Automatic dataset shift identification to support safe deployment of medical imaging AI

Mélanie Roschewitz, Raghav Mehta, Charles Jones, and Ben Glocker

Imperial College London
m.roschewitz21@imperial.ac.uk

Abstract. Shifts in data distribution can substantially harm the performance of clinical AI models and lead to misdiagnosis. Hence, various methods have been developed to detect the presence of such shifts at deployment time. However, root causes of dataset shifts are varied, and the choice of shift mitigation strategies highly depends on the precise type of shift encountered at test time. As such, *detecting* test-time dataset shift is not sufficient: precisely *identifying* which type of shift has occurred is critical. In this work, we propose the first unsupervised dataset shift identification framework, effectively distinguishing between *prevalence shift*, *covariate shift* and *mixed* shifts. We show the effectiveness of the proposed shift identification framework across three different imaging modalities (chest radiography, digital mammography, and retinal fundus images) on five types of real-world dataset shifts, using five large publicly available datasets. Code is publicly available at https://github.com/biomedia-mira/shift_identification.

Keywords: Dataset shift · Monitoring · Distribution shift

1 Introduction

Machine learning models are notoriously sensitive to changes in the input data distribution, a phenomenon commonly referred to as dataset shift [28]. This is particularly problematic in clinical settings, where dataset shift is a common occurrence and may arise from various factors [3]. Changes in the frequency of disease positives over time or across geographical regions cause *prevalence shift* [9] (also known as label shift). The use of different acquisition protocols or scanners [22,23], or a change in patient demographics [26] can induce shifts in image characteristics, known as *covariate shift*. We illustrate examples of real-world shifts in Fig. 1. Dataset shift can dramatically affect the performance of AI, lead to clinical errors such as misdiagnosis and is recognised as the fundamental barrier hindering AI adoption [21,18,8,20]. It is hence crucial to implement safeguards allowing not only effective detection of the presence of shifts, but importantly, reliable identification of the root causes. Comprehensive shift detection and identification frameworks are key for the safe deployment and continuous monitoring of AI in clinical practice.

Dataset shifts can be detected at deployment time by using statistical testing to compare the distributions of incoming test data to the distribution of

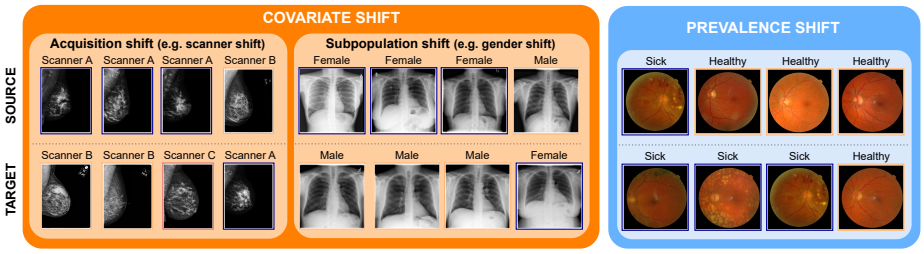

Fig. 1. Examples of dataset shifts in medical imaging. Reliably detecting and identifying the nature of the shift is crucial to enable the safe deployment of ML systems.

the reference data (representative of the data used to validate the deployed AI model). Significant progress has been made in this field where state-of-the-art methods can detect various types of real-world shifts [14,17]. Shifts between test and reference data can either be detected at the output level by comparing distributions of model outputs, or at the input level by comparing low-dimensional feature representations of input images [17]. In this study, we also show that different types of shifts require different shift detection approaches, and that self-supervised (SSL) image encoders [4], yield excellent low-dimensional feature representations for shift detection.

