum softmax response (MSR) and the predictive entropy computed from the classifier's softmax output, three predictive uncertainty measures based on Monte-Carlo Dropout (MCD) [8], namely mean softmax (MCD-MSR), predictive entropy (MCD-PE) and expected entropy (MCD-EE), ConfidNet [5], which is trained as an extension to the classifier, DeepGamblers (DG) that learns a confidence like reservation score (DG-Res) [22] and the work of DeVries et al. [6]. **Training Settings:** On each dataset, we employ the classifier behind the respective leading results in literature: For chest X-ray data we use DenseNet121 [12], for dermoscopy data we use EfficientNet-B4 [29] and for fluorescence cell microscopy and lung nodule CT data we us DenseNet161 [12]. We select the initial learning rate between 10^{-3} and 10^{-5} and weight decay between 0 and 10^{-5} via grid search and optimize for validation accuracy. All models were trained with dropout. All hyperparameters can be found in Appendix Table 3.

4 Results

4.1 Silent Failure Prevention Benchmark

Table 1 shows the results of our benchmark for silent failure prevention in the biomedical domain and provides the first overview of the current state of the reliability of classification systems in high-stake biomedical applications.

None of the Evaluated Methods from the Literature Beats the Maximum Softmax Response Baseline Across a Realistic Range of Failure Sources. This result is generally consistent with previous findings in Bernhard et al. [3] and Jaeger et al. [15], but is shown for the first time for a diverse range of realistic biomedical failure sources. Previously proposed methods do not outperform MSR baselines even in the settings they have been proposed for, e.g. Devries et al. under distribution shifts, or ConfidNet and DG-RES for i.i.d. testing.

MCD and Loss Attenuation are Able to Improve the MSR. MCD-MSR is the overall best performing method indicating that MCD generally improves the confidence scoring ability of softmax outputs on these tasks. Interestingly, the DG loss attenuation applied to MCD-MSR, DG-MCD-MSR, which has not been part of the original DG publication but was first tested in Jaeger et al. [15], shows the best results on i.i.d. testing on 3 out of 4 tasks. However, the method is not reliable across all settings, falling short on manifestation shifts and corruptions on the lung nodule CT dataset.

Table 1. Silent failure prevention benchmark results measured in AURC[%] **(score range: [0, 100], lower is better).** The coloring is normalized by column, while lighter colors depict better scores. All values denote an average of three runs. "cor" denotes the average over all corruption types and intensities levels. Similarly, "acq"/"man" denote averages over all acquisition/manifestation shifts per dataset. "iid" denotes scenarios without distribution shifts. Results with further metrics are reported in Appendix Table 2

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

Effects of Particular Shifts on the Reliability of a CSF Might Be Interdependent. When looking beyond the averages displayed in Table 1 and analyzing the results of individual clinical centers, corruptions and manifestation shifts, one remarkable pattern can be observed: In various cases, the same CSF showed opposing behavior between two variants of the same shift on the same dataset. For instance, Devries et al. outperforms all other CSFs for one clinical site (MSKCC) as target domain, but falls short on the other one (HCB). On the Chest X-ray dataset, MCD worsens the performance for darkening corruptions across all CSFs and intensity levels, whereas the opposite is observed for brightening corruptions. Further, on the lung nodule CT dataset, DG-MCD-RES performs best on bright/dark corruptions and the spiculation manifestation shift, but worst on noise corruption and falls behind on the texture manifestation shift. These observations indicate trade-offs, where, within one distribution shift, reliability against one domain might induce susceptibility to other domains.

Current Systems are Not Generally Reliable Enough for Clinical Application. Although CSFs can mitigate the rate of silent failures (see Appendix Fig. 3), the reliability of the resulting classification systems is not sufficient for high-stake applications in the biomedical domain, with substantial rates of silent failure in three out of four tasks. Therefore, a deeper understanding of the root causes of these failures is needed.

