Data AUDIT: Identifying Attribute Utility- and Detectability-Induced Bias in Task Models

Mitchell Pavlak[1,2](✉), Nathan Drenkow[1], Nicholas Petrick[2],
Mohammad Mehdi Farhangi[2], and Mathias Unberath[1]

[1] The Johns Hopkins University, Baltimore, MD, USA
mpavlak1@jhu.edu

[2] Center for Devices and Radiological Health, U.S. Food and Drug Administration,
Silver Spring, MD, USA

Abstract. To safely deploy deep learning-based computer vision models for computer-aided detection and diagnosis, we must ensure that they are robust and reliable. Towards that goal, algorithmic auditing has received substantial attention. To guide their audit procedures, existing methods rely on heuristic approaches or high-level objectives (e.g., non-discrimination in regards to protected attributes, such as sex, gender, or race). However, algorithms may show bias with respect to various attributes beyond the more obvious ones, and integrity issues related to these more subtle attributes can have serious consequences. To enable the generation of actionable, data-driven hypotheses which identify specific dataset attributes likely to induce model bias, we contribute a first technique for the rigorous, quantitative screening of medical image *datasets*. Drawing from literature in the causal inference and information theory domains, our procedure decomposes the risks associated with dataset attributes in terms of their detectability and utility (defined as the amount of information the attribute gives about a task label). To demonstrate the effectiveness and sensitivity of our method, we develop a variety of datasets with synthetically inserted artifacts with different degrees of association to the target label that allow evaluation of inherited model biases via comparison of performance against true counterfactual examples. Using these datasets and results from hundreds of trained models, we show our screening method reliably identifies nearly imperceptible bias-inducing artifacts. Lastly, we apply our method to the natural attributes of a popular skin-lesion dataset and demonstrate its success. Our approach provides a means to perform more systematic algorithmic audits and guide future data collection efforts in pursuit of safer and more reliable models. Full code is available at https://github.com/mpavlak25/data-audit.

M. Pavlak and N. Drenkow—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_43.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 442–452, 2023.
https://doi.org/10.1007/978-3-031-43898-1_43

Keywords: Bias · shortcut learning · fairness · algorithmic auditing · datasets

1 Introduction

Continual advancement of deep learning algorithms for medical image analysis has increased the potential for their adoption at scale. Across a wide range of medical applications including skin lesion classification [8,31], detection of diabetic retinopathy in fundus images [14], detection of large vessel occlusions in CT [20], and detection of pneumonia in chest x-ray [24], deep learning algorithms have pushed the boundaries close to or beyond human performance.

However, with these innovations has come increased scrutiny of the integrity of these models in safety critical applications. Prior work [7,10,17] has found that deep neural networks are capable of exploiting spurious features and other shortcuts in the data that are not causally linked to the task of interest such as using dermascopic rulers as cues to predict melanoma [2,35,36] or associating the presence of a chest drain with pneumothorax in chest X-ray analysis [21]. The exploitation of such shortcuts by DNNs may have serious bias/fairness implications [11,12] and negative ramifications for model generalization [7,21].

As attention to these issues grows, recent legislation has been proposed that would require the algorithmic auditing and impact assessment of ML-based automated decision systems [37]. However, without clearly defined strategies for selecting attributes to audit for bias, impact assessments risk being constrained to only legally protected categories and may miss more subtle shortcuts and data flaws that prevent the achievement of important model goals [23,28]. Our goal in this work is to develop objective methods for generating data-driven hypotheses about the relative level of risk of various attributes to better support the efficient, comprehensive auditing of any model trained on the same data.

Our method generates targeted hypotheses for model audits by assessing (1) how feasible it is for a downstream model to detect and exploit the presence of a given attribute from the image alone (detectability), and (2) how much information the model would gain about the task labels if said attribute were known (utility). Causally irrelevant attributes with high utility and detectability become top priorities when performing downstream model audits. We demonstrate high utility complicates attempts to draw conclusions about the detectability of attributes and show our approach succeeds where unconditioned approaches fail. We rigorously validate our approach using a range of synthetic artifacts which allow us to expedite the auditing of models via the use of true counterfactuals. We then apply our method to a popular skin lesion dataset where we identify a previously unreported potential shortcut.

2 Related Work

Issues of bias and fairness are of increasing concern in the research community. Recent works such as [13,29,30] identify cases where trained DNNs exhibit performance disparities across protected groups for chest x-ray classification tasks.