While *detecting* dataset shifts is important, it is insufficient for the safe deployment of AI. Besides knowing that there is a problem, we need to be able to *identify* the precise type of shift to take the necessary actions, implement preventive measures, and safeguard against harm caused by AI errors. Indeed, many domain adaptation techniques are shift-specific: applying the wrong mitigation technique may, in the best case, be ineffective in resolving the shift or, in the worst case, severely harm model performance or calibration. For example, prevalence shifts can often be mitigated with lightweight output recalibration techniques [1,24], but these rely on the assumption that no other types of shift are present. Applying such label shift adaptation methods when the shift is actually caused by covariate shift may drastically degrade model calibration and clinical metrics. In contrast, covariate shifts require more advanced domain adaptation techniques or model fine-tuning [29,12,25]. For example, image-harmonisation techniques (e.g. [12]) or automatic correction methods (e.g. [18]) effectively mitigate effects of acquisition shifts on model performance but will fail in the case of prevalence shift. The difficulty lies in the fact that a change in image characteristics may cause similar changes in the distribution over model outputs as a change in disease prevalence [18], and determining the cause of an observed shift can be challenging. Despite its importance, automatic dataset shift identification has remained an open problem.

In this work, we address this issue by proposing a dataset shift *identification* framework capable of identifying the root cause of the underlying shift, effectively separating (i) prevalence shift, (ii) covariate shift and (iii) mixed shift (both prevalence and covariate shifts). To the best of our knowledge, this is the first

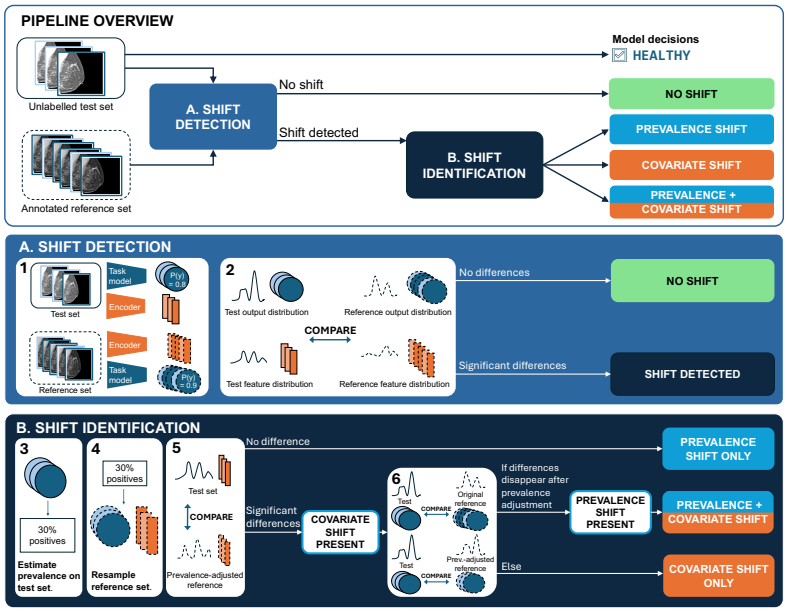

Fig. 2. Proposed dataset shift identification pipeline. Contrarily to previous works, we do not simply detect the *presence* of shifts but also *identify* the nature of the shift, leveraging both task model outputs and features from self-supervised encoders.

framework able to identify the type of test-time shifts in an unsupervised manner for imaging data, beyond solely detecting shifts. An in-depth evaluation across three different clinical applications (chest radiography, digital mammography, and retinal fundus images) on five types of real-world dataset shifts demonstrates that our framework accurately distinguishes between prevalence shifts, covariate shifts, and mixed shifts across various scenarios.

2 Background

Shift definitions. Formally, let X denote the input image and Y the target. *Label shift* [16] (or prevalence shift) occurs when label distribution changes across domains i.e. $P_{ref}(Y) \neq P_{test}(Y)$, the conditional distributions are preserved, i.e. $P_{ref}(X|Y) = P_{test}(X|Y)$, where P_{ref} and P_{test} denote distributions on reference and target domains respectively. Conversely, *covariate shift* occurs when $P_{ref}(X) \neq P_{test}(X)$, while conditional distributions are preserved [16]. Acquisition and subpopulation shifts are cases of covariate shifts as they directly affect image appearance.

