Subgroup Performance Analysis in Hidden Stratifications

Alceu Bissoto[1,2,3(✉)], Trung-Dung Hoang[1,2,3], Tim Flühmann[1,2,3], Susu Sun[4], Christian F. Baumgartner[4,5], and Lisa M. Koch[1,2,3(✉)]

[1] University of Bern, Switzerland
[2] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Bern University Hospital, Switzerland
[3] Diabetes Center Berne, Switzerland
[4] Cluster of Excellence: Machine Learning - New Perspectives for Science, University of Tübingen, Germany
[5] Faculty of Health Sciences and Medicine, University of Lucerne, Switzerland
{alceu.bissoto,lisa.koch}@unibe.ch

Abstract. Machine learning (ML) models may suffer from significant performance disparities between patient groups. Identifying such disparities by monitoring performance at a granular level is crucial for safely deploying ML to each patient. Traditional subgroup analysis based on metadata can expose performance disparities only if the available metadata (e.g., patient sex) sufficiently reflects the main reasons for performance variability, which is not common. Subgroup discovery techniques that identify cohesive subgroups based on learned feature representations appear as a potential solution: They could expose hidden stratifications and provide more granular subgroup performance reports. However, subgroup discovery is challenging to evaluate even as a standalone task, as ground truth stratification labels do not exist in real data. Subgroup discovery has thus neither been applied nor evaluated for the application of subgroup performance monitoring. Here, we apply subgroup discovery for performance monitoring in chest x-ray and skin lesion classification.We propose novel evaluation strategies and show that a simplified subgroup discovery method without access to classification labels or metadata can expose larger performance disparities than traditional metadata-based subgroup analysis. We provide the first compelling evidence that subgroup discovery can serve as an important tool for comprehensive performance validation and monitoring of trustworthy AI in medicine[1].

Keywords: Subgroup discovery · performance monitoring

1 Introduction

Machine learning (ML) models often perform systematically differently across patient subgroups [18,6,13,15]. This has hampered past attempts at safely deploying medical AI in particular in underserved populations [18,6]. Model performance can depend on many factors [9], including patient attributes (e.g., sex,

[1] Code available at `https://github.com/alceubissoto/hidden-subgroup-perf`

age, ethnicity) and image characteristics (e.g., image quality, artifacts, device manufacturer). Subgroup analysis based on such metadata could identify disparate outcomes in patient groups. However, limited metadata typically exists, and available metadata may not adequately capture the data's true variability nor incorporate concepts important to ML models. Hidden stratifications therefore often exist, which can lead to systematic performance disparities that go unnoticed in the evaluation of ML models [15].

Recently, subgroup discovery methods have emerged for algorithmically identifying systematically different subgroups in computer vision tasks [7,8,20]. These techniques appear as a potential solution for more comprehensive model validation as they could expose hidden stratifications and enable more detailed subgroup performance analyses. However, subgroup discovery is challenging to evaluate even as a standalone task, as labels for ground truth stratifications inherently do not exist in real data. The lack of labels hinders its application to performance monitoring, for which it remains surprisingly underexplored. As a result, current evaluation approaches are limited to (1) less realistic synthetic datasets, where factors of variations can be fully controlled, or (2) measuring alignment with known characteristics such as patient sex or age, which we realistically cannot expect to characterise the main factors of variation in heterogeneous data distributions.

In this paper, we apply and evaluate subgroup discovery in the downstream application of subgroup performance analysis (Fig. 1). While validation remains challenging, we propose novel evaluation metrics and provide evidence on synthetic and real-world medical image classification tasks that subgroup discovery can expose systematic performance gaps. We argue that subgroup discovery can be an effective and easily implemented tool to enhance the performance validation and monitoring of ML systems in medicine. Our main contributions are:

- We provide the first comprehensive evidence that subgroup discovery can systematically expose performance gaps in medical imaging, identifying meaningful subgroups in both synthetic and real-world settings.
- We introduce novel metrics to evaluate the quality of discovered subgroups.
- We demonstrate that discovered subgroups exhibit significantly larger performance disparities than conventional demographic metadata, revealing critical gaps missed by traditional fairness auditing.