Of interest to this work, [22] used Mutual Information-based analysis to examine the robustness of DNNs on dermascopy data and observed performance disparities with respect to typical populations of interest (i.e., age, sex) as well as less commonly audited dataset properties (e.g., image hue, saturation). In addition to observing biased performance in task models, [11,12] show that patient race (and potentially other protected attributes) may be implicitly encoded in representations extracted by DNNs on chest x-ray images. A more general methodology for performing algorithmic audits in medical imaging is also proposed in [18]. In contrast to our work, these methods focus on individual, biased task models without considering the extent to which those biases are induced by the causal structure of the training/evaluation data.

In addition to model auditing methods, a number of metrics have been proposed to quantify bias [9]. A recent study [1] compared several and recommend normalized pointwise mutual information due to its ability to measure associations in the data while accounting for chance. Also relevant to this work, [16] provides an analysis of fairness metrics and guidelines for metric selection in the presence of dataset bias. However, these studies focus primarily on biases identifiable through dataset attributes alone and do not consider whether those attributes are detectable in the image data itself.

Lastly, [28] found pervasive *data cascades* where data quality issues compound and cause adverse downstream impacts for vulnerable groups. However, their study was qualitative and no methods for automated dataset auditing were introduced. Bissoto et al. [3,4] consider the impact of bias in dermatological data by manipulating images to remove potential causally-relevant features while measuring a model's ability to still perform the lesion classification task. Closest to our work, [25] takes a causal approach to shortcut identification by using conditional dependence tests to determine whether DNNs rely on specific dataset attributes for their predictions. In contrast, our work focuses on screening *datasets* for attributes that induce bias in task models. As a result, we directly predict attribute values to act as a strong upper bound on detectability and use normalized, chance-adjusted dependence measures to obtain interpretable metrics that we show correlate well with the performance of task models.

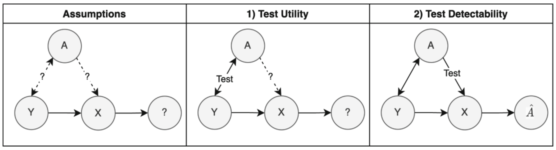


Fig. 1. Relationships assessed in the attribute screening protocol. DNNs are trained to predict $\hat{A}$ from X for use in estimating attribute detectability.

3 Methods

To audit at the dataset level, we perform a form of causal discovery to identify likely relationships between the task labels, dataset attributes represented as image metadata, and features of the images themselves (as illustrated in Fig. 1).

We start from a set of labels $\{Y\}$, attributes $\{A\}$, and images $\{X\}$. We assume that Y (the disease) is the causal parent of X (the image) given that the disease affects the image appearance but not vice versa [6]. Then the dataset auditing procedure aims to assess the existence and relative strengths of the following two relationships: (1) *Utility*: $A \leftrightarrow Y$ and (2) *Detectability*: $A \rightarrow X$. The utility measures whether a given attribute shares any relationship with the label. The presence of this relationship for A that are not clinically relevant (e.g., sensor type or settings) represents increased potential for biased outcomes. However, not every such attribute carries the same risk for algorithmic bias. Crucially, relationship (2) relates to the detectability of the attribute itself. If our test for (2) finds the existence of relationship $A \rightarrow X$ is probable, we consider the attribute detectable. Dataset attributes identified as having positive utility with respect to the label (1) and detectable in the image (2) are classified as potential shortcuts and pose the greatest risk to models trained on this dataset.

Causal Discovery with Mutual Information. Considering attributes in isolation, we assess attribute utility and detectability from an information theoretic perspective. In particular, we recognize first that the presence of a relationship between A and Y can be measured via their Mutual Information: $MI(A;Y) = H(Y) - H(Y|A)$. MI measures the information gained (or reduction in uncertainty) about Y by observing A (or vice versa) and $MI(A;Y) = 0$ occurs when A and Y are independent. We rely on the faithfulness assumption which implies that a causal relationship exists between A and Y when $MI(A;Y) > 0$. From an auditing perspective, we aim to identify the presence and relative magnitude of the relationship but not necessarily the nature of it.