Dataset shift detection. Several paradigms have been proposed for dataset shift detection. The simplest method to implement consists of comparing distributions of a classifier's outputs between the reference and test domain, proposed by

Rabanser et al. [17] and referred to as 'Black Box Shift Detection' (BBSD). Specifically, softmax model outputs are collected for all samples in the reference and test sets. Then, for each class, a separate univariate Kolmogorov-Smirnov (K-S) test is run to determine if the class-wise predicted probabilities distributions differ between reference and test domain, the overall significance of the shift is then determined after applying Bonferroni [7] correction for multiple testing. Rabanser et al. [17] also propose another type of shift detector where reference and test data input distribution are compared using a feature-based approach. In this test, the input sample (image) first gets projected to a smaller dimension, e.g. through a pretrained neural network encoder and the shift is then measured using the 'Maximum Mean Discrepancy' (MMD) permutation test. Other shift detection approaches consist of training a domain classifier to classify samples between the reference and test domains [15,10] and using the accuracy of this classifier as a proxy for measuring distances between distributions. One drawback of this approach is its high computational cost: for every test set, a new domain classifier must be trained, which is highly impractical in continuous monitoring scenarios.

3 Methods

3.1 Dataset shift identification pipeline

We propose a framework for identifying whether dataset shift is caused by prevalence shift, by covariate shift or by a mix of both. The approach is divided in two stages (Fig. 2). First, we perform standard dataset shift detection to separate the 'shift' from the 'no shift' cases (Fig. 2, A). For this step, we use a dual detection approach, combining signals from task model outputs and features from self-supervised (SSL) encoders to detect shifts. In details, we independently run the BBSD and the MMD shift detection tests; this yields C p-values for the BBSD test (one per class), and one p-value for the MMD permutation test, we then apply Bonferroni correction on the $C + 1$ p-values to get the overall significance. If a shift is detected we then proceed to shift *identification*. This starts with estimating the prevalence in the test set (Fig. 2, B.3). For this, we can leverage the prevalence shift adaptation literature, where various methods have been proposed to estimate the density ratio $P_{ref}(Y)/P_{test}(Y)$ [1,19,24]. Here, we use the state-of-the-art 'class probability matching with calibrated networks' (CPMCN) method [24] to estimate this ratio and the test set prevalence. Next, we resample the reference set to match this estimated prevalence (Fig. 2, B.4). We then first compare feature distributions between prevalence-adjusted reference and test set (Fig. 2, B.5): if differences are no longer significant after adjusting the prevalence, the shift is attributed to prevalence shift. Conversely, if differences persist after adjusting the prevalence, then covariate shift is necessarily present. In this case, we compare model output distributions with BBSD to determine whether prevalence shift is also present (Fig. 2, B.6). Precisely, if there were significant differences in model output distributions before adjusting the prevalence, but this shift disappears after adjusting the prevalence, we know

that prevalence shift is also responsible for the observed shift, in this case we conclude that the observed shift is a case of mixed shift (prevalence + covariate shift). Else, we conclude that the shift is attributed to covariate shift only.

3.2 Datasets and shift generation

We evaluate our methods on four different datasets covering three different imaging modalities: (i) chest radiography using PadChest [2] (collected in Spain, with two scanners); (ii) mammography using the EMBED [11] dataset (mammograms acquired in the US with six different scanners), and (iii) fundus images for which we create 'RETINA' a multi-domain dataset by combining three different public datasets: the Kaggle Diabetic Retinopathy Detection dataset [6], the Kaggle Aptos Blindness Detection dataset [13] and the Messidor-v2 dataset [5] (covering three countries US, India and France and various acquisition devices from high quality scanners to phone pictures). To study prevalence shift detection, we associate each dataset with a downstream task. For chest radiography datasets we focus on pneumonia detection, for mammography on breast density assessment (4 classes), and for retinal images on binary diabetic retinopathy classification. Then, we study various types of covariate shifts. For PadChest, we study gender shift by varying the proportion of female patients in the test set. Moreover, PadChest contains scans acquired with two scanners, 'Phillips' (40%) and 'Imaging' (60%). This allows to simulate different levels of acquisition shift, by varying the proportion of Phillips scans in the test set. Similarly, for EMBED [11], we study acquisition shift by varying the distribution of scanners in the test set. This dataset offers a complementary view to PadChest, with a multi-class task of interest and providing even more flexibility for simulating diverse acquisition shifts (six scanners). Note that in EMBED each exam comprises 4 mammograms (left/right breasts and MLO/CC views). We excluded all exams that did not contain exactly four images and kept exactly one exam per patient, and ensure that test set sampling was done at the exam level. Finally, for the RETINA dataset, we simulate covariate shifts by varying the proportion of samples coming from each underlying dataset (Aptos [13], Kaggle DR [6] and Messidor [5]).