2 Methods

We seek subgroup divisions that expose large systematic performance gaps of a target classification model while preserving subgroup cohesion, so that the model performance in a subgroup can be attributed to a shared characteristic. Metadata-based subgroups are inherently cohesive since the division is provided by a semantic concept such as patient age or sex. However, we hypothesise that these attributes do not adequately reflect the main factors of variation affecting model performance, which often results in relatively small performance gaps

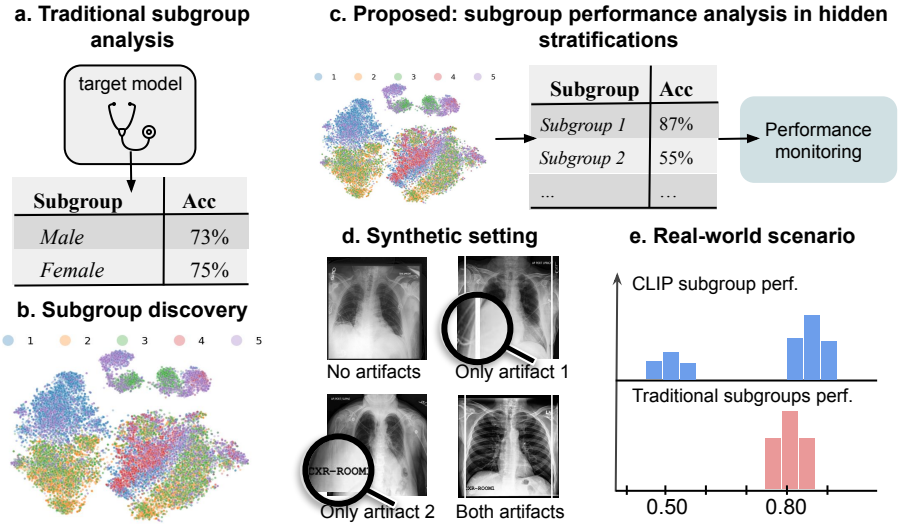

Fig. 1: (a) Traditional subgroup analysis detects disparate patient outcomes, but it is limited to annotated metadata. (b) Subgroup discovery reveals hidden stratifications but lacks performance validation. (c) We bridge this gap by applying subgroup discovery for performance analysis in both (d) controlled synthetic settings and (e) real-world scenarios with unknown subgroups.

(Fig. 1 **a**). Instead, we propose to use subgroup discovery techniques (Fig. 1 **b**) for subgroup performance analysis in hidden stratifications (Fig. 1 **c**).

We propose a two-tiered evaluation approach to tackle the difficult challenge of validating hidden subgroups. First, we inject synthetic artifacts to create clinically-inspired subgroups where ground-truth model performance is available (Fig. 1 **d**). Finally, we propose a strategy to evaluate subgroup discovery for performance analysis for the first time in a real-world data distribution (Fig. 1 **e**).

2.1 Preliminaries: Subgroup Discovery Algorithm

We use DOMINO [8], a simple yet effective approach for subgroup discovery. First, a feature representation $z(x)$ is extracted from each image x using an external pretrained model such as CLIP [16] followed by dimensionality reduction using principal component analysis. In addition, softmax predictions $\hat{y}(x)$ are obtained from the target classification model. While the model predictions encapsulate characteristics important for the classification task, the external model helps identify task-agnostic features such as artifacts. Next, the samples are clustered into subgroups $\mathcal{S}$ using a generalised Gaussian Mixture Model (GMM) by minimising the following objective (similar to [8]):

$$\ell(\phi) = \sum_{i=1}^{n_{samples}} \log \sum_{j=1}^{|\mathcal{S}|} P_{\phi_S}(\mathcal{S}^{(j)}=1) P_{\phi_Z}(Z=z(x_i) \,|\, \mathcal{S}^{(j)}=1) P_{\phi_{\hat{Y}}}(\hat{Y}=\hat{y}(x_i) \,|\, \mathcal{S}^{(j)}=1)^{\gamma}, \quad (1)$$

where γ balances the influence of predicted labels $\hat{y}(x)$ and embeddings $z(x)$ in the slicing decision. In contrast to the original DOMINO [8], we remove classification labels in the GMM, enabling subgroup discovery in post-deployment scenarios with unlabeled test sets. We use explicit validation and test sets separations, fitting DOMINO on validation, and inferring subgroups on the test set.