Attributes identified as having a relationship with Y are then assessed for their detectability (i.e., condition (2)). We determine detectability by training a DNN on the data to predict attribute values. Because we wish to audit the entire dataset for bias, we cannot rely on a single train/val/test split. Instead, we partition the dataset into k folds (typically 3) and finetune a sufficiently expressive DNN on the train split of each fold to predict the given attribute A. We then generate unbiased predictions for the entire dataset by taking the output $\hat{A}$ from each DNN evaluated on their respective test split. We measure the Conditional Mutual Information over all predictions: $CMI(\hat{A}; A|Y) = H(\hat{A}|Y) - H(\hat{A}|A, Y)$. $CMI(A, \hat{A}|Y)$ measures information shared between attribute A and its prediction $\hat{A}$ when controlling for information provided by Y. Since relationship A and Y was established via $MI(A;Y)$, we condition on label Y to understand the extent to which attribute A can be predicted from images when accounting for features associated with Y that may also improve the prediction of A. Similar to MI, $CMI(\hat{A}; A|Y) > 0$ implies $A \rightarrow \hat{A}$ exists.

To determine independence and account for bias and dataset specific effects, we include permutation-based shuffle tests from [26,27]. These approaches replace values of A with close neighbors to approximate the null hypothesis that the given variables are conditionally independent. By calculating the percentile of $CMI(A;\hat{A}|Y)$ among all $CMI(A_\pi;\hat{A}|Y)$ (where A_π are permutations of A), we estimate the probability our samples are independent while adjusting for estimator bias and dataset-specific effects. To make CMI and MI statistics interpretable for magnitude-based comparison between attributes, we include adjustments for underlying distribution entropy and chance as per [1,34] (See supplement).

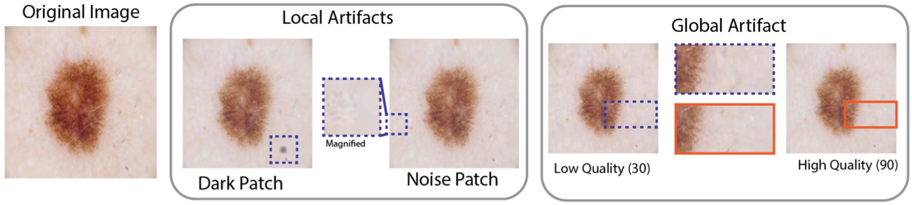


Fig. 2. Examples of synthetic artifacts with varying image effects.

4 Experiments and Discussion

To demonstrate the effectiveness of our method, we first conduct a series of experiments using synthetically-altered skin lesion data from the HAM10000 dataset where we precisely create, control, and assess biases in the dataset. After establishing the accuracy and sensitivity of our method on synthetic data, we apply our method to the natural attributes of HAM10000 in Experiment 5 (Sect. 4.5).

Datasets: We use publicly available skin lesion data from the HAM10000 [33] dataset with additional public metadata from [2]. The dataset consists of 10,015 dermascopic images collected from two sites, we filter so only one image per lesion is retained, leaving 7,387 images. The original dataset has seven diagnostic categories: we focus on predicting lesion malignancy as a challenging and practical task. While we recognize the importance of demonstrating the applicability of the methodology over many datasets, here we use trials where we perturb this dataset with a variety of synthetic, realistic artifacts (e.g., Fig. 2), and control association with the malignant target label. With this procedure, we create multiple variants of the dataset with attributes that have known utility and detectability as well as ground truth counterfactuals for task model evaluation. Further details are available in the supplementary materials.

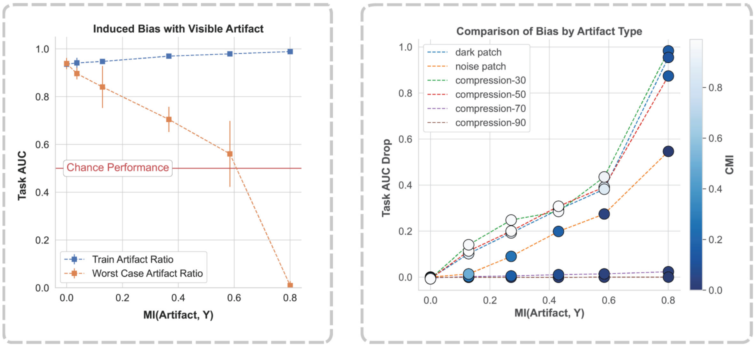
Training Protocol: For attribute prediction networks used by our detectability procedure, we finetune ResNet18 [15] models with limited data augmentation. For the malignancy prediction task, we use Swin Transformer [19] tiny models with RandAugment augmentation to show detectability results generalize to stronger architectures. All models were trained using class-balanced sampling with a batch size of 128 and the AdamW optimizer with a learning rate of 5e−5, linear decay schedule, and default weight decay and momentum parameters. For each trial, we use three-fold cross-validation and subdivide each training fold in a (90:10) `train:validation` split to select the best models for the relevant test fold. By following this procedure, we get unbiased artifact predictions over the entire dataset for use by MI estimators by aggregating predictions over all test folds. We generally measure model performance via the Receiver Operating Characteristic Area Under the Curve (AUC).