Implementation details

4 Results

4.1 Different shifts require different shift detectors

Prior to diving into shift *identification*, we first investigate which types of shifts are successfully *detected* by prominent dataset shift detection methods. We compare two families of shift detectors: model output-based (BBSD) and feature-based detectors (MMD). We additionally test a dual approach that combines both approaches for improved shift detection ('Duo'). For feature-based shift detection, any pretrained network can be used as feature extractor, we could

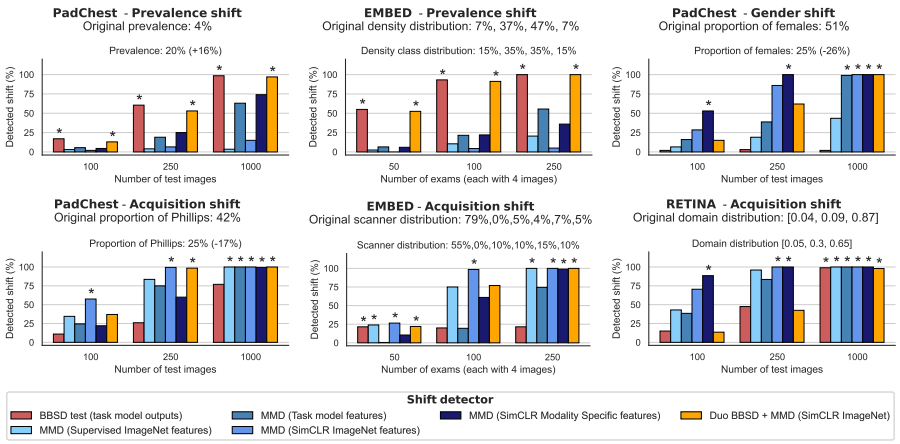

Fig. 3. Shift detectors comparison. We studied prevalence shift as well as two subtypes of covariate shifts: subpopulation and acquisition shift. Shift detection rate is computed over 200 bootstrap samples. Feature-based detection (MMD) is best for covariate shift detection and model-output based detection (BBSD) is best for prevalence shift. For each test set size, the best detector along with all detectors not significantly different from it, are denoted with $*$ (Fisher's exact test, at level .05, with Bonferroni).

for example use the task model directly. However, this may not be the best choice as learned features will be heavily skewed towards encoding characteristics specifically relevant to that task as opposed to encoding more generic image representations [27]. Hence, we here explore the potential of SSL image encoders for shift detection as these encoders learn to effectively summarise the information encoded in a given image. We compare the performance of MMD detection for four different encoders: (i) 'Supervised ImageNet' trained to perform classification on ImageNet; (ii) 'Task model' trained to perform the downstream classification task; (iii) 'SSL ImageNet' trained on ImageNet data only; (iv) 'SSL Modality Specific' trained on the same modality as the reference dataset.