2.2 Synthetic scenario with generated artifacts

We first evaluate subgroup discovery in a simulated scenario where ground truth subgroups and subgroup performances are known. To simulate performance disparities, we add artifacts spuriously correlated with the positive disease label, similar to standard practice in shortcut learning research [17,2,3,19]. In particular, we introduce a simulated scenario where we synthetically add two artifacts independently correlated with the label: one is a known attribute for traditional subgroup analysis, but the other is hidden and could potentially be exposed by subgroup discovery. The artifacts are inserted on positive samples with probability, or bias level, p, and on negative samples with probability $1 - p$, resulting in four ground truth subgroups. Training and validation sets are generated from this biased version of the data and are used for training the target classification model and selecting its hyperparameters. The validation set is also used to fit DOMINO. For testing, we use an unbiased test set ($p = 0.5$), facilitating fair comparisons across training bias levels.

2.3 Real-world data distribution: unknown hidden stratifications

Next, we assess the ability of subgroup discovery to reveal hidden performance gaps in real-world data where no labels exist for hidden stratifications. Following the same procedure as in the synthetic setting, we train the target classification model and DOMINO based on the training and validation set and infer subgroups on the test set. In the absence of ground truth labels for hidden subgroups, we use measured metadata (e.g. patient age, sex) as a baseline stratification method, which reflects current standard practice for subgroup performance analysis. Each metadata attribute (e.g. patient sex) defines a different subgroup division (male vs. female), assigning each sample its corresponding attribute performance. We average the performance values across all its metadata attributes to obtain an overall performance metric for each sample. For subgroup discovery, we can extend the same idea to marginalize over the stochastic effects caused by the use of different random seeds, providing a more robust estimation of the discovered subgroup performances.

2.4 Evaluation metrics

An ideal stratification leads to subgroups with systematic performance differences. Identifying large performance gaps across cohesive groups may provide actionable insights into the failure modes of the target classification model. We

propose two new metrics to evaluate the quality of discovered subgroups: performance gap and average purity. We measure the **performance gap** of a subgroup division S as $\Delta(S) = \max_{s \in S} M(s) - \min_{s \in S} M(s)$, where $M(s)$ is the model performance in subgroup s, e.g. accuracy. **Average purity** measures subgroup cohesion by calculating how well subgroups align with known attributes, such as the presence of artifacts or patient characteristics. For subgroup s, let $n_{s,a}$ be the number of samples with attribute a and n_s the total samples. The purity of s is the fraction of samples in its majority attribute, corrected by a term c for robustness to small subgroups. Then, the average purity is given by $AP(S) = \frac{1}{|A|} \sum_{a \in A} \max_{s \in S_a} \left(\frac{n_{s,a}}{n_s + c} \right)$, where S_a is the set of subgroups whose majority attribute is a.

2.5 Datasets

We selected datasets that provide comprehensive coverage of metadata. CheXpert-Plus [5] is an extension of CheXpert [11] and provides metadata that allows for a challenging comparison to our discovered subgroups. The metadata includes patient demographics (e.g., sex, age), comorbidities (e.g., edema, fracture), and artifacts, totalling 20 attributes. Our training, validation, and test set follow an 80/10/10 division, with a total of 178,684 / 22,263 / 22,281 images respectively.

SLICE-3D [14] is a recent skin lesion classification dataset. Apart from patient details (e.g., sex, age), it includes lesion-specific visual traits, enabling analysis of subgroups aligned with diagnostic-relevant features (e.g., lesion hue and size). Due to the dataset's imbalance, we allocated more samples to the validation and test sets to ensure an adequate number of positive cases. We divided Patient IDs in a 60/20/20 scheme, resulting in 252,047 / 80,516 / 68,496 images.

While we use both datasets for our real-world experiments, we adapt CheXpertPlus with two clinically-inspired artifacts following previous work [19] (Fig. 1 **b**): a *hospital tag* on the bottom left, and vertical lines of *hyperintense signal*.

3 Results

3.1 Experimental setup

For all experiments, we trained ResNet-50 [10] classification models using SGD and searched over learning rates of $\{10^{-5}, 10^{-4}, 10^{-3}\}$ with the weight decay of 10^{-4}. Models were selected based on validation balanced accuracy for CheXpert-Plus and thresholded AUC for SLICE-3D. For CheXpertPlus, we chose the task of "cardiomegaly vs. all", while for SLICE-3D we followed the original problem of "malignant vs. benign". For subgroup discovery, we always used 15 subgroups. The external model is the pretrained CLIP [16] for all scenarios, and we included BiomedCLIP [21] for our real-world CheXpertPlus experiments. In the synthetic scenario (Sec. 2.2), we considered hyperintensities as a known attribute and hospital tags as a hidden stratification and varied the bias level p between 0.6, 0.7, 0.8. We use accuracy as our primary metric for measuring performance gaps and subgroup performances.