4.1 Experiment 1: Induced Bias Versus Relationship Strength

For this experiment, we select an artifact that *we are certain is visible* (JPEG compression at quality 30 applied to 1000 images), and seek to understand how the relationship between attribute and task label influences the task model's reliance on the attribute. The artifact is introduced with increasing utility such that the probability of the artifact is higher for cases that are malignant. Then, we create a worst case counterfactual set, where each malignant case does not have the artifact, and each benign case does. In Fig. 3a, we see performance rapidly declines *below random chance* as utility increases.

4.2 Experiment 2: Detectability of Known Invisible Artifacts

In the previous section, we showed that the utility $A \leftrightarrow Y$ directly impacts the task model bias, *given A is visible in images.* However, it is not always



(a) (b)

Fig. 3. (a) Performance of task models trained on data with known detectable artifacts introduced with various positive correlations to the malignant class and evaluated on our worst-case counterfactual set. (b) Performance drop on worst-case counterfactual test set of models trained with various artifacts of unknown detectability. For each, $MI(A, Y)$, $CMI(A, \hat{A}|Y)$ values are estimated empirically with normalization and adjustment for chance applied.

obvious whether an attribute is visible. In *Reading Race*, Gichoya et al. showed racial identity can be predicted with high AUC from medical images where this information is not expected to be preserved. Here, we show CMI represents a promising method for determining attribute detectability while controlling for attribute information communicated through labels and not through images.

Specifically, we consider the case of an "invisible artifact". We make no changes to the images, but instead create a set of randomized labels for our non-existent artifact that have varying correlation with the task labels ($A \leftrightarrow Y$). As seen in Fig. 4, among cases where the invisible artifact and task label have a reasonable association, models tasked with predicting the invisible artifact perform well above random chance, seemingly indicating that these artifacts are visible in images. However, by removing the influence of task label and related image features by calculating $MI(A, \hat{A}|Y)$, we clearly see that the artifact predictions are independent of the labels, meaning there is no visible attribute. Instead, all information about the attribute is inferred from the task label.

4.3 Experiment 3: Conditioned Detectability Versus Ground Truth

To verify that the conditional independence testing procedure does not substantially reduce our ability to correctly identify artifacts that truly are visible, we introduce Gaussian noise with standard deviation decreasing past human perceptible levels. In Experiment 2 (Fig 4A) the performance of detecting artifact presence is artificially inflated because of a relationship between disease and artifact. Here the artifact is introduced at random so AUC is an unbiased measure of detectability. In Table 1, we see the drop in CMI percentile from conditioning is minimal, indicating sensitivity even to weakly detectable attributes.

4.4 Experiment 4: Relationship and Detectability vs Induced Bias

Next, we consider how utility and detectability together relate to bias. We introduce a variety of synthetic artifacts and levels of bias and measure the drop in

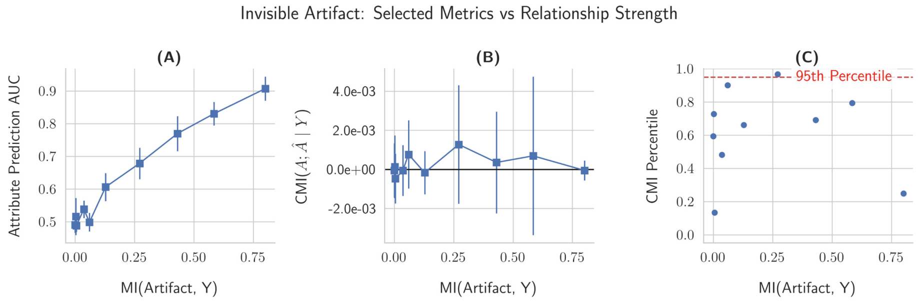


Fig. 4. (**A**) AUC for attribute prediction vs utility ($MI(A;Y)$). The models learn to fit a non-existent artifact given sufficient utility $A \leftrightarrow Y$. (**B**) $CMI(A;\hat{A}|Y)$ reported with 95% CIs (calculated via bootstrap) for the same models and predictions, each interval includes zero, suggesting conditional independence. (**C**) CMI statistic percentile vs 1000 trials with data permuted to be conditionally independent.