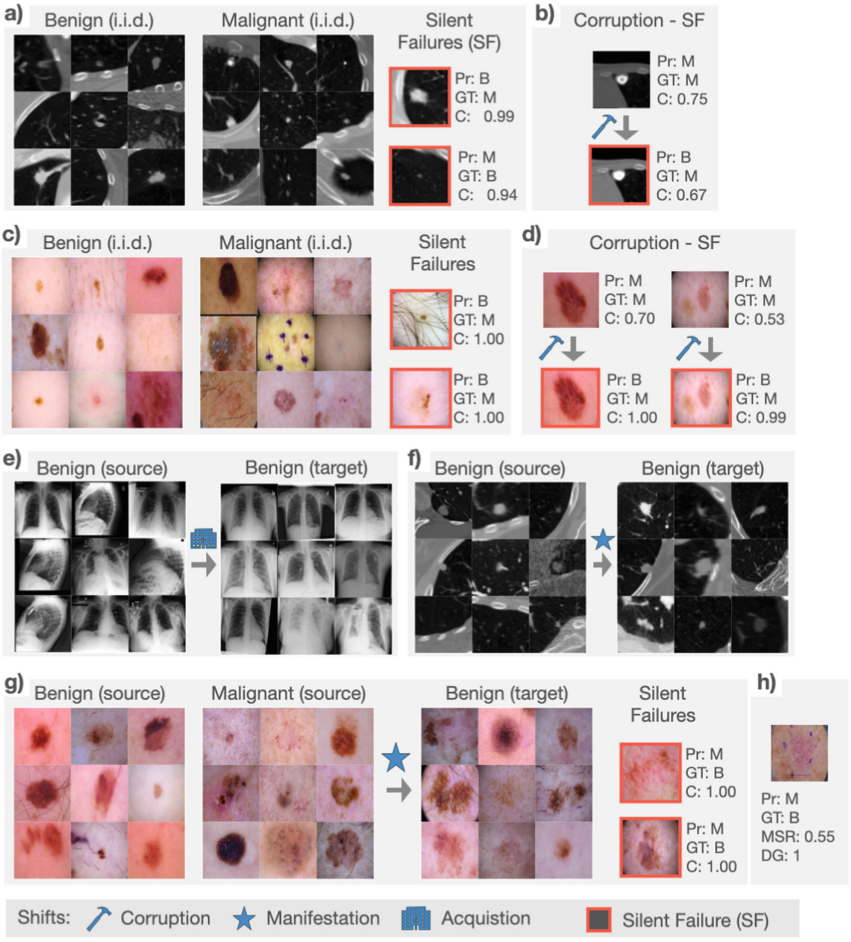

Fig. 2. Various Examples of how the SF-Visuals tool fosters a deeper understanding of root causes of silent failures. Abbreviations: i.i.d: Independent and identically distributed, Pr.: Prediction. GT: Ground Truth, C: Confidence Score, Source: Source domain, Target: Target domain.

4.2 Investigation of Silent Failure Sources

SF-Visuals Enables Comprehensive Analysis of Silent Failures. Figure 1 vividly demonstrates the added benefit of the proposed tool. First, an Interactive Scatter Plot (Fig. 1b, left) provides an overview of the MSKCC acquisition shift on the dermoscopy dataset and reveals a severe change of the data distribution. For instance, some malignant lesions of the target domain (purple dots) are located deep within the "benign" cluster. Figure 1c provides a Concept Cluster Plot that visually confirms how some of these lesions (purple dot) share characteristics of the benign cluster of the source domain (turquoise dot), such

as being smaller, brighter, and rounder compared to malignant source-lesions (blue dot). The right-hand plot of Fig. 1b reveals that these cases have in fact caused silent failures (red crosses) and visual inspection (see arrow and Fig. 1a) confirms the hypothesis that these failures have been caused by the fact that the acquisition shift introduced malignant target-lesions that exhibit benign characteristics. Figure 1b (right) further provides insights about the general behavior of the CSF: Silent failures occur for both classes and are either located at the cluster border (i.e. decision boundary), deeper inside the opposing cluster center (severe class confusions), or represent outliers. Most silent failures occur at the boundary, where the CSF should reflect class ambiguities by low scores, hinting at general misbehavior or overconfidence in this area. Further towards the cluster boundary, the ambiguity in images seems to increase, as the CSF is able to detect the failures (light blue layer of dots). A layer of "false alarms" follows (brown colored dots), where decisions are correct, but confidence is still low.

SF-Visuals Generates Insights Across Tasks and Distribution Shifts. i.i.d. (No Shift): This analysis reveals how simple class clustering (no distribution shifts involved) can help to gain intuition on the most severe silent failures (examples selected as the two highest-confidence failures). On the lung nodule CT data (Fig. 2a), we see how the classifier and CSF break down when a malignant sample (typically: small bright, round) exhibits characteristics typical to benign lesions (larger, less cohesive contour, darker) and vice versa. This pattern of contrary class characteristics is also observed on the dermoscopy dataset (2c). The failure example at the top is particularly severe, and localization in the scatter plot reveals a position deep inside the 'benign' cluster indicating either a severe sampling error in the dataset (e.g. underrepresented lesion subtype) or simply a wrong label. **Corruption shift:** Figs. 2b and 2d show for the Lung Nodule CT data and the dermoscopy data, respectively, how corruptions can lead to silent failures in low-confident predictions. In both examples, 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 Nodule CT data). **Acquisition shift:** Additionally to the example in Fig. 1, Fig. 2e shows how the proposed tool visualizes an acquisition shift on the chest X-ray data. While this reveals an increased blurriness in the target domain, it is difficult to derive further insights involving specific pathologies without a clinical expert. Figure 2h shows a classification failure from a target clinical center together with the model's confidence as measured by MSR and DG. While MSR assigns the prediction low confidence thereby catching the failure, DG assigns high confidence for the same model and prediction, causing a silent failure. This example shows how the tool allows the comparison of CSFs and can help to identify failure modes specific to each CSF. **Manifestation shift:** On the dermoscopy data (Fig. 2g), we see how a manifestation shift can cause silent failures. The benign lesions in the target domain are similar to the malignant lesions in the source domain (rough skin, irregular shapes), and indeed the two failures in the target domain seem to fall into this trap. On the lung nodule CT data (Fig. 2f), we observe a visual distinction

between the spiculated target domain (spiked surface) and the non-spiculated source domain (smooth surface).