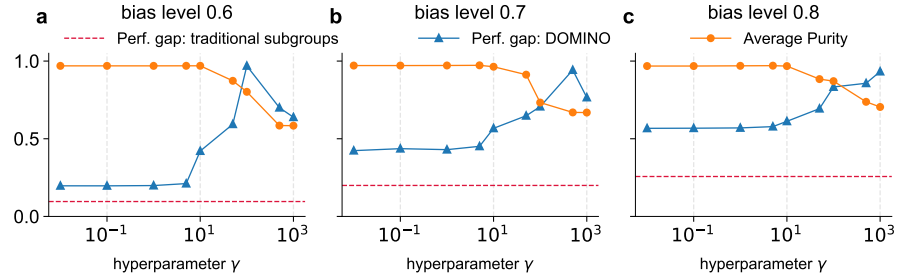

Fig. 2: Performance gap and purity of subgroups across different γ and bias levels.

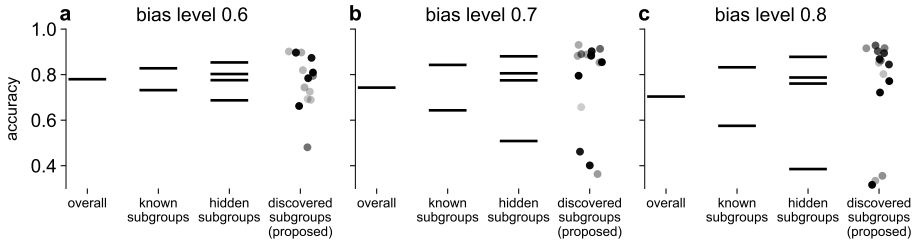

Fig. 3: Detailed subgroup accuracies for our synthetic scenario. Purer subgroups performances (darker dots) capture the true performance gap characterized by hidden subgroups, which are overlooked by traditional subgroup analysis with access to a single artifact (known subgroups), and by overall performance.

3.2 Subgroup discovery uncovers large performance disparities while maintaining cohesive subgroups

Across all experiments in both synthetic (Fig. 2) and real-world settings (Fig. 4 **a, c**), subgroup discovery consistently exposed performance gaps larger than traditional subgroups (red dash line in Figs. 2 **a-c**) without sacrificing cohesion. While the performance gap and purity competed, performance gaps increased before purity declined when increasing γ. This allowed substantial performance disparities to be exposed without sacrificing the cohesiveness of the subgroups. We chose the "elbow" point before a sharp purity decrease, resulting in $\gamma = 10$ for synthetic and real-world CheXpertPlus, and $\gamma = 50$ for the SLICE-3D, as shown in Figs. 2 and 4 **a, c**.

3.3 Subgroup discovery captures actual subgroup performance

In our synthetic scenario with one known and one unknown artifact, subgroup analysis based on the known artifact unsurprisingly revealed increasing performance gaps when increasing the bias level from 0.6 to 0.8, but missed the much