Table 1. Detectability of Gaussian noise with varying strength vs independence testing-based percentile and CMI(A; $\hat{A}|Y$), normalized and adjusted for chance. Anecdotally, $\sigma = 0.05$ is the minimum level that is visible (see supplement).

Noise Detectability vs Ground Truth									
Metric	Gaussian Noise σ								
	.5	.4	.3	.2	.1	.05	.01	.001	0
Attribute Prediction AUC	1.0 ± 0	1.0 ± 0	1.0 ± 0	1.0 ± 0	1.0 ± 0	1.0 ± 0	$0.71 \pm .075$	$0.53 \pm .017$	$0.52 \pm .036$
CMI(A; $\hat{A}|Y$)	1.0	1.0	1.0	1.0	0.997	0.997	0.0496	$-4.44\text{e}{-4}$	$-4.7\text{e}{-5}$
CMI Statistic Percentile	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.089	0.595

AUC that occurs when evaluated on a test set with artifacts introduced in the same ratio as training versus the worst case ratio as defined in Experiment 1.

Of 36 unique attribute-bias combinations trialed, 32/36 were correctly classified as visible via permutation test with 95% cutoff percentile. The remaining four cases were all compression at quality 90 and had negligible impact on task models (mean drop in AUC of $-0.0003 \pm .0006$). In Fig. 3b, we see the relative strength of the utility ($A \leftrightarrow Y$) correlates with the AUC drop observed. This implies utility represents a useful initial metric to predict the risk of an attribute. The detectability, $CMI(A; \hat{A}|Y)$, decreases as utility, $MI(A; Y)$, increases, implying the two are not independent. Intuitively, when A and Y are strongly related (Utility is high), knowledge of the task label means A is nearly determined, so learning $\hat{A}$ does not convey much new information and detectability is smaller. To combat this, we use a conditional permutation method [27] for judging whether or not an artifact is present. Further, detectability among attributes with equal utility for each level above 0 have statistically significant correlations with drops in AUC (Kendall's τ of 0.800, 0.745, 0.786, 0.786, 0.716 respectively). From this, we expect that for attributes with roughly equal utility, more detectable attributes are more likely to result in biased task models.

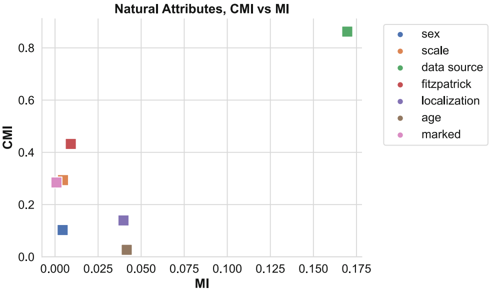


Fig. 5. Detectability vs. utility for natural attributes in HAM10000. Attributes with high CMI and MI which are non-causally related to the disease pose the greatest risk.

4.5 Experiment 5: HAM10000 Natural Attributes

Last, we run our screening procedure over the natural attributes of HAM10000 and find that all pass the conditional independence tests of detectability. Based on our findings, we place the attributes in the following order of concern: (1) Data source, (2) Fitzpatrick Skin Scale, (3) Ruler Presence, (4) Gentian marking presence (we skip localization, age and sex due to clinical relevance [5]). From Fig. 5 we see data source is both more detectable and higher in utility than other variables of interest, representing a potential shortcut. To the best of our knowledge, we are the first to document this concern, though recent independent work supports our result that differences between the sets are detectable [32].

5 Conclusions

Our proposed method marks a positive step forward in anticipating and detecting unwanted bias in machine learning models. By focusing on dataset screening, we aim to prevent downstream models from inheriting biases already present and exploitable in the data. While our screening method naturally includes common auditing hypotheses (e.g., bias/fairness for vulnerable groups), it is capable of generating targeted hypotheses on a much broader set of attributes ranging from sensor information to clinical collection site. Future work could develop unsupervised methods for discovering additional high risk attributes without annotations. The ability to identify and investigate these hypotheses provides broad benefit for research, development, and regulatory efforts aimed at producing safe and reliable AI models.