Fig. 3 show the shift detection rates for various dataset-shift combination across shift detectors. We can see that for prevalence shifts, output-based shift detection performs significantly better than all feature-based tests. This is intuitive as a change in prevalence should directly be reflected by a change in the distribution of task model outputs. The exact opposite is true for covariate shifts, where most feature-based detectors perform substantially better than output-based detectors, regardless of whether we look at acquisition or subpopulation shifts. The results also show that for optimal detection of more subtle covariate shifts, it is crucial to use an SSL encoder, as SSL encoders offer substantially better detection rates than their supervised counterparts. The SSL model trained on ImageNet data was particularly effective and, in some cases even better than the modality-specific SSL model. Given the orthogonal behaviour of output-based and feature-based tests, we introduce a dual detection approach

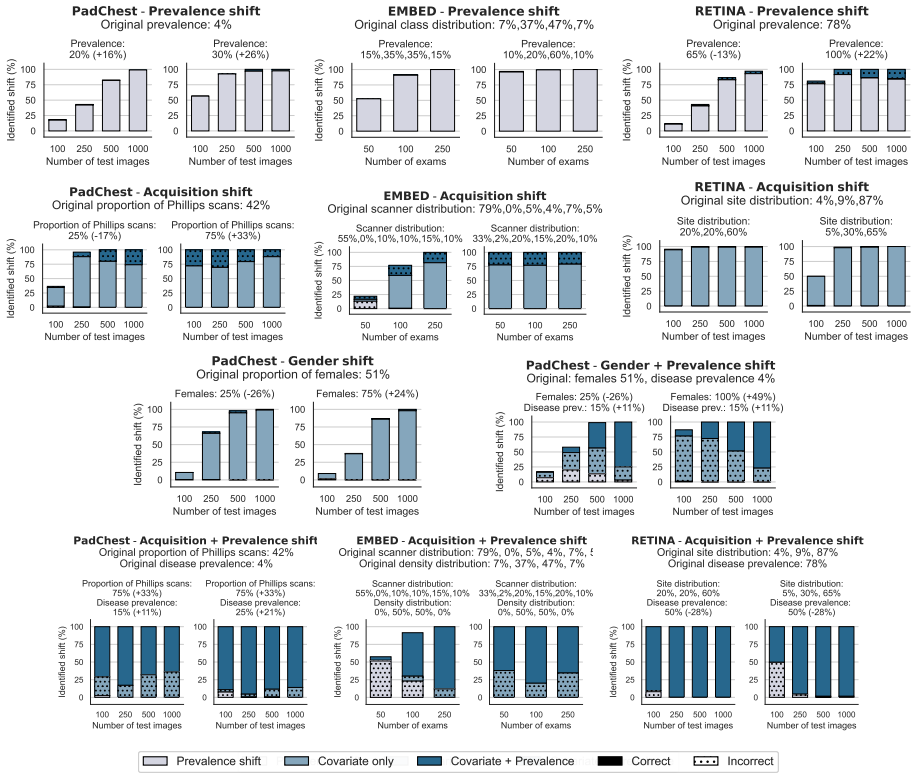

Fig. 4. Shift identification accuracy. Across all datasets, the shift identification framework is able to successfully detect and identify all types of shifts high accuracy.

combining responses from output-based and SSL feature-based shift detectors (see Section 3). Results in Fig. 3 confirm that this 'Duo' approach performs best overall across shifts and datasets. For example, in PadChest, with 1000 test cases: MMD (ImageNet-SSL features) has 65% detection rate of prevalence shift, 100% for gender and acquisition shift, i.e. an average of 88% detection across shifts, BBSD has an average of 60% across shifts, and Duo has an average of detection rate close to 100%. We employ this 'Duo' approach for the detection step of our shift identification pipeline.

4.2 Shift identification accuracy

In this work, we propose a framework able to precisely identify the type of shift present in the test dataset (see Section 3). Evaluation results in Fig. 4 show that our framework is capable of distinguishing between prevalence shifts, covariate shifts and mixed shifts with high accuracy across all datasets and types of shifts. For prevalence shifts, the average shift identification rate, across datasets and shift levels, is 91% with 500 test images (top row). Similarly, for covariate shifts

(acquisition and gender shifts), when using a test set size of 500 images (250 exams on EMBED), the identification accuracy is greater than 80% for any shift level and dataset and the average accuracy is 87%. Importantly, the framework is also able to accurately distinguish between cases of covariate shift only and cases of covariate and prevalence shifts. With a test set of 500 images (resp. 250 exams on EMBED), mixed gender and prevalence shifts are correctly identified as mixed shifts with an average accuracy of 75% across all tested scenarios. Overall, more subtle shifts are best detected with larger test sets whereas larger shifts can be detected with smaller test sets.