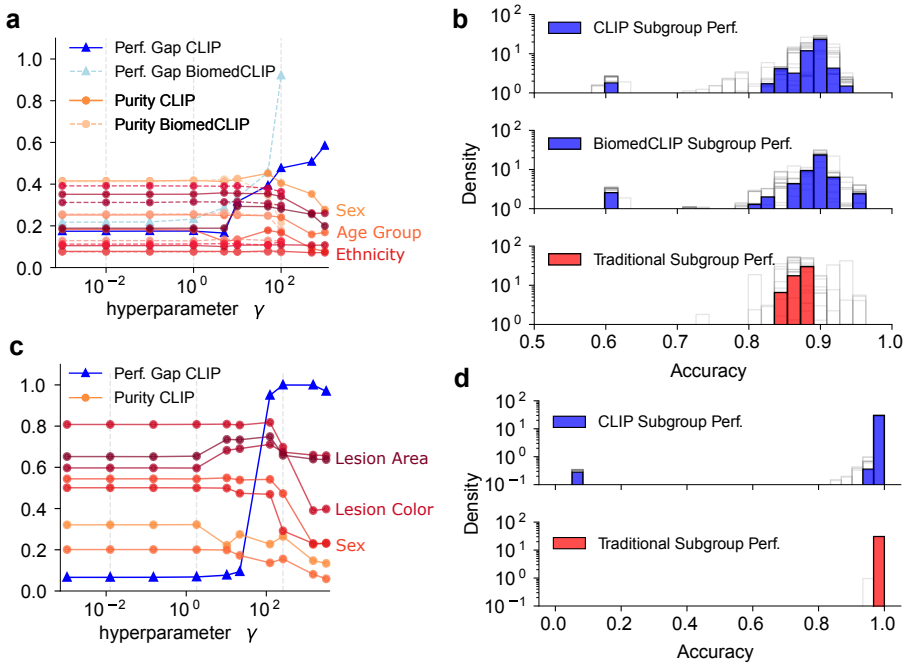

Fig. 4: (a,c): Performance gaps and metadata-based purity for different γ. (b,d): Histograms of subgroup performances for different subgroup divisions: In blue, subgroup discovery with different external CLIP models, averaged over different random seeds (gray transparent bars). In red, our baseline of subgroups defined by different metadata, averaged over their attributes (gray). Top (a,b) and bottom (c,d) rows show CheXpertPlus and SLICE-3D results, respectively.

larger, hidden performance gaps caused by the second artifact (Fig. 3 **a-c**). Subgroup discovery without access to either artifact annotations successfully found subgroups that captured the hidden subgroup performances (dots in Fig. 3 **a-c**).

As our simulated artifacts were added to real data where factors unknown to us could additionally affect performance, discovered subgroups exposed additional performance disparities. For example, in Fig. 3 **b**, one discovered subgroup neither aligned with the hidden subgroups in terms of performance, nor in terms of purity (reflected by light grey colour).

3.4 Subgroup discovery exposes higher performance gaps than traditional subgroup analysis in real-world scenarios

Finally, we applied subgroup discovery for performance analysis in two real-world applications without artificial artifacts in chest x-ray and skin lesion analysis. In both cases, hidden biases were likely present but not annotated [12,4]. Subgroup discovery identified higher performance gaps than traditional metadata-based

analysis (see performance histograms in Fig. 4 **b, d**). On CheXpertPlus (Fig. 4 **b**), subgroup discovery consistently found underperforming subgroups with less than 60% accuracy, while the majority of subgroups achieved around 90% accuracy. In contrast, metadata-based analysis did not expose such low-performing subgroups and led to a narrower range of performances overall. For skin lesion analysis, subgroup discovery found a subgroup with 721 negatives and 17 positives with only 5% accuracy (Fig. 4 **d**).

3.5 Discovered subgroups in real-world scenarios do not capture patient demographics, but align well with visual features

In the CheXpertPlus dataset, the discovered subgroups did not align well with concepts described by the available metadata, leading to subgroups with low purity concerning attributes such as patient sex, age or ethnicity (Fig. 4 **a**). This confirms that available metadata often does not reflect the main factors of variability in real-world data distributions.

In contrast, the SLICE-3D skin lesion dataset contained annotations of visual features such as lesion area or colour. The discovered subgroups were well stratified by these visual features. This was reflected by high purity across a wide range of DOMINO configurations (Fig. 4 **c**). Demographic attributes such as patient sex remained at a low purity level, similar as in the CheXpert experiments. While some annotated lesion characteristics (e.g. area, colour) are related to lesion malignancy [1], subgroup analysis based on these attributes did not expose the performance disparities we observed with discovered subgroups.

3.6 Feature extractors trained on natural images are sufficient for exposing meaningful performance gaps

Finally, we used BiomedCLIP [21] as a feature extractor for subgroup discovery in CheXpert to investigate whether representations learned from biomedical data led to better stratification of disease-related features in medical images. However, BiomedCLIP and original CLIP led to similar subgroup purities (Fig. 4 **a**) and performance disparities (Fig. 4 **a, b**). This indicates that even feature extractors trained on natural images can expose meaningful performance gaps in real-world data distributions, where factors of variation may be more visually subtle than the simulated artifacts we introduced in our synthetic experiments.