5 Conclusion

We see two major opportunities for this work to make an impact on the community. 1) We hope the revealed shortcomings of current systems on biomedical tasks in combination with the deeper understanding of CSF behaviors granted by SF-Visuals will catalyze research towards a new generation of more reliable CSFs. 2) This study shows that in order to progress towards reliable ML systems, a deeper understanding of the data itself is required. SF-Visuals can help to bridge this gap and equip researchers with a better intuition of when and how to employ ML systems for a particular task.

Acknowledgements. This work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

References

1. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans: the LIDC/IDRI thoracic CT database of lung nodules. Med. Phys. **38**(2), 915–931 (2011). https://doi.org/10.1118/1.3528204
2. Band, N., Rudner, T.G.J., Feng, Q., Filos, A., Nado, Z., et al.: Benchmarking Bayesian deep learning on diabetic retinopathy detection tasks. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), January 2022
3. Bernhardt, M., Ribeiro, F.D.S., Glocker, B.: Failure detection in medical image classification: a reality check and benchmarking testbed, October 2022. https://doi.org/10.48550/arXiv.2205.14094
4. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nat. Commun. **11**(1), 3673 (2020). https://doi.org/10.1038/s41467-020-17478-w
5. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. In: NeurIPS, vol. 32. Curran Associates, Inc. (2019)
6. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks, February 2018. https://doi.org/10.48550/arXiv.1802.04865
7. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. arXiv:2106.03004 [cs], July 2021
8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: ICML, pp. 1050–1059. PMLR, June 2016
9. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. arXiv:1705.08500 [cs], June 2017
10. Geifman, Y., El-Yaniv, R.: SelectiveNet: a deep neural network with an integrated reject option. arXiv:1901.09192 [cs, stat], June 2019
11. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv:1610.02136 [cs], October 2018

12. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 4700–4708 (2017)
13. Idrissi, B.Y., Bouchacourt, D., Balestriero, R., Evtimov, I., Hazirbas, C., et al.: ImageNet-X: understanding model mistakes with factor of variation annotations, November 2022. https://doi.org/10.48550/arXiv.2211.01866
14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, January 2019. https://doi.org/10.48550/arXiv.1901.07031
15. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification, November 2022. https://doi.org/10.48550/arXiv.2211.15259
16. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**(1), 160035 (2016). https://doi.org/10.1038/sdata.2016.35
17. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J. Biomed. Health Inform. **23**(2), 538–546 (2019). https://doi.org/10.1109/JBHI.2018.2824327
18. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? arXiv:1703.04977 [cs], March 2017
19. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., et al.: WILDS: a benchmark of in-the-wild distribution shifts, July 2021. https://doi.org/10.48550/arXiv.2012.07421
20. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS, vol. 31. Curran Associates, Inc. (2018)
21. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv:1706.02690 [cs, stat], August 2020
22. Liu, Z., Wang, Z., Liang, P.P., Salakhutdinov, R.R., Morency, L.P., et al.: Deep gamblers: learning to abstain with portfolio theory. In: NeurIPS, vol. 32. Curran Associates, Inc. (2019)
23. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: NeurIPS, vol. 31. Curran Associates, Inc. (2018)
24. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J.: PH2 - a dermoscopic image database for research and benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437–5440, July 2013. https://doi.org/10.1109/EMBC.2013.6610779
25. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., et al.: Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In: NeurIPS, vol. 32. Curran Associates, Inc. (2019)
26. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci. Data **8**(1), 34 (2021). https://doi.org/10.1038/s41597-021-00815-z
27. Santurkar, S., Tsipras, D., Madry, A.: BREEDS: benchmarks for subpopulation shift. In: International Conference on Learning Representations, February 2022
28. Sypetkowski, M., Rezanejad, M., Saberian, S., Kraus, O., Urbanik, J., et al.: RxRx1: a dataset for evaluating experimental batch correction methods, January 2023. https://doi.org/10.48550/arXiv.2301.05768
29. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: ICML, pp. 6105–6114. PMLR, May 2019

30. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 180161 (2018). https://doi.org/10.1038/sdata.2018.161
31. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., et al.: ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR, pp. 3462–3471, July 2017. https://doi.org/10.1109/CVPR.2017.369
32. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., et al.: Contrastive training for improved out-of-distribution detection. arXiv:2007.05566 [cs, stat], July 2020
33. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., et al.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI, pp. 242–252. LNCS. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_24