5 Discussion

Shift identification is critical for monitoring AI systems in deployment and for the root cause analysis of AI performance drift. By analysing common dataset shift detection paradigms, we find that different types of shifts require different types of shift detectors. Our analysis also demonstrates that encoders trained in a self-supervised manner yield features with substantially higher shift detection power than supervised counterparts. Maybe surprisingly, our results show that generic encoders trained in a self-supervised manner on ImageNet provide highly discriminative features for medical image dataset shift detection, transportable across datasets. Following these findings, we propose a lightweight and practical shift identification framework, capable of detecting and identifying important real-world dataset shifts, showing high accuracy in correctly identifying the type of shift present in the test set across all modalities, tasks and various levels of shift intensity. Our results demonstrate the practical value of the shift identification method, which does not require any training at test time, nor any ground truth labels or annotations on the test domain data. The use of a readily available SSL encoder trained on ImageNet data for feature extraction dispenses us from training any additional model for shift detection and identification purposes. Combining signals from both feature-based and model output-based shift detectors yields reliable and consistent detection and identification across all types of shifts. Importantly, our method not only separates covariate shifts from prevalence shifts but also reliably detects when both types of shifts are present in the test set, rendering the proposed method applicable to many real-world deployment scenarios.

In terms of limitations, we note that in the case of covariate shift, on its own, the proposed framework does not allow for a more fine-grained identification of the origin of shift, e.g. the distinction between population and acquisition shifts. To allow for an even more precise sub-type shift identification, integrating metadata statistics in the pipeline could complement the framework. However, relying solely on metadata monitoring is insufficient as it only enables the detection of shifts affecting the specific attributes collected at test time. The proposed framework can help uncover shifts that are not detectable by means of simply tracking population metadata. In this context, our shift identification framework offers an important safeguard for deploying AI models in clinical practice.

Acknowledgments

M.R. is funded by an Imperial College London President's PhD Scholarship and a Google PhD Fellowship. R.M. is funded through the European Union's Horizon Europe research and innovation programme under grant agreement 10108030. C.J. is supported by Microsoft Research and EPSRC through the Microsoft PhD Scholarship Programme. B.G. acknowledges support from the Royal Academy of Engineering as part of his Kheiron Medical Technologies/RAEng Research Chair in Safe Deployment of Medical Imaging AI.

Disclosure of interests

B.G. is part-time employee of DeepHealth. No other competing interests.

References

1. Alexandari, A., Kundaje, A., Shrikumar, A.: Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation. In: Proceedings of the 37th International Conference on Machine Learning. pp. 222–232. PMLR (Nov 2020), iSSN: 2640-3498
2. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66**, 101797 (Dec 2020)
3. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (Jul 2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. PMLR (Nov 2020)
5. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J.C.: Feedback on a publicly distributed image database: the messidor database. Image Analysis and Stereology **33**(3), 231–234 (Aug 2014), number: 3
6. Dugas, E., Jared, J., Cukierski, W.: Diabetic Retinopathy Detection Kaggle Challenge (2015), https://kaggle.com/competitions/diabetic-retinopathy-detection
7. Dunn, O.J.: Multiple Comparisons among Means. Journal of the American Statistical Association **56**(293), 52–64 (Mar 1961)
8. Finlayson Samuel G., Subbaswamy Adarsh, Singh Karandeep, Bowers John, Kupke Annabel, Zittrain Jonathan, Kohane Isaac S., Saria Suchi: The Clinician and Dataset Shift in Artificial Intelligence. New England Journal of Medicine **385**(3), 283–286 (Jul 2021)
9. Godau, P., Kalinowski, P., Christodoulou, E., Reinke, A., Tizabi, M., Ferrer, L., Jäger, P.F., Maier-Hein, L.: Deployment of Image Analysis Algorithms Under Prevalence Shifts. In: MICCAI 2023. pp. 389–399. Springer (Oct 2023)
10. Jang, S., Park, S., Lee, I., Bastani, O.: Sequential Covariate Shift Detection Using Classifier Two-Sample Tests. In: Proceedings of the 39th International Conference on Machine Learning. pp. 9845–9880. PMLR (Jun 2022), iSSN: 2640-3498

11. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., Woo, M., McAdams, C.R., Newell, M.S., Banerjee, I., Gichoya, J., Trivedi, H.: The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images. Radiology: Artificial Intelligence **5**(1), e220047 (Jan 2023). https://doi.org/10.1148/ryai.220047
12. Kang, H., Luo, D., Feng, W., Zeng, S., Quan, T., Hu, J., Liu, X.: StainNet: A Fast and Robust Stain Normalization Network. Frontiers in Medicine **8** (2021)
13. Karthik, M., Sohier, D.: APTOS 2019 Blindness Detection Kaggle Challenge (2019), https://kaggle.com/competitions/aptos2019-blindness-detection
14. Koch, L.M., Baumgartner, C.F., Berens, P.: Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. npj Digital Medicine **7**(1), 1–11 (May 2024)
15. Lopez-Paz, D., Oquab, M.: Revisiting Classifier Two-Sample Tests. In: ICLR (Jul 2017)
16. Murphy, K.P.: Probabilistic Machine Learning: Advanced Topics. MIT Press (2023)
17. Rabanser, S., Günnemann, S., Lipton, Z.: Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
18. Roschewitz, M., Khara, G., Yearsley, J., Sharma, N., James, J.J., Ambrozay, E., Heroux, A., Kecskemethy, P., Rijken, T., Glocker, B.: Automatic correction of performance drift under acquisition shift in medical image classification. Nature Communications **14**(1), 6608 (Oct 2023)
19. Saerens, M., Latinne, P., Decaestecker, C.: Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Computation **14**(1), 21–41 (Jan 2002)
20. Sahiner, B., Chen, W., Samala, R.K., Petrick, N.: Data drift in medical machine learning: implications and potential remedies. British Journal of Radiology **96**(1150), 20220878 (Oct 2023)
21. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**(12), 2176–2182 (Dec 2021)
22. Sharma, N., Ng, A.Y., James, J.J., Khara, G., Ambrozay, E., Austin, C.C., Forrai, G., Fox, G., Glocker, B., Heindl, A., Karpati, E., Rijken, T.M., Venkataraman, V., Yearsley, J.E., Kecskemethy, P.D.: Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. BMC Cancer **23**(1) (May 2023)
23. Stacke, K., Eilertsen, G., Unger, J., Lundstrom, C.: Measuring Domain Shift for Deep Learning in Histopathology. IEEE journal of biomedical and health informatics **25**(2), 325–336 (Feb 2021)
24. Wen, H., Betken, A., Hang, H.: Class Probability Matching with Calibrated Networks for Label Shift Adaption. In: Proceedings of The Twelfth International Conference on Learning Representations (Apr 2024)
25. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning Semantic Representations for Unsupervised Domain Adaptation. In: Proceedings of the 35th International Conference on Machine Learning. pp. 5423–5432. PMLR (Jul 2018)
26. Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M.: The limits of fair medical imaging AI in real-world generalization. Nature Medicine pp. 1–11 (Jun 2024)

27. Zamzmi, G., Venkatesh, K., Nelson, B., Prathapan, S., Yi, P., Sahiner, B., Delfino, J.G.: Out-of-Distribution Detection and Radiological Data Monitoring Using Statistical Process Control. Journal of Imaging Informatics in Medicine (Sep 2024)
28. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain Generalization: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4396–4415 (Apr 2023)
29. Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A.: Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. NeuroImage **243**, 118569 (Nov 2021)