4 Discussion

We demonstrate that hidden stratifications in synthetic and real-world data can lead to performance disparities, which often cannot be detected by traditional metadata-based subgroup analysis. Meanwhile, subgroup discovery exposed substantial and systematic performance disparities between cohesive subgroups. In the synthetic scenario, discovered subgroups accurately captured artificial ground truth subgroups (Sec. 3.3). In real-world data, where the true factors of

variation in data might be more visually subtle, we showed evidence that the factors guiding subgroup discovery are not necessarily low-level perceptual features (Sec. 3.5). For skin lesion analysis, the lesion color, which is clinically relevant for the diagnosis of melanoma, indirectly influenced the subgroup discovery, resulting in high average purity. While no ground truth stratification labels exist for real data, our results were robust and consistent across datasets, hyperparameter configurations and random seeds. We conclude that subgroup discovery should be highly relevant as a performance monitoring and reporting tool, and argue that it should accompany traditional subgroup analysis as an additional safeguard during real-world ML validation and deployment.

Future work could further investigate subgroup discovery robustness, facilitating their adoption by ML practitioners. Beyond their use in safe deployment, our subgroup performance analysis approach could be useful for developing unbiased ML models. There, discovered subgroups could replace or augment existing subgroup labels, e.g. when reporting worst-group performance.

Acknowledgments. This project was supported by the Diabetes Center Berne (AB, TH, TF, LK), the Carl Zeiss Foundation in the project "Certification and Foundations of Safe Machine Learning Systems in Healthcare" (SS), the Deutsche Forschungsgemeinschaft (DFG) – EXC number 2064/1 – Project number 39072764 (SS, CB), and strategic funding of the medical faculty of the University of Bern (TH). Calculations were performed on UBELIX, the HPC cluster at the University of Bern.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abbasi, N.R., Shaw, H.M., Rigel, D.S., Friedman, R.J., McCarthy, W.H., Osman, I., Kopf, A.W., Polsky, D.: Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. Jama **292**(22), 2771–2776 (2004)
2. Bayasi, N., Fayyad, J., Bissoto, A., Hamarneh, G., Garbi, R.: Biaspruner: Debiased continual learning for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 90–101. Springer (2024)
3. Bissoto, A., Barata, C., Valle, E., Avila, S.: Even small correlation and diversity shifts pose dataset-bias issues. Pattern Recognition Letters **45** (2024)
4. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? not so fast. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020)
5. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. arXiv preprint arXiv:2405.19538 (2024)
6. Christodoulou, E., Reinke, A., Houhou, R., Kalinowski, P., Erkan, S., Sudre, C.H., Burgos, N., Boutaj, S., Loizillon, S., Solal, M., et al.: Confidence intervals uncovered: Are we ready for real-world medical imaging ai? In: International Conference

on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 124–132. Springer (2024)
7. d'Eon, G., d'Eon, J., Wright, J.R., Leyton-Brown, K.: The spotlight: A general method for discovering systematic errors in deep learning models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1962–1981 (2022)
8. Eyuboglu, S., Varma, M., Saab, K.K., Delbrouck, J.B., Lee-Messer, C., Dunnmon, J., Zou, J., Re, C.: Domino: Discovering systematic errors with cross-modal embeddings. In: International Conference on Learning Representations (2022)
9. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., Saria, S.: The Clinician and Dataset Shift in Artificial Intelligence. New England Journal of Medicine **385**(3), 283–286 (Jul 2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI Conference on Artificial Intelligence (2019)
12. Jiménez-Sánchez, A., Juodelyte, D., Chamberlain, B., Cheplygina, V.: Detecting shortcuts in medical images-a case study in chest x-rays. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
13. 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 (2024)
14. Kurtansky, N.R., D'Alessandro, B.M., Gillis, M.C., Betz-Stablein, B., Cerminara, S.E., Garcia, R., Girundi, M.A., Goessinger, E.V., Gottfrois, P., Guitera, P., et al.: The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. Scientific Data **11**(1), 884 (2024)
15. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Re, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: ACM Conference on Health, Inference, and Learning (2020)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
17. Roschewitz, M., Mehta, R., Jones, C., Glocker, B.: Automatic dataset shift identification to support root cause analysis of ai performance drift. arXiv preprint arXiv:2411.07940 (2024)
18. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., 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 (2021)
19. Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In: Proc. Medical Image Computing and Computer Assisted Interventions MICCAI (2023)
20. Yenamandra, S., Ramesh, P., Prabhu, V., Hoffman, J.: Facts: First amplify correlations and then slice to discover bias. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4794–4804 (2023)
21. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024)