Acknowledgements. This project was supported in part by an appointment to the Research Participation Program at the U.S. Food and Drug Administration administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

References

1. Aka, O., Burke, K., Bauerle, A., Greer, C., Mitchell, M.: Measuring model biases in the absence of ground truth. In: AAAI/ACM AIES. ACM (2021)
2. Bevan, P., Atapour-Abarghouei, A.: Skin deep unlearning: artefact and instrument debiasing in the context of melanoma classification (2021)
3. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (De) constructing bias on skin lesion datasets. In: IEEE CVPRW (2019)
4. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? Not so fast. In: IEEE CVPRW, pp. 740–741 (2020)
5. Carr, S., Smith, C., Wernberg, J.: Epidemiology and risk factors of melanoma. Surg. Clin. North Am. **100**, 1–12 (2020)
6. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nat. Commun. **11**, 3673 (2020). https://doi.org/10.1038/s41467-020-17478-w

7. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell. **3**, 610–619 (2021). https://doi.org/10.1038/s42256-021-00338-7
8. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)
9. Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., Kompatsiaris, I.: A survey on bias in visual datasets. Comput. Vis. Image Underst. **223**, 103552 (2022)
10. Geirhos, R., et al.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**, 665–673 (2020)
11. Gichoya, J.W., et al.: AI recognition of patient race in medical imaging: a modelling study. Lancet Digit. Health **4**, e406–e414 (2022)
12. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in image-based models for disease detection (2021)
13. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Risk of bias in chest x-ray foundation models, September 2022
14. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**, 2402–2410 (2016)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVPR, pp. 770–778 (2016)
16. Henry Hinnefeld, J., Cooman, P., Mammo, N., Deese, R.: Evaluating fairness metrics in the presence of dataset bias, September 2018
17. Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M.W., Wiens, J.: Deep learning applied to chest X-rays: exploiting and preventing shortcuts. In: Machine Learning for Healthcare Conference, pp. 750–782. PMLR (2020)
18. Liu, X., Glocker, B., McCradden, M.M., Ghassemi, M., Denniston, A.K., Oakden-Rayner, L.: The medical algorithmic audit. Lancet Digit Health **4**, e384–e397 (2022)
19. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE/CVPR (2021)
20. Murray, N.M., Unberath, M., Hager, G.D., Hui, F.K.: Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. J. NeuroInterv. Surg. **12**, 156–164 (2020)
21. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 151–159 (2020)
22. O'Brien, M., Bukowski, J., Hager, G., Pezeshk, A., Unberath, M.: Evaluating neural network robustness for melanoma classification using mutual information. In: Medical Imaging 2022: Image Processing. SPIE (2022)
23. Raji, I.D., Kumar, I.E., Horowitz, A., Selbst, A.: The fallacy of AI functionality. In: ACM Conference on Fairness, Accountability, and Transparency. ACM (2022)
24. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
25. Reimers, C., Penzel, N., Bodesheim, P., Runge, J., Denzler, J.: Conditional dependence tests reveal the usage of ABCD rule features and bias variables in automatic skin lesion classification. In: IEEE CVPRW (2021)
26. Runge, J.: Causal network reconstruction from time series: from theoretical assumptions to practical estimation. Chaos **28**, 075310 (2018)
27. Runge, J.: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: AISTATS. PMLR (2018)

28. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M.: "everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In: ACM CHI. ACM (2021)
29. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: CheXclusion: fairness gaps in deep chest X-ray classifiers. In: Pacific Symposium on Biocomputing (2021)
30. 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. Nat. Med. **27**, 2176–2182 (2021)
31. Soenksen, L.R., et al.: Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Sci. Transl. Med. **13**, eabb3652 (2021)
32. Somfai, E., et al.: Handling dataset dependence with model ensembles for skin lesion classification from dermoscopic and clinical images. Int. J. Imaging Syst. Technol. **33**(2), 556–571 (2023)
33. 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**, 180161 (2018)
34. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. JMLR **11**, 2837–2854 (2010)
35. Winkler, J.K., et al.: Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. **155**, 1135–1141 (2019)
36. Winkler, J.K., et al.: Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. Eur. J. Cancer **145**, 146–154 (2021)
37. Wyden, R., Booker, C., Clarke, Y.: Algorithmic accountability act of 2022 (